Cisco Unified Communications System Release 9.0 SRND

April 30, 2013

Americas Headquarters
Cisco Systems, Inc.
170 West Tasman Drive
San Jose, CA 95134-1706
USA
http://www.cisco.com
Tel: 408 526-4000
800 553-NETS (6387)
Fax: 408 527-0883

Text Part Number: OL-27282-05

Any Internet Protocol (IP) addresses and phone numbers used in this document are not intended to be actual addresses and phone numbers. Any examples, command display output, network topology diagrams, and other figures included in the document are shown for illustrative purposes only. Any use of actual IP addresses or phone numbers in illustrative content is unintentional and coincidental.

Cisco Unified Communications System 9.0 SRND

CONTENTS

PART 3 **Unified Communications Call Control**

CHAPTER 16

CHAPTER 17

Preface

Revised: April 30, 2013; OL-27282-05

This document provides design considerations and guidelines for deploying Cisco Unified Communications System Release 9.0, including Cisco Unified Communications Manager 9.0 and other components of the Cisco Unified Communications System.

This document should be used in conjunction with other documentation available at the following locations:

- For other Solution Reference Network Design (SRND) documents:

  http://www.cisco.com/go/ucsrnd

- For more information about the Cisco Unified Communications System:

  http://www.cisco.com/go/unified-techinfo

- For more information about Cisco Unified Communications Manager:

  http://www.cisco.com/en/US/products/sw/voicesw/ps556/tsd_products_support_series_home.html

- For other Cisco design guides:

  http://www.cisco.com/go/designzone

- For all Cisco products and documentation:

  http://www.cisco.com

New or Changed Information for This Release

Within each chapter of this guide, new and revised information is listed in a section titled *What's New in This Chapter.*

Note Unless stated otherwise, the information in this document applies only to Release 9.0 of the Cisco Unified Communications System and its components. For information on later 9.*x* releases of Cisco Unified Communications, refer to the *Cisco Collaboration 9.x SRND* available at http://www.cisco.com/en/US/docs/voice_ip_comm/cucm/srnd/collab09/clb09.html.

For additional information about Cisco Unified Communications and Collaboration solutions, refer to the documentation available at:

http://www.cisco.com/go/ucsrnd

Revision History

Note This document is no longer being updated. For the latest information on Cisco Unified Communications 9.*x* system releases, refer to the *Cisco Collaboration 9.x SRND* available at http://www.cisco.com/en/US/docs/voice_ip_comm/cucm/srnd/collab09/clb09.html.

The following table lists the revision history for this document.

Revision Date	Document Part Number	Comments
April 30, 2013	OL-27282-05	Minor corrections and changes. For details, in each chapter see *What's New in This Chapter.*
October 31, 2012	OL-27282-04	Periodic update of various chapters. For details, in each chapter see *What's New in This Chapter.*
September 28, 2012	OL-27282-03	Periodic update of various chapters. For details, in each chapter see *What's New in This Chapter.*
August 31, 2012	OL-27282-02	Periodic update of various chapters. For details, in each chapter see *What's New in This Chapter.*
June 28, 2012	OL-27282-01	Initial version of this document for Cisco Unified Communications System Release 9.0.

Obtaining Documentation and Submitting a Service Request

For information on obtaining documentation, submitting a service request, and gathering additional information, see the monthly *What's New in Cisco Product Documentation*, which also lists all new and revised Cisco technical documentation, at:

http://www.cisco.com/en/US/docs/general/whatsnew/whatsnew.html

Subscribe to the *What's New in Cisco Product Documentation* as a Really Simple Syndication (RSS) feed and set content to be delivered directly to your desktop using a reader application. The RSS feeds are a free service and Cisco currently supports RSS Version 2.0.

Cisco Product Security Overview

This product contains cryptographic features and is subject to United States and local country laws governing import, export, transfer and use. Delivery of Cisco cryptographic products does not imply third-party authority to import, export, distribute, or use encryption. Importers, exporters, distributors and users are responsible for compliance with U.S. and local country laws. By using this product you agree to comply with applicable laws and regulations. If you are unable to comply with U.S. and local laws, return this product immediately.

Further information regarding U.S. export regulations may be found at:

http://www.access.gpo.gov/bis/ear/ear_data.html

Conventions

This document uses the following conventions:

Convention	Indication
bold font	Commands and keywords and user-entered text appear in **bold** font.
italic font	Document titles, new or emphasized terms, and arguments for which you supply values are in *italic* font.
[ ]	Elements in square brackets are optional.
{x | y | z }	Required alternative keywords are grouped in braces and separated by vertical bars.
[x | y | z]	Optional alternative keywords are grouped in brackets and separated by vertical bars.
string	A nonquoted set of characters. Do not use quotation marks around the string or the string will include the quotation marks.
`courier` font	Terminal sessions and information the system displays appear in `courier` font.
< >	Non-printing characters such as passwords are in angle brackets.
[ ]	Default responses to system prompts are in square brackets.
!, #	An exclamation point (!) or a pound sign (#) at the beginning of a line of code indicates a comment line.

Note Means *reader take note*.

Tip Means *the following information will help you solve a problem*.

Caution Means *reader be careful*. In this situation, you might perform an action that could result in equipment damage or loss of data.

Timesaver Means *the described action saves time*. You can save time by performing the action described in the paragraph.

Warning **Means *reader be warned*. In this situation, you might perform an action that could result in bodily injury.**

CHAPTER 1

Introduction

Revised: November 15, 2010; OL-27282-05

The Cisco Unified Communications System delivers fully integrated communications by enabling data, voice, and video to be transmitted over a single network infrastructure using standards-based Internet Protocol (IP). Leveraging the framework provided by Cisco IP hardware and software products, the Cisco Unified Communications System delivers unparalleled performance and capabilities to address current and emerging communications needs in the enterprise environment. The Cisco Unified Communications family of products is designed to optimize feature functionality, reduce configuration and maintenance requirements, and provide interoperability with a wide variety of other applications. The Cisco Unified Communications System provides this capability while maintaining a high level of availability, quality of service (QoS), and security for your network.

The Cisco Unified Communications System incorporates and integrates the following major communications technologies:

- IP telephony

  IP telephony refers to technology that transmits voice communications over a network using IP standards. Cisco Unified Communications includes a wide array of hardware and software products such as call processing agents, IP phones (both wired and wireless), voice messaging systems, video devices, and many special applications.

- Customer contact center

  Cisco Unified Contact Center products are a combination of strategy and architecture that promote efficient and effective customer communications across a globally capable network by enabling organizations to draw from a broader range of resources to service customers. They include access to a large pool of agents and multiple channels of communication as well as customer self-help tools.

- Video telephony

  The Cisco Unified Video Advantage products enable real-time video communications and collaboration using the same IP network and call processing agent as Cisco Unified Communications. With Cisco Unified Video Advantage, making a video call is as easy as dialing a phone number.

- Rich-media conferencing

  Cisco Unified MeetingPlace, Cisco Unified Videoconferencing, and Cisco WebEx Software as a Service enhance the virtual meeting environment with a integrated set of IP-based tools for voice, video, and web conferencing.

- Mobility

  Cisco wireless and mobility solutions enable users to increase productivity and responsiveness by enabling access to network resources and applications securely, regardless of location or client device.

- TelePresence

  Cisco TelePresence delivers real-time, face-to-face interactions between people and places in their work and personal lives using advanced visual, audio, and collaboration technologies. These technologies transmit life-size, high-definition images and spatial discrete audio that make users feel like they are in the same room even when they are half a world away.

- Applications

  Cisco provides numerous embedded applications and also works with leading-edge companies to provide the broadest selection of innovative third-party unified communications applications and products focused on critical business needs such messaging, customer care, and workforce optimization.

The remainder of this document focuses on system design considerations for deploying these technologies and applications in the Cisco Unified Communications System.

For information about other aspects of the Cisco Unified Communications System, refer to the documentation available at the following locations:

http://www.cisco.com/go/ucsrnd

http://www.cisco.com/go/unified-techinfo

You can also find additional documentation for the Cisco Unified Communications family of products at the following location:

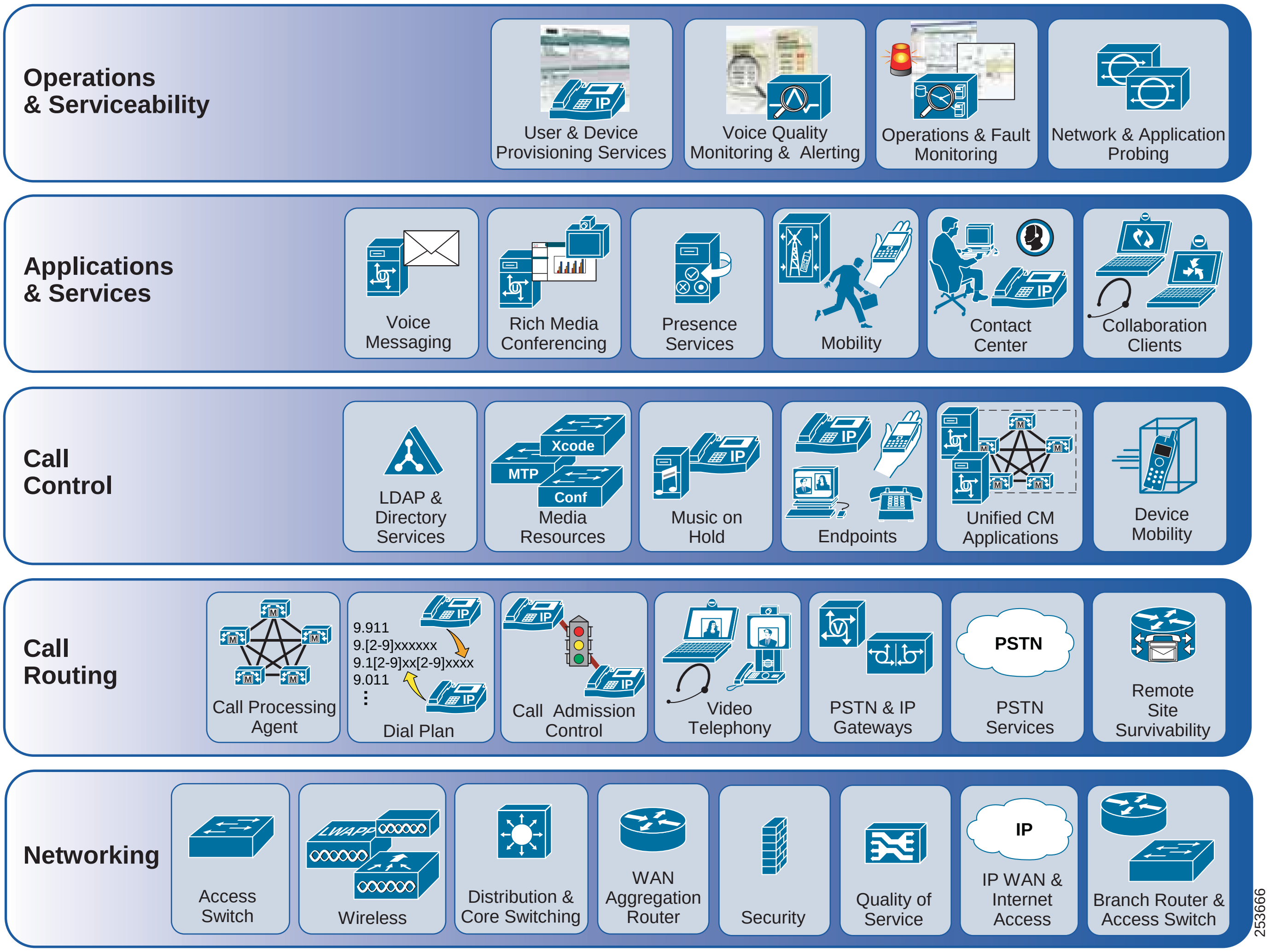
http://www.cisco.com

Note The design guidance in this document applies to Cisco customers and partners who want to deploy an Enterprise Unified Communications solution. For those interested in hosted or managed Unified Communication solutions, please refer to http://www.cisco.com/go/hostedcollab for more information.

Cisco Unified Communications System Architecture

Figure 1-1 illustrates the layered architecture of the Cisco Unified Communications System.

Figure 1-1 Architecture of the Cisco Unified Communications System



The various layers of the Cisco Unified Communications System perform the following major tasks and roles:

- Networking

  This layer forms the foundation for the Unified Communications network. It includes components that provide the following functions and capabilities:

  – Network infrastructure ensures a redundant and resilient network foundation with Quality of Service (QoS) enabled for Unified Communications applications.

  – Voice security ensures a general security policy and a hardened and secure networking foundation for Unified Communications applications.

  - Unified Communications deployment models provide tested models as well as best practices and design guidelines for deploying a Unified Communications System.
  - IP telephony migration options provide guidelines on how to plan and approach a migration from standalone voice, video, and collaboration systems to an integrated Cisco Unified Communications System.

  For more information on the Networking layer, see the Overview of Cisco Unified Communications Networking, page 2-1.

- Call Routing

  This layer handles the processing and routing of calls throughout the system. It includes components that provide the following functions and capabilities:

  - Call processing agents provides telephony services and call routing capabilities.
  - The dial plan provides endpoint numbering, dialed digits analysis, and classes of restriction to limit types of calls that a user can make.
  - Call admission control provides mechanisms for preventing oversubscription of network bandwidth by limiting the number of calls that are allowed on the network at a given time, based on overall call capacity of the call processing components and network bandwidth.
  - Video telephony services provide the ability to provision and register video endpoints as well as to set up, route, and maintain video calls on the network.
  - PSTN gateways and provider voice and data services provide access to voice and data networks outside the enterprise, including the PSTN, Internet, and service provider IP-based trunks.
  - Remote site survivability provides continuation of basic telephony services at remote sites when the central-site telephony services are unavailable due to failed or flapping network connectivity.

  For more information on the Call Routing layer, see the Overview of Cisco Unified Communications Call Routing, page 7-1.

- Call Control

  This layer enables users to initiate and manage calls. It includes components that provide the following functions and capabilities:

  - Integration with central Lightweight Directory Access Protocol (LDAP) directories enables companies to centralize all user information in a single repository available to Unified Communications applications, with a reduction in maintenance costs through the ease of adds, moves, and changes.
  - Access to media resources provides media processing functions such as conferencing, media termination, transcoding, echo cancellation, signaling, packetization of a stream, streaming audio (annunciation), and so forth.
  - Music on hold provides music (or advertising) to callers when their call is placed on hold, transferred, parked, or added to an ad-hoc conference.
  - Unified Communications endpoints and feature sets range from gateways that support ordinary analog phones in an IP environment to an extensive set of native IP phones offering a range of capabilities for the end user.
  - Device mobility features enable mobile users to roam from one site to another with their endpoint devices and to acquire the dynamically allocated settings of their roaming site for call routing, codec section, media resource selection, and so forth.

  - Applications embedded in the call control software provide features such as click-to-call dialing, manager-assistant applications, and the ability for users to log in to any phone, as well as support for web-based applications that can run directly on the user's desktop phone.

  For more information on the Call Control layer, see the Overview of Cisco Unified Communications Call Control, page 15-1.

- Applications and Services

  This layer contains numerous applications and services that can be deployed on top of an existing Cisco Unified Communications infrastructure to add enhanced user features to the system. It includes components that provide the following functions and capabilities:

  - Voice messaging provides voicemail services and message waiting indication.
  - Rich media conferencing provides audio and video conferencing as well as web-based application and document sharing.
  - Presence services provide user availability tracking across user devices and clients.
  - Mobility services provide enterprise-level Unified Communications features and functionality to users outside the enterprise.
  - Contact center applications provide call handling, queuing, and monitoring for large call volumes.
  - Collaboration client services provide integration to Unified Communications services and leveraging of various applications.

  For more information on the Applications and Services layer, see the Overview of Cisco Unified Communications Applications and Services, page 20-1.

- Operations and Serviceability

  This layer contains system-level services for monitoring and managing the Unified Communications network and applications. It includes components that provide the following functions and capabilities:

  - User and device provisioning services provide centralized provisioning and configuration of users and devices for Unified Communications applications and services.
  - Voice quality monitoring and alerting provide the ability to monitor various call flows within the system to determine whether voice quality is acceptable and to alert administrators when the voice quality is not acceptable.
  - Operations and fault monitoring provides the ability to monitor all application and service operations and to issue alerts to administrators regarding network and application failures.
  - Network and application probing provides the ability to probe and collect network and application traffic information at various locations throughout the deployment and to allow administrators to access and retrieve this information from a central location.

  For more information on the Operations and Serviceability layer, see the Overview of Cisco Unified Communications Operations and Serviceability, page 27-1.

Note The design recommendations in this guide have been reviewed and found to be consistent with the Cisco Borderless Network Smart Business Architecture (SBA). Contact your Cisco representative for more information on SBA.

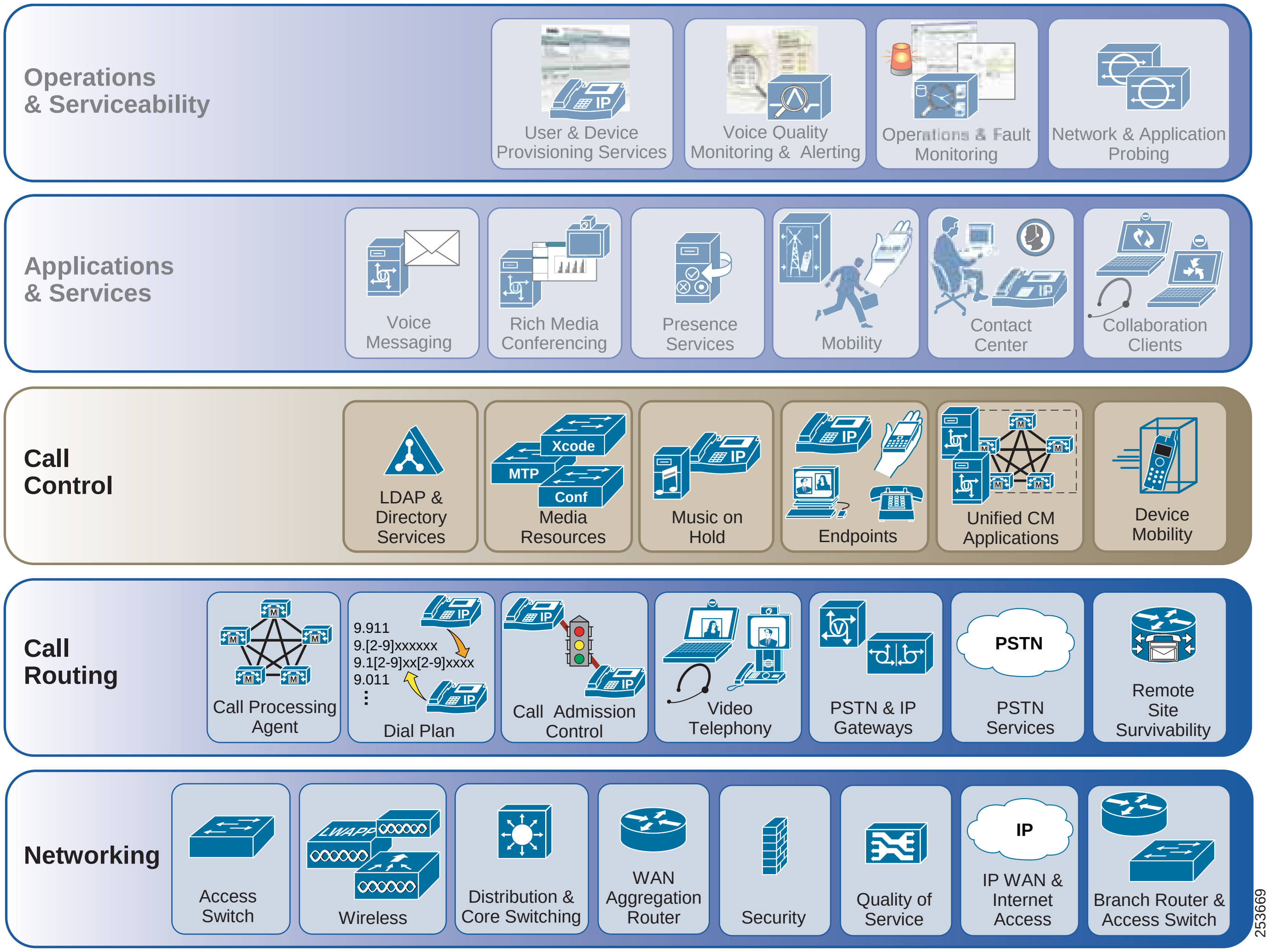
How to Use This Design Guide

This document provides design considerations, guidelines, and best practices for deploying a Cisco Unified Communications System. As discussed in the previous section, the architecture of the Cisco Unified Communications System consists of five layers. This document is divided into five parts corresponding to the five architectural layers. Each part of this document contains chapters that describe the components and design guidelines for the corresponding architectural layer.

The process for building a good Unified Communications system is similar to building a house: first you have to establish a solid infrastructure and foundation upon which to build all the other layers. And the other layers must be added in a particular sequence, usually from the bottom up. (For example, you have to build the walls of a house before you can put a roof on it.) In the case of a Unified Communications system, the networking layer provides the infrastructure, and the other layers must be added from the bottom up in the sequence shown in Figure 1-1. The parts and chapters of this guide are organized in that same sequence to help you establish a logical process for designing your Unified Communications system.

The first chapter in each part of this guide presents an overview of the information contained in that part. The overview includes an illustration of the five architectural layers of the Cisco Unified Communications System, and the layer being discussed in that part of the guide is highlighted. For example, Figure 1-2 is the illustration from the Call Control part of this guide. The Call Control layer is highlighted in a different color to show that it is the layer being discussed in that part of the guide. The layers below it (Networking and Call Routing) are visible because they must already be in place before the Call Control layer can be implemented. The layers above it (Applications & Services and Operations & Serviceability) are shaded to indicate that they cannot be implemented until the current layer (Call Control in this example) is in place.

Figure 1-2 Cisco Unified Communications Call Control Architecture



If you are designing a new Unified Communications system, Cisco recommends that you develop your design according to the sequence and guidelines presented in this document. If you already have some layers of the system in place and you want to add other layers to it, Cisco recommends that you at least review the sections of this guide that pertain to the existing layers to ensure that your system complies with all the guidelines.

PART 1

Unified Communications Networking

CHAPTER 2

Overview of Cisco Unified Communications Networking

Revised: February 29, 2012; OL-27282-05

A solid network infrastructure is required to build a successful Unified Communications system in an enterprise environment. Other key aspects of the network architecture include voice security, unified communications deployment models, and migration strategies.

Unified Communications – including IP telephony, rich media, collaboration, and many other functions – places strict requirements on IP packet loss, packet delay, and delay variation (or jitter). Therefore, you need to enable most of the Quality of Service (QoS) mechanisms available on Cisco switches and routers throughout the network. For the same reasons, redundant devices and network links that provide quick convergence after network failures or topology changes are also important to ensure a highly available infrastructure. The following aspects are essential to the topic of Unified Communications networking and are specifically organized here in order of importance and relevance to one another:

- Network Infrastructure — Ensures a redundant and resilient foundation with QoS enabled for Unified Communications applications.
- Voice Security — Ensures a general security policy for Unified Communications applications and a hardened and secure networking foundation for them to rely upon.
- Unified Communications Deployment Models — Provide tested models in which to deploy Unified Communications call control and applications, as well as best practices and design guidelines to apply to Unified Communications deployments.
- IP Telephony Migration Options — Provide guidelines on how to plan and approach a migration from separate standalone voice, video, and collaboration systems to an integrated Cisco Unified Communications System.

The chapters in this part of the SRND cover the networking subjects mentioned above. Each chapter provides an introduction to the subject matter, followed by discussions surrounding architecture, high availability, capacity planning, and design considerations. The chapters focus on design-related aspects rather than product-specific support and configuration information, which is covered in the related product documentation.

This part of the SRND includes the following chapters:

- Network Infrastructure, page 3-1

  This chapter describes the requirements of the network infrastructure needed to build a Cisco Unified Communications System in an enterprise environment. The sections in this chapter describe the network infrastructure features as they relate to LAN, WAN, and wireless LAN infrastructures. The chapters treat the areas of design, high availability, quality of service, and bandwidth provisioning as is pertinent to each infrastructure.

- Unified Communications Security, page 4-1

  This chapter presents guidelines and recommendations for securing Unified Communications networks. The topics in this chapter range from general security, such as policy and securing the infrastructure, to phone security in VLANs, on switch ports, and with QoS. Other security aspects covered in this chapter include access control lists, securing gateways and media resources, firewalls, data center designs, securing application servers, and network virtualization.

- Unified Communications Deployment Models, page 5-1

  This chapter describes the deployment models for Cisco Unified Communications Manager as they relate to the various network infrastructures such as a single site or campus, multi-site environments, and data center solutions. This chapter covers these deployment models and the best practices and design considerations for each model, including many other subtopics pertinent to the model discussed.

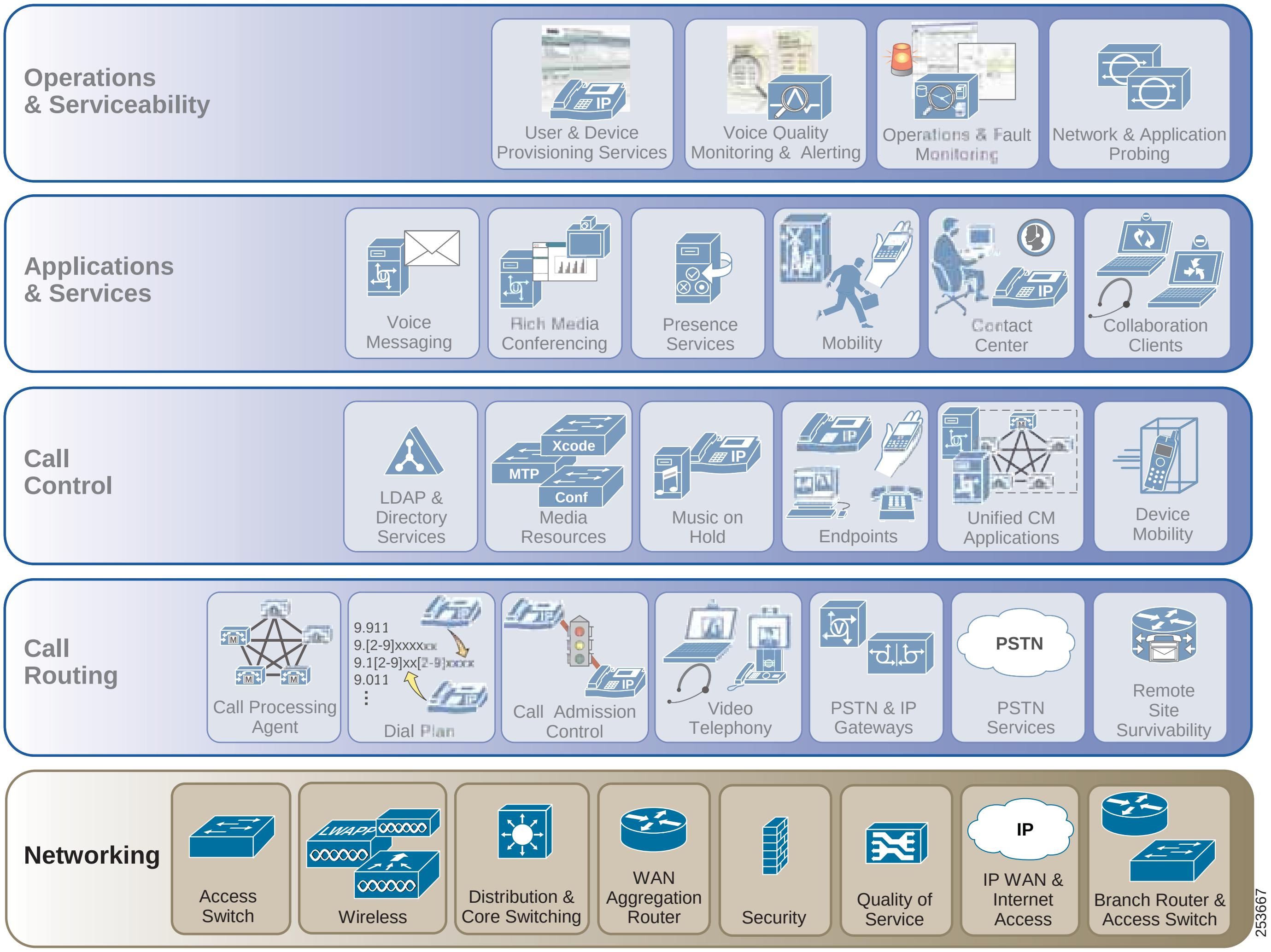
- IP Telephony Migration Options, page 6-1

  This chapter describes several methods for migrating from separate standalone voice, video, and collaboration systems to an integrated Cisco Unified Communications System. It discusses the pros and cons of both phased migration and parallel cutover. It also describes the services needed to connect a private branch exchange (PBX) to a new Unified Communications system. The major topics discussed in this chapter include IP telephony migration, video migration, and migration of voice and desktop collaboration systems.

Architecture

The networking architecture lays the foundation upon which all other layers of the Unified Communications System are deployed. Figure 2-1 shows the logical location of the networking layer in the overall Cisco Unified Communications System architecture.

Figure 2-1 *Cisco Unified Communications Networking Architecture*



All other layers of the Unified Communications System architecture, including call routing, call control, applications and services, and operations and serviceability, rely heavily on the readiness of the network to support their services. The networking layer is the single most important aspect of a solid Unified Communications foundation in that it provides the quality of service needed to ensure applications have uncompromised access to network services. The networking layer also ensures the correct deployment of servers and the proper bandwidth for endpoints and services to communicate effectively and securely.

High Availability

Proper design of the network infrastructure requires building a robust and redundant network from the bottom up. By structuring the LAN as a layered model (access, distribution, and core layers) and developing the LAN infrastructure one step of the model at a time, you can build a highly available, fault tolerant, and redundant network. Proper WAN infrastructure design is also extremely important for normal IP telephony operation on a converged network. Proper infrastructure design requires following basic configuration and design best-practices for deploying a WAN that is as highly available as possible and that provides guaranteed throughput. Furthermore, proper WAN infrastructure design requires deploying end-to-end QoS on all WAN links.

Wireless LAN infrastructure design becomes important when IP telephony is added to the wireless LAN (WLAN) portions of a converged network. With the addition of wireless Unified Communications endpoints such as the Cisco Unified Wireless IP Phones 7921G and 7925G, voice traffic has moved onto the WLAN and is now converged with the existing data traffic there. Just as with wired LAN and wired WAN infrastructures, the addition of voice in the WLAN requires following basic configuration and design best-practices for deploying a highly available network. In addition, proper WLAN infrastructure design requires understanding and deploying QoS on the wireless network to ensure end-to-end voice quality on the entire network.

After designing and implementing the network infrastructure properly, you can add network and application services successfully across the network, thus providing a highly available foundation upon which your Unified Communications services can run.

Capacity Planning

Scaling your network infrastructure to handle the Unified Communications applications and services that it must support requires providing adequate available bandwidth and the capability to handle the additional traffic load created by the applications.

For a complete discussion of system sizing, capacity planning, and deployment considerations related to sizing, refer to the chapter on Unified Communications Design and Deployment Sizing Considerations, page 29-1.

CHAPTER 3

Network Infrastructure

Revised: April 30, 2013; OL-27282-05

This chapter describes the requirements of the network infrastructure needed to build a Cisco Unified Communications System in an enterprise environment. Figure 3-1 illustrates the roles of the various devices that form the network infrastructure, and Table 3-1 summarizes the features required to support each of these roles.

Unified Communications places strict requirements on IP packet loss, packet delay, and delay variation (or jitter). Therefore, you need to enable most of the Quality of Service (QoS) mechanisms available on Cisco switches and routers throughout the network. For the same reasons, redundant devices and network links that provide quick convergence after network failures or topology changes are also important to ensure a highly available infrastructure

The following sections describe the network infrastructure features as they relate to:

- LAN Infrastructure, page 3-4
- WAN Infrastructure, page 3-34
- Wireless LAN Infrastructure, page 3-54

Figure 3-1 Typical Campus Network Infrastructure

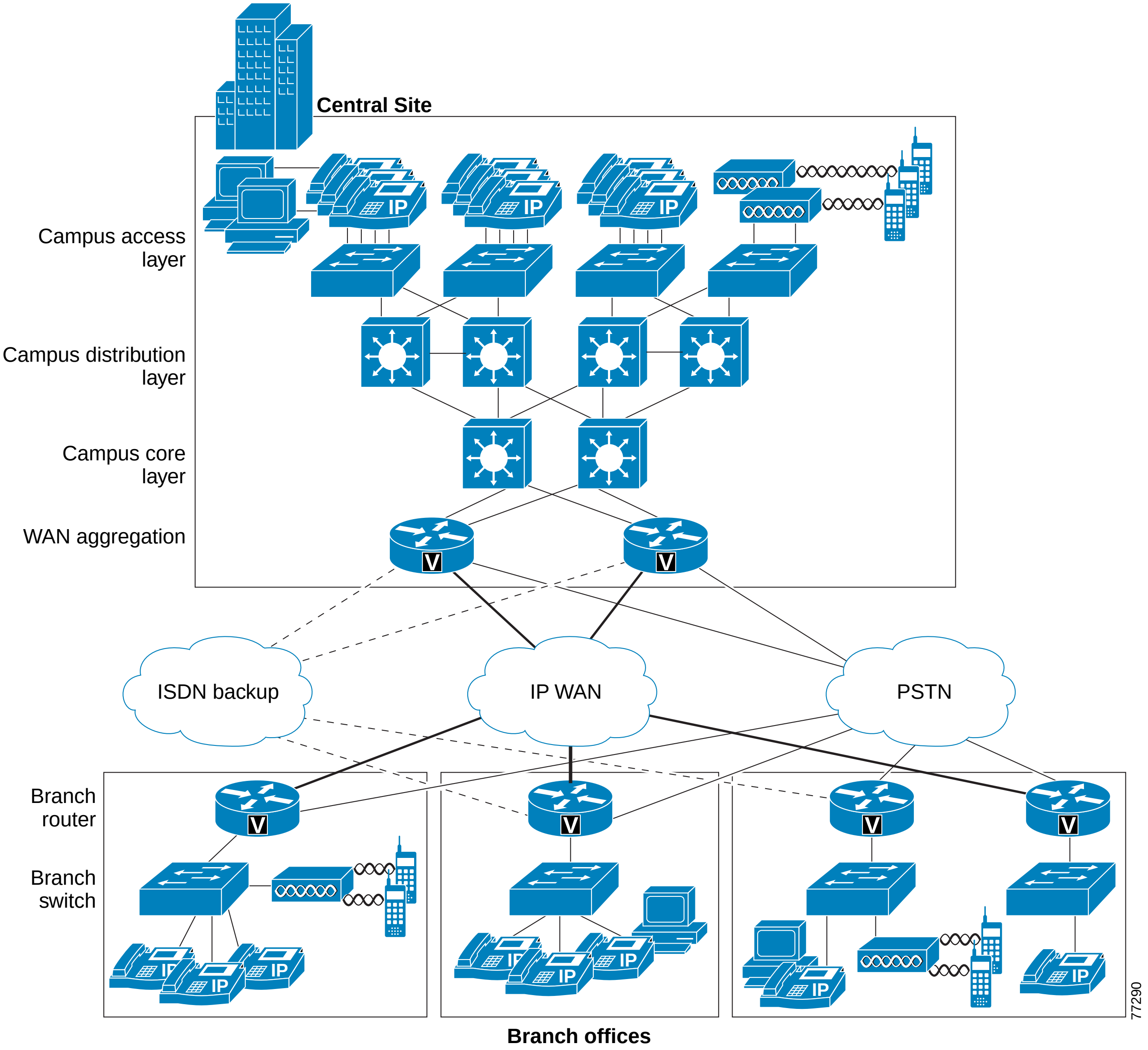


Table 3-1 Required Features for Each Role in the Network Infrastructure

Infrastructure Role	Required Features
Campus Access Switch	• In-Line Power[1] • Multiple Queue Support • 802.1p and 802.1Q • Fast Link Convergence
Campus Distribution or Core Switch	• Multiple Queue Support • 802.1p and 802.1Q • Traffic Classification • Traffic Reclassification
WAN Aggregation Router (Site that is at the hub of the network)	• Multiple Queue Support • Traffic Shaping • Link Fragmentation and Interleaving (LFI)[2] • Link Efficiency • Traffic Classification • Traffic Reclassification • 802.1p and 802.1Q
Branch Router (Spoke site)	• Multiple Queue Support • LFI[2] • Link Efficiency • Traffic Classification • Traffic Reclassification • 802.1p and 802.1Q
Branch or Smaller Site Switch	• In-Line Power[1] • Multiple Queue Support • 802.1p and 802.1Q

[1] Recommended.

[2] For link speeds less than 786 kbps.

What's New in This Chapter

Table 3-2 lists the topics that are new in this chapter or that have changed significantly from previous releases of this document.

Table 3-2 New or Changed Information Since the Previous Release of This Document

New or Revised Topic	Described in	Revision Date
Centralized TFTP	Centralized TFTP in a Mixed Environment, with Servers Running Different Releases of Cisco Unified CM, page 3-33	April 30, 2013
QoS design considerations for virtual Unified Communications	QoS Design Considerations for Virtual Unified Communications with Cisco UCS B-Series Blade Servers, page 3-19	September 28, 2012
Minor updates for wireless LAN infrastructure	Design Considerations for Voice and Video over WLAN, page 3-60	August 31, 2012
No changes for Cisco Unified Communications System Release 9.0		June 28, 2012

LAN Infrastructure

Campus LAN infrastructure design is extremely important for proper Unified Communications operation on a converged network. Proper LAN infrastructure design requires following basic configuration and design best practices for deploying a highly available network. Further, proper LAN infrastructure design requires deploying end-to-end QoS on the network. The following sections discuss these requirements:

- LAN Design for High Availability, page 3-4
- LAN Quality of Service (QoS), page 3-15

LAN Design for High Availability

Properly designing a LAN requires building a robust and redundant network from the top down. By structuring the LAN as a layered model (see Figure 3-1) and developing the LAN infrastructure one step of the model at a time, you can build a highly available, fault tolerant, and redundant network. Once these layers have been designed correctly, you can add network services such as DHCP and TFTP to provide additional network functionality. The following sections examine the infrastructure layers and network services:

- Campus Access Layer, page 3-5
- Campus Distribution Layer, page 3-10
- Campus Core Layer, page 3-12
- Network Services, page 3-22

For more information on campus design, refer to the *Design Zone for Campus* at

http://www.cisco.com/go/designzone

Campus Access Layer

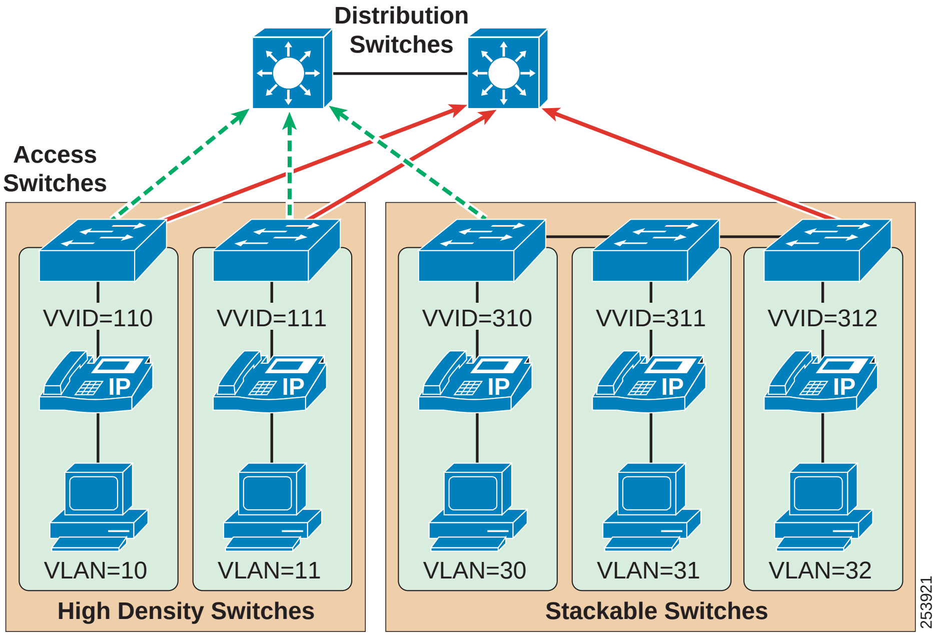
The access layer of the Campus LAN includes the portion of the network from the desktop port(s) to the wiring closet switch. Access layer switches have traditionally been configured as Layer 2 devices with Layer 2 uplinks to the distribution layer. The Layer 2 and spanning tree recommendations for Layer 2 access designs are well documented and are discussed briefly below. For newer Cisco Catalyst switches supporting Layer 3 protocols, new routed access designs are possible and offer improvements in convergence times and design simplicity. Routed access designs are discussed in the section on Routed Access Layer Designs, page 3-7.

Layer 2 Access Design Recommendations

Proper access layer design starts with assigning a single IP subnet per virtual LAN (VLAN). Typically, a VLAN should not span multiple wiring closet switches; that is, a VLAN should have presence in one and only one access layer switch (see Figure 3-2). This practice eliminates topological loops at Layer 2, thus avoiding temporary flow interruptions due to Spanning Tree convergence. However, with the introduction of standards-based IEEE 802.1w Rapid Spanning Tree Protocol (RSTP) and 802.1s Multiple Instance Spanning Tree Protocol (MISTP), Spanning Tree can converge at much higher rates. More importantly, confining a VLAN to a single access layer switch also serves to limit the size of the broadcast domain. There is the potential for large numbers of devices within a single VLAN or broadcast domain to generate large amounts of broadcast traffic periodically, which can be problematic. A good rule of thumb is to limit the number of devices per VLAN to about 512, which is equivalent to two Class C subnets (that is, a 23-bit subnet masked Class C address). For more information on the campus access layer, refer to the documentation on available at
http://www.cisco.com/en/US/products/hw/switches/index.html.

Note The recommendation to limit the number of devices in a single Unified Communications VLAN to approximately 512 is not solely due to the need to control the amount of VLAN broadcast traffic. Installing Unified CM in a VLAN with an IP subnet containing more than 1024 devices can cause the Unified CM server ARP cache to fill up quickly, which can seriously affect communications between the Unified CM server and other Unified Communications endpoints.

Figure 3-2 Access Layer Switches and VLANs for Voice and Data



When you deploy voice, Cisco recommends that you enable two VLANs at the access layer: a native VLAN for data traffic (VLANs 10, 11, 30, 31, and 32 in Figure 3-2) and a voice VLAN under Cisco IOS or Auxiliary VLAN under CatOS for voice traffic (represented by VVIDs 110, 111, 310, 311, and 312 in Figure 3-2).

Separate voice and data VLANs are recommended for the following reasons:

- Address space conservation and voice device protection from external networks

  Private addressing of phones on the voice or auxiliary VLAN ensures address conservation and ensures that phones are not accessible directly through public networks. PCs and servers are typically addressed with publicly routed subnet addresses; however, voice endpoints may be addressed using RFC 1918 private subnet addresses.

- QoS trust boundary extension to voice devices

  QoS trust boundaries can be extended to voice devices without extending these trust boundaries and, in turn, QoS features to PCs and other data devices.

- Protection from malicious network attacks

  VLAN access control, 802.1Q, and 802.1p tagging can provide protection for voice devices from malicious internal and external network attacks such as worms, denial of service (DoS) attacks, and attempts by data devices to gain access to priority queues through packet tagging.

- Ease of management and configuration

  Separate VLANs for voice and data devices at the access layer provide ease of management and simplified QoS configuration.

To provide high-quality voice and to take advantage of the full voice feature set, access layer switches should provide support for:

- 802.1Q trunking and 802.1p for proper treatment of Layer 2 CoS packet marking on ports with phones connected
- Multiple egress queues to provide priority queuing of RTP voice packet streams

- The ability to classify or reclassify traffic and establish a network trust boundary
- Inline power capability (Although inline power capability is not mandatory, it is highly recommended for the access layer switches.)
- Layer 3 awareness and the ability to implement QoS access control lists (These features are recommended if you are using certain Unified Communications endpoints such as a PC running a softphone application that cannot benefit from an extended trust boundary.)

Spanning Tree Protocol (STP)

To minimize convergence times and maximize fault tolerance at Layer 2, enable the following STP features:

- PortFast

  Enable PortFast on all access ports. The phones, PCs, or servers connected to these ports do not forward bridge protocol data units (BPDUs) that could affect STP operation. PortFast ensures that the phone or PC, when connected to the port, is able to begin receiving and transmitting traffic immediately without having to wait for STP to converge.

- Root guard or BPDU guard

  Enable root guard or BPDU guard on all access ports to prevent the introduction of a rogue switch that might attempt to become the Spanning Tree root, thereby causing STP re-convergence events and potentially interrupting network traffic flows. Ports that are set to **errdisable** state by BPDU guard must either be re-enabled manually or the switch must be configured to re-enable ports automatically from the errdisable state after a configured period of time.

- UplinkFast and BackboneFast

  Enable these features where appropriate to ensure that, when changes occur on the Layer 2 network, STP converges as rapidly as possible to provide high availability. When using Cisco stackable switches, enable Cross-Stack UplinkFast (CSUF) to provide fast failover and convergence if a switch in the stack fails.

- UniDirectional Link Detection (UDLD)

  Enable this feature to reduce convergence and downtime on the network when link failures or misbehaviors occur, thus ensuring minimal interruption of network service. UDLD detects, and takes out of service, links where traffic is flowing in only one direction. This feature prevents defective links from being mistakenly considered as part of the network topology by the Spanning Tree and routing protocols.

Note With the introduction of RSTP 802.1w, features such as PortFast and UplinkFast are not required because these mechanisms are built in to this standard. If RSTP has been enabled on the Catalyst switch, these commands are not necessary.

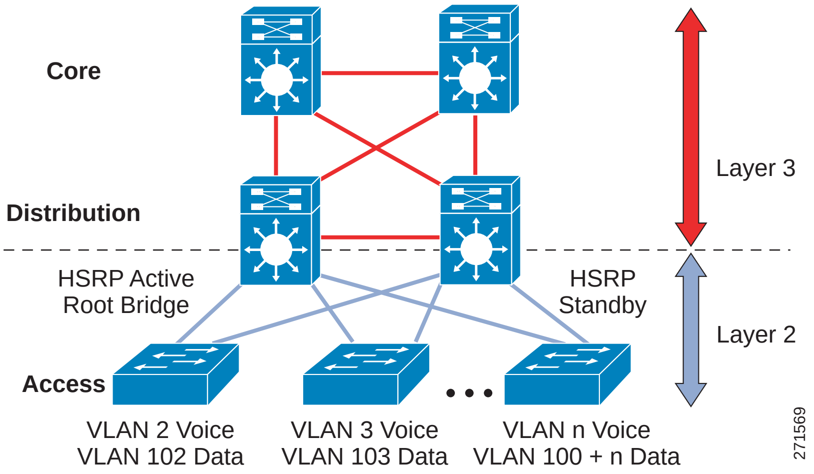
Routed Access Layer Designs

For campus designs requiring simplified configuration, common end-to-end troubleshooting tools, and the fastest convergence, a hierarchical design using Layer 3 switching in the access layer (routed access) in combination with Layer 3 switching at the distribution layer provides the fastest restoration of voice and data traffic flows.

Migrating the L2/L3 Boundary to the Access Layer

In the typical hierarchical campus design, the distribution layer uses a combination of Layer 2, Layer 3, and Layer 4 protocols and services to provide for optimal convergence, scalability, security, and manageability. In the most common distribution layer configurations, the access switch is configured as a Layer 2 switch that forwards traffic on high-speed trunk ports to the distribution switches. The distribution switches are configured to support both Layer 2 switching on their downstream access switch trunks and Layer 3 switching on their upstream ports toward the core of the network, as shown in Figure 3-3.

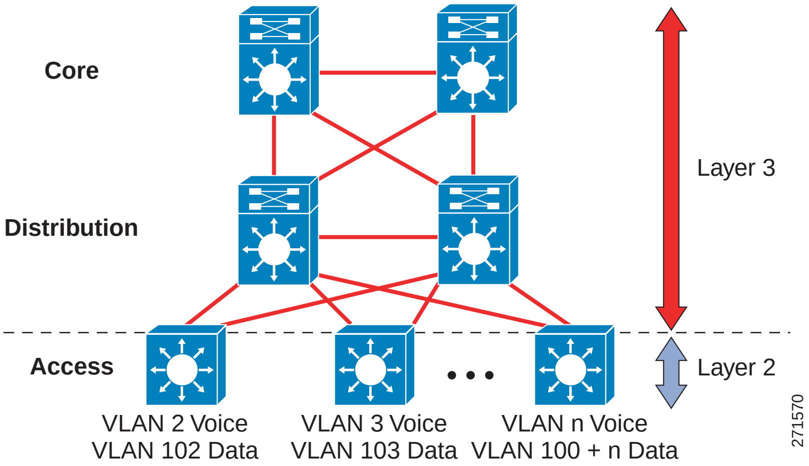
Figure 3-3 Traditional Campus Design — Layer 2 Access with Layer 3 Distribution



The purpose of the distribution switch in this design is to provide boundary functions between the bridged Layer 2 portion of the campus and the routed Layer 3 portion, including support for the default gateway, Layer 3 policy control, and all the multicast services required.

An alternative configuration to the traditional distribution layer model illustrated in Figure 3-3 is one in which the access switch acts as a full Layer 3 routing node (providing both Layer 2 and Layer 3 switching) and the access-to-distribution Layer 2 uplink trunks are replaced with Layer 3 point-to-point routed links. This alternative configuration, in which the Layer 2/3 demarcation is moved from the distribution switch to the access switch (as shown in Figure 3-4), appears to be a major change to the design but is actually just an extension of the current best-practice design.

Figure 3-4 Routed Access Campus Design — Layer 3 Access with Layer 3 Distribution



In both the traditional Layer 2 and the Layer 3 routed access designs, each access switch is configured with unique voice and data VLANs. In the Layer 3 design, the default gateway and root bridge for these VLANs is simply moved from the distribution switch to the access switch. Addressing for all end stations and for the default gateway remains the same. VLAN and specific port configurations remain unchanged on the access switch. Router interface configuration, access lists, "ip helper," and any other configuration for each VLAN remain identical but are configured on the VLAN Switched Virtual Interface (SVI) defined on the access switch instead of on the distribution switches.

There are several notable configuration changes associated with the move of the Layer 3 interface down to the access switch. It is no longer necessary to configure a Hot Standby Router Protocol (HSRP) or Gateway Load Balancing Protocol (GLBP) virtual gateway address as the "router" interfaces because all the VLANs are now local. Similarly, with a single multicast router, for each VLAN it is not necessary to perform any of the traditional multicast tuning such as tuning PIM query intervals or ensuring that the designated router is synchronized with the active HSRP gateway.

Routed Access Convergence

The many potential advantages of using a Layer 3 access design include the following:

- Improved convergence
- Simplified multicast configuration
- Dynamic traffic load balancing
- Single control plane
- Single set of troubleshooting tools (for example, ping and traceroute)

Of these advantages, perhaps the most significant is the improvement in network convergence times possible when using a routed access design configured with Enhanced Interior Gateway Routing Protocol (EIGRP) or Open Shortest Path First (OSPF) as the routing protocol. Comparing the convergence times for an optimal Layer 2 access design (either with a spanning tree loop or without a loop) against that of the Layer 3 access design, you can obtain a four-fold improvement in convergence times, from 800 to 900 msec for the Layer 2 design to less than 200 msec for the Layer 3 access design.

For more information on routed access designs, refer to the document on *High Availability Campus Network Design – Routed Access Layer using EIGRP or OSPF*, available at

http://www.cisco.com/application/pdf/en/us/guest/netsol/ns432/c649/ccmigration_09186a0080811468.pdf

Campus Distribution Layer

The distribution layer of the Campus LAN includes the portion of the network from the wiring closet switches to the next-hop switch. For more information on the campus distribution layer switches, refer to the product documentation available at

http://www.cisco.com/en/US/products/hw/switches/index.html

At the distribution layer, it is important to provide redundancy to ensure high availability, including redundant links between the distribution layer switches (or routers) and the access layer switches. To avoid creating topological loops at Layer 2, use Layer 3 links for the connections between redundant Distribution switches when possible.

First-Hop Redundancy Protocols

In the campus hierarchical model, where the distribution switches are the L2/L3 boundary, they also act as the default gateway for the entire L2 domain that they support. Some form of redundancy is required because this environment can be large and a considerable outage could occur if the device acting as the default gateway fails.

Gateway Load Balancing Protocol (GLBP), Hot Standby Router Protocol (HSRP), and Virtual Router Redundancy Protocol (VRRP) are all first-hop redundancy protocols. Cisco initially developed HSRP to address the need for default gateway redundancy. The Internet Engineering Task Force (IETF) subsequently ratified Virtual Router Redundancy Protocol (VRRP) as the standards-based method of providing default gateway redundancy. More recently, Cisco developed GLBP to overcome some the limitations inherent in both HSRP and VRRP.

HSRP and VRRP with Cisco enhancements both provide a robust method of backing up the default gateway, and they can provide failover in less than one second to the redundant distribution switch when tuned properly.

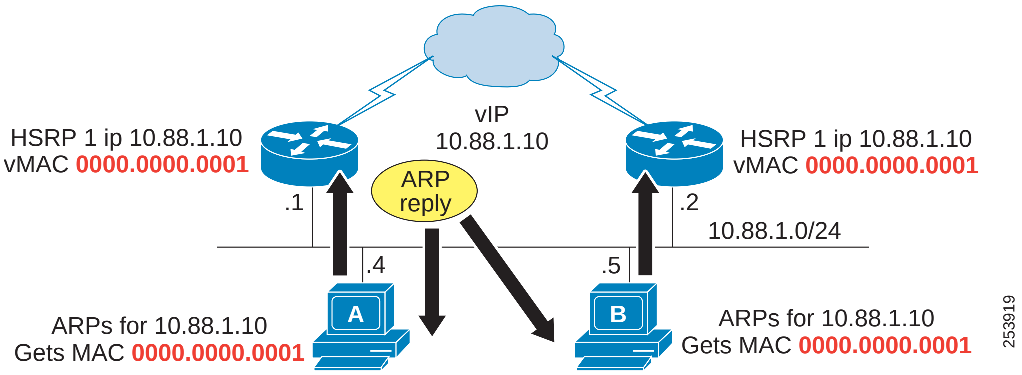
Gateway Load Balancing Protocol (GLBP)

Like HSRP and VRRP, Cisco's Gateway Load Balancing Protocol (GLBP) protects data traffic from a failed router or circuit, while also allowing packet load sharing between a group of redundant routers. When HSRP or VRRP are used to provide default gateway redundancy, the backup members of the peer relationship are idle, waiting for a failure event to occur for them to take over and actively forward traffic.

Before the development of GLBP, methods to utilize uplinks more efficiently were difficult to implement and manage. In one technique, the HSRP and STP/RSTP root alternated between distribution node peers, with the even VLANs homed on one peer and the odd VLANs homed on the alternate. Another technique used multiple HSRP groups on a single interface and used DHCP to alternate between the multiple default gateways. These techniques worked but were not optimal from a configuration, maintenance, or management perspective.

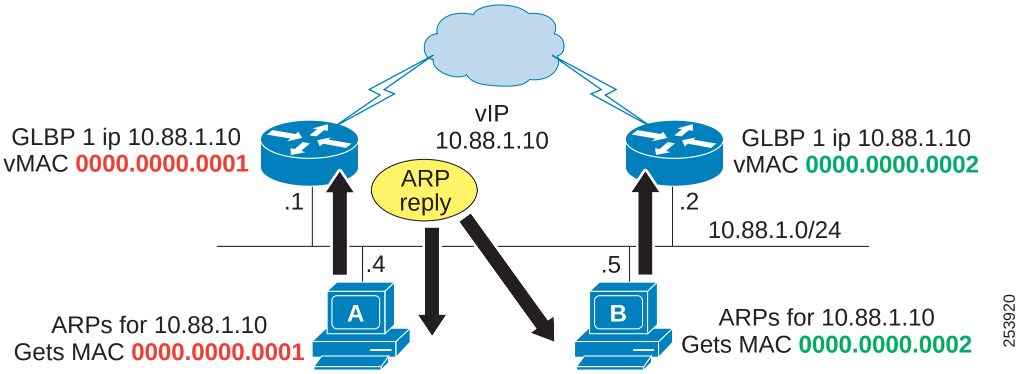
GLBP is configured and functions like HSRP. For HSRP, a single virtual MAC address is given to the endpoints when they use Address Resolution Protocol (ARP) to learn the physical MAC address of their default gateways (see Figure 3-5).

Figure 3-5 HSRP Uses One Virtual MAC Address



Two virtual MAC addresses exist with GLBP, one for each GLBP peer (see Figure 3-6). When an endpoint uses ARP to determine its default gateway, the virtual MAC addresses are checked in a round-robin basis. Failover and convergence work just like with HSRP. The backup peer assumes the virtual MAC address of the device that has failed, and begins forwarding traffic for its failed peer.

Figure 3-6 GLBP Uses Two Virtual MAC Addresses, One for Each GLBP Peer



The end result is that a more equal utilization of the uplinks is achieved with minimal configuration. As a side effect, a convergence event on the uplink or on the primary distribution node affects only half as many hosts, giving a convergence event an average of 50 percent less impact.

For more information on HSRP, VRRP, and GLBP, refer to the *Campus Network for High Availability Design Guide*, available at

> http://www.cisco.com/application/pdf/en/us/guest/netsol/ns431/c649/ccmigration_09186a008093b876.pdf

Routing Protocols

Configure Layer 3 routing protocols such as OSPF and EIGRP at the distribution layer to ensure fast convergence, load balancing, and fault tolerance. Use parameters such as routing protocol timers, path or link costs, and address summaries to optimize and control convergence times as well as to distribute traffic across multiple paths and devices. Cisco also recommends using the **passive-interface** command to prevent routing neighbor adjacencies via the access layer. These adjacencies are typically unnecessary, and they create extra CPU overhead and increased memory utilization because the routing protocol keeps track of them. By using the **passive-interface** command on all interfaces facing the access layer, you prevent routing updates from being sent out on these interfaces and, therefore, neighbor adjacencies are not formed.

Campus Core Layer

The core layer of the Campus LAN includes the portion of the network from the distribution routers or Layer 3 switches to one or more high-end core Layer 3 switches or routers. Layer 3-capable Catalyst switches at the core layer can provide connectivity between numerous campus distribution layers. For more details on the campus core layer switches, refer to the documentation on available at http://www.cisco.com/en/US/products/hw/switches/index.html.

At the core layer, it is again very important to provide the following types of redundancy to ensure high availability:

- Redundant link or cable paths

  Redundancy here ensures that traffic can be rerouted around downed or malfunctioning links.

- Redundant devices

  Redundancy here ensures that, in the event of a device failure, another device in the network can continue performing tasks that the failed device was doing.

- Redundant device sub-systems

  This type of redundancy ensures that multiple power supplies and modules are available within a device so that the device can continue to function in the event that one of these components fails.

The Cisco Catalyst Virtual Switching System (VSS) is a method to ensure redundancy in all of these areas by pooling together two Catalyst supervisor engines to act as one. For more information regarding VSS, refer to the product documentation available at

http://www.cisco.com/en/US/products/ps9336/index.html

Routing protocols at the core layer should again be configured and optimized for path redundancy and fast convergence. There should be no STP in the core because network connectivity should be routed at Layer 3. Finally, each link between the core and distribution devices should belong to its own VLAN or subnet and be configured using a 30-bit subnet mask.

Data Center and Server Farm

Typically, Cisco Unified Communications Manager (Unified CM) cluster servers, including media resource servers, reside in a firewall-secured data center or server farm environment. In addition, centralized gateways and centralized hardware media resources such as conference bridges, DSP or transcoder farms, and media termination points may be located in the data center or server farm. The placement of firewalls in relation to Cisco Unified Communications Manager (Unified CM) cluster servers and media resources can affect how you design and implement security in your network. For design guidance on firewall placement in relation to Unified Communications systems and media resources, see Firewalls, page 4-22.

Because these servers and resources are critical to voice networks, Cisco recommends distributing all Unified CM cluster servers, centralized voice gateways, and centralized hardware resources between multiple physical switches and, if possible, multiple physical locations within the campus. This distribution of resources ensures that, given a hardware failure (such as a switch or switch line card failure), at least some servers in the cluster will still be available to provide telephony services. In addition, some gateways and hardware resources will still be available to provide access to the PSTN and to provide auxiliary services. Besides being physically distributed, these servers, gateways, and hardware resources should be distributed among separate VLANs or subnets so that, if a broadcast storm or denial of service attack occurs on a particular VLAN, not all voice connectivity and services will be disrupted.

Power over Ethernet (PoE)

PoE (or inline power) is 48 Volt DC power provided over standard Ethernet unshielded twisted-pair (UTP) cable. Instead of using wall power, IP phones and other inline powered devices (PDs) such as the Aironet Wireless Access Points can receive power provided by inline power-capable Catalyst Ethernet switches or other inline power source equipment (PSE). Inline power is enabled by default on all inline power-capable Catalyst switches.

Deploying inline power-capable switches with uninterruptible power supplies (UPS) ensures that IP phones continue to receive power during power failure situations. Provided the rest of the telephony network is available during these periods of power failure, then IP phones should be able to continue making and receiving calls. You should deploy inline power-capable switches at the campus access layer within wiring closets to provide inline-powered Ethernet ports for IP phones, thus eliminating the need for wall power.

⚠ **Caution** The use of power injectors or power patch panels to deliver PoE can damage some devices because power is always applied to the Ethernet pairs. PoE switch ports automatically detect the presence of a device that requires PoE before enabling it on a port-by-port basis.

In addition to Cisco PoE inline power, Cisco now supports the IEEE 802.3af PoE standard. The majority of Cisco switches and Cisco Unified IP Phones comply with the 802.3af standard. For information about which Cisco Unified IP Phones support the 802.3af PoE standard, refer to the product documentation for your particular phone models (available at http://www.cisco.com).

Energy Conservation for IP Phones

Cisco EnergyWise Technology provides intelligent management of energy usage for devices on the IP network, including Unified Communications endpoints that use Power over Ethernet (PoE). Cisco EnergyWise architecture can turn power on and off to devices connected with PoE on EnergyWise enabled switches, based on a configurable schedule. For more information on EnergyWise, refer to the documentation at

http://www.cisco.com/en/US/products/ps10195/index.html

When the PoE switch powers off IP phones for EnergyWise conservation, the phones are completely powered down. EnergyWise shuts down inline power on the ports that connect to IP phones and does so by a schedule or by commands from network management tools. When power is disabled, no verification occurs to determine whether a phone has an active call. The power is turned off and any active call is torn down. The IP phone loses registration from Cisco Unified Communications Manager and no calls can be made to or from the phone. There is no mechanism on the phone to power it on, therefore emergency calling will not be available on that phone.

The IP phone can be restarted only when the switch powers it on again. After power is restored, the IP phones will reboot and undergo a recovery process that includes requesting a new IP address, downloading a configuration file, applying any new configuration parameters, downloading new firmware or locales, and registering with Cisco Unified CM.

The EnergyWise schedule is configured and managed on the Cisco Network Infrastructure. It does not require any configuration on the IP phone or on Cisco Unified CM. However, power consumption on the phone can also be managed by a device profile configured on Unified CM. The energy saving options provided by Unified CM include the following:

- Power Save Plus Mode, page 3-14
- Power Save Mode, page 3-14

Power Save Plus Mode

In Power Save Plus mode, the phone on and off times and the idle timeout periods can be configured on the IP phones. The Cisco IP Phones' EnergyWise Power Save Plus configuration options specify the schedule for the IP phones to sleep (power down) and wake (power up). This mode requires an EnergyWise enabled network. If EnergyWise is enabled, then the sleep and wake times, as well as other parameters, can be used to control power to the phones. The Power Save Plus parameters are configured in the product-specific device profile in Cisco Unified CM Administration and sent to the IP phones as part of the phone configuration XML file.

During the configured power off period in this power saving mode, the IP phone sends a request to the switch asking for a wake-up at a specified time. If the switch is EnergyWise enabled, it accepts the request and reduces the power to the phone port, putting the phone to sleep. The sleep mode reduces the power consumption of the phone to 1 watt or less. The phone is not completely powered off in this case. When the phone is sleeping, the PoE switch provides minimal power that illuminates the Select key on the phone. A user can wake up the IP phone by using the Select button. The IP phone does not go into sleep mode if a call is active on the phone. Audio and visual alerts can optionally be configured to warn users before a phone enters the Power Save Plus mode. While the phone is in sleep mode, it is not registered to Cisco Unified CM and cannot receive any inbound calls. Use the Forward Unregistered setting in the phone's device configuration profile to specify how to treat any inbound calls to the phone's number.

Note The Cisco EnergyWise Power Save Plus mode is supported in Unified CM 8.6 and later releases, and it requires phone firmware version 9.(2)1 or later. It is available on the Cisco Unified IP Phone 6900, 8900, and 9900 Series.

Power Save Mode

In Power Save mode, the backlight on the screen is not lit when the phone is not in use. The phone stays registered to Cisco Unified CM in this mode and can receive inbound calls and make outbound calls. Cisco Unified CM Administration has product-specific configuration options to turn off the display at a designated time on some days and all day on other days. The phone remains in Power Save mode for the scheduled duration or until the user lifts the handset or presses any button. An EnergyWise enabled network is not required for the Power Save mode. Idle times can be scheduled so that the display remains on until the timeout and then turns off automatically. The phone is still powered on in this mode and can receive inbound calls.

The Power Save mode can be used together with the Power Save Plus mode. Using both significantly reduces the total power consumption by Cisco Unified IP Phones.

For information on configuring these modes, refer to the administration guides for the Cisco Unified IP Phones, available at the following locations:

- Cisco Unified IP Phones 9900 Series

  http://www.cisco.com/en/US/products/ps10453/prod_maintenance_guides_list.html
- Cisco Unified IP Phones 8900 Series

  http://www.cisco.com/en/US/products/ps10451/prod_maintenance_guides_list.html
- Cisco Unified IP Phones 6900 Series

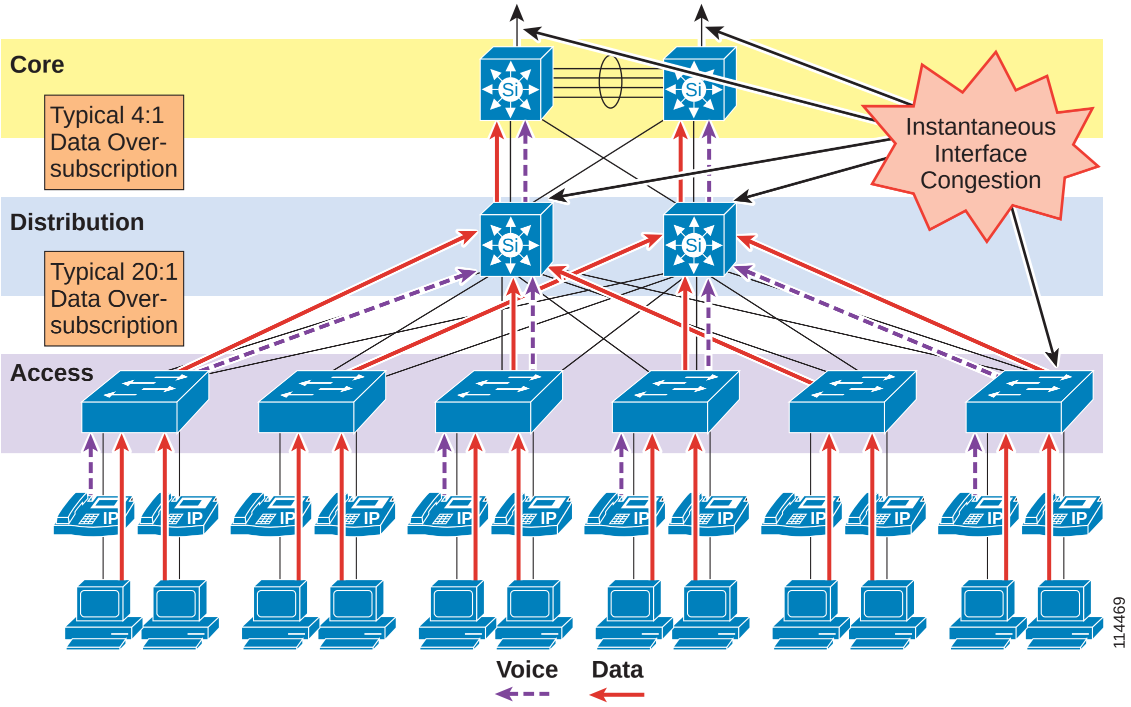
  http://www.cisco.com/en/US/products/ps10326/prod_maintenance_guides_list.html

LAN Quality of Service (QoS)

Until recently, quality of service was not an issue in the enterprise campus due to the asynchronous nature of data traffic and the ability of network devices to tolerate buffer overflow and packet loss. However, with new applications such as voice and video, which are sensitive to packet loss and delay, buffers and not bandwidth are the key QoS issue in the enterprise campus.

Figure 3-7 illustrates the typical oversubscription that occurs in LAN infrastructures.

Figure 3-7 Data Traffic Oversubscription in the LAN


Core
Typical 4:1 Data Over-subscription
Distribution
Typical 20:1 Data Over-subscription
Access
Instantaneous Interface Congestion
Si
IP
Voice
Data
114469

This oversubscription, coupled with individual traffic volumes and the cumulative effects of multiple independent traffic sources, can result in the egress interface buffers becoming full instantaneously, thus causing additional packets to drop when they attempt to enter the egress buffer. The fact that campus switches use hardware-based buffers, which compared to the interface speed are much smaller than those found on WAN interfaces in routers, merely increases the potential for even short-lived traffic bursts to cause buffer overflow and dropped packets.

Applications such as file sharing (both peer-to-peer and server-based), remote networked storage, network-based backup software, and emails with large attachments, can create conditions where network congestion occurs more frequently and/or for longer durations. Some of the negative effects of recent worm attacks have been an overwhelming volume of network traffic (both unicast and broadcast-storm based), increasing network congestion. If no buffer management policy is in place, loss, delay, and jitter performance of the LAN may be affected for all traffic.

Another situation to consider is the effect of failures of redundant network elements, which cause topology changes. For example, if a distribution switch fails, all traffic flows will be reestablished through the remaining distribution switch. Prior to the failure, the load balancing design shared the load between two switches, but after the failure all flows are concentrated in a single switch, potentially causing egress buffer conditions that normally would not be present.

For applications such as voice, this packet loss and delay results in severe voice quality degradation. Therefore, QoS tools are required to manage these buffers and to minimize packet loss, delay, and delay variation (jitter).

The following types of QoS tools are needed from end to end on the network to manage traffic and ensure voice quality:

- Traffic classification

  Classification involves the marking of packets with a specific priority denoting a requirement for class of service (CoS) from the network. The point at which these packet markings are trusted or not trusted is considered the trust boundary. Trust is typically extended to voice devices (phones) and not to data devices (PCs).

- Queuing or scheduling

  Interface queuing or scheduling involves assigning packets to one of several queues based on classification for expedited treatment throughout the network.

- Bandwidth provisioning

  Provisioning involves accurately calculating the required bandwidth for all applications plus element overhead.

The following sections discuss the use of these QoS mechanisms in a campus environment:

- Traffic Classification, page 3-16
- Interface Queuing, page 3-18
- Bandwidth Provisioning, page 3-19
- Impairments to IP Communications if QoS is Not Employed, page 3-19

Traffic Classification

It has always been an integral part of the Cisco network design architecture to classify or mark traffic as close to the edge of the network as possible. Traffic classification is an entrance criterion for access into the various queuing schemes used within the campus switches and WAN interfaces. Cisco IP Phones mark voice control signaling and voice RTP streams at the source, and they adhere to the values presented in Table 3-3. As such, the IP phone can and should classify traffic flows.

Table 3-3 lists the traffic classification requirements for the LAN infrastructure.

Table 3-3 Traffic Classification Guidelines for Various Types of Network Traffic

	Layer-3 Classification			**Layer-2 Classification**
Application	**Type of Service (ToS) IP Precedence (IPP)**	**Per-Hop Behavior (PHB)**	**Differentiated Services Code Point (DSCP)**	**Class of Service (CoS)**
Routing	6	CS6	48	6
Voice Real-Time Transport Protocol (RTP)	5	EF	46	5
Videoconferencing	4	AF41	34	4
Streaming video	4	CS4	32	4
Call signaling[1]	3	CS3 (currently) AF31 (previously)	24 (currently) 26 (previously)	3
Transactional data	2	AF21	18	2
Network management	2	CS2	16	2
Scavenger	1	CS1	8	1
Best effort	0	0	0	0

[1] The recommended DSCP/PHB marking for call control signaling traffic has been changed from 26/AF31 to 24/CS3. A marking migration has occurred within Cisco to reflect this change, however some products still mark signaling traffic as 26/AF31. Therefore, in the interim, Cisco recommends that both AF31 and CS3 be reserved for call signaling.

For more information about traffic classification, refer to the *Enterprise QoS Solution Reference Network Design (SRND)*, available at

http://www.cisco.com/go/designzone

Traffic Classification for Video Telephony

The main classes of interest for IP Video Telephony are:

- Voice

  Voice is classified as CoS 5 (IP Precedence 5, PHB EF, or DSCP 46).

- Videoconferencing

  Videoconferencing is classified as CoS 4 (IP Precedence 4, PHB AF41, or DSCP 34).

- Call signaling

  Call signaling for voice and videoconferencing is now classified as CoS 3 (IP Precedence 3, PHB CS3, or DSCP 24) but was previously classified as PHB AF31 or DSCP 26.

Cisco highly recommends these classifications as *best practices* in a Cisco Unified Communications network.

QoS Marking Differences Between Video Calls and Voice-Only Calls

The voice component of a call can be classified in one of two ways, depending on the type of call in progress. A voice-only telephone call would have its media classified as CoS 5 (IP Precedence 5 or PHB EF), while the voice channel of a video conference would have its media classified as CoS 4 (IP Precedence 4 or PHB AF41). All the Cisco IP Video Telephony products adhere to the Cisco

Corporate QoS Baseline standard, which requires that the audio and video channels of a video call both be marked as CoS 4 (IP Precedence 4 or PHB AF41). The reasons for this recommendation include, but are not limited to, the following:

- To preserve lip-sync between the audio and video channels
- To provide separate classes for audio-only calls and video calls

The signaling class is applicable to all voice signaling protocols (such as SCCP, MGCP, and so on) as well as video signaling protocols (such as SCCP, H.225, RAS, CAST, and so on).

Given the recommended classes, the first step is to decide where the packets will be classified (that is, which device will be the first to mark the traffic with its QoS classification). There are essentially two places to mark or classify traffic:

- On the originating endpoint — the classification is then trusted by the upstream switches and routers
- On the switches and/or routers — because the endpoint is either not capable of classifying its own packets or is not trustworthy to classify them correctly

QoS Enforcement Using a Trusted Relay Point (TRP)

A Trusted Relay Point (TRP) can be used to enforce and/or re-mark the DSCP values of media flows from endpoints. This feature allows QoS to be enforced for media from endpoints such as softphones, where the media QoS values might have been modified locally.

A TRP is a media resource based upon the existing Cisco IOS media termination point (MTP) function.

Endpoints can be configured to "Use Trusted Relay Point," which will invoke a TRP for all calls.

For QoS enforcement, the TRP uses the configured QoS values for media in Unified CM's Service Parameters to re-mark and enforce the QoS values in media streams from the endpoint.

TRP functionality is supported by Cisco IOS MTPs and transcoding resources. (Use Unified CM to check "Enable TRP" on the MTP or transcoding resource to activate TRP functionality.)

Interface Queuing

After packets have been marked with the appropriate tag at Layer 2 (CoS) and Layer 3 (DSCP or PHB), it is important to configure the network to schedule or queue traffic based on this classification, so as to provide each class of traffic with the service it needs from the network. By enabling QoS on campus switches, you can configure all voice traffic to use separate queues, thus virtually eliminating the possibility of dropped voice packets when an interface buffer fills instantaneously.

Although network management tools may show that the campus network is not congested, QoS tools are still required to guarantee voice quality. Network management tools show only the average congestion over a sample time span. While useful, this average does not show the congestion peaks on a campus interface.

Transmit interface buffers within a campus tend to congest in small, finite intervals as a result of the bursty nature of network traffic. When this congestion occurs, any packets destined for that transmit interface are dropped. The only way to prevent dropped voice traffic is to configure multiple queues on campus switches. For this reason, Cisco recommends always using a switch that has at least two output queues on each port and the ability to send packets to these queues based on QoS Layer 2 and/or Layer 3 classification. The majority of Cisco Catalyst Switches support two or more output queues per port. For more information on Cisco Catalyst Switch interface queuing capabilities, refer to the documentation at http://www.cisco.com/en/US/products/hw/switches/index.html

Bandwidth Provisioning

In the campus LAN, bandwidth provisioning recommendations can be summarized by the motto, *Over provision and under subscribe*. This motto implies careful planning of the LAN infrastructure so that the available bandwidth is always considerably higher than the load and there is no steady-state congestion over the LAN links.

The addition of voice traffic onto a converged network does not represent a significant increase in overall network traffic load; the bandwidth provisioning is still driven by the demands of the data traffic requirements. The design goal is to avoid extensive data traffic congestion on any link that will be traversed by telephony signaling or media flows. Contrasting the bandwidth requirements of a single G.711 voice call (approximately 86 kbps) to the raw bandwidth of a FastEthernet link (100 Mbps) indicates that voice is not a source of traffic that causes network congestion in the LAN, but rather it is a traffic flow to be protected from LAN network congestion.

Impairments to IP Communications if QoS is Not Employed

If QoS is not deployed, packet drops and excessive delay and jitter can occur, leading to impairments of the telephony services. When media packets are subjected to drops, delay, and jitter, the user-perceivable effects include clicking sound, harsh-sounding voice, extended periods of silence, and echo.

When signaling packets are subjected to the same conditions, user-perceivable impairments include unresponsiveness to user input (such as delay to dial tone), continued ringing upon answer, and double dialing of digits due to the user's belief that the first attempt was not effective (thus requiring hang-up and redial). More extreme cases can include endpoint re-initialization, call termination, and the spurious activation of SRST functionality at branch offices (leading to interruption of gateway calls).

These effects apply to all deployment models. However, single-site (campus) deployments tend to be less likely to experience the conditions caused by sustained link interruptions because the larger quantity of bandwidth typically deployed in LAN environments (minimum links of 100 Mbps) allows for some residual bandwidth to be available for the IP Communications system.

In any WAN-based deployment model, traffic congestion is more likely to produce sustained and/or more frequent link interruptions because the available bandwidth is much less than in a LAN (typically less than 2 Mbps), so the link is more easily saturated. The effects of link interruptions can impact the user experience, whether or not the voice media traverses the packet network, because signaling traffic between endpoints and the Unified CM servers can also be delayed or dropped.

QoS Design Considerations for Virtual Unified Communications with Cisco UCS B-Series Blade Servers

With a virtualized Unified Communications solution, Cisco Unified Communications products can run as virtual machines on a select set of supported hypervisor, server, and storage products. The most important component in a virtual Unified Communications solution is the Cisco Unified Computing System (UCS) Platform along with hypervisor virtualization technology. Virtualized Unified Communications designs have specific considerations with respect to QoS, as discussed below. For more information on the Cisco Unified Computing System (UCS) architecture, hypervisor technology for application virtualization, and Storage Area Networking (SAN) concepts, see Deploying Unified Communications on Virtualized Servers, page 5-46.

In a virtualized environment, Unified Communications applications such as Cisco Unified Communications Manager (Unified CM) run as virtual machines on top of the VMware Hypervisor. These Unified Communications virtual machines are connected to a virtual software switch rather than a hardware-based Ethernet switch for Media Convergence Server (MCS) deployments. The following types of virtual software switches are available:

- VMware vSphere Standard Switch

  Available with all VMware vSphere editions and independent of the type of VMware licensing scheme. The vSphere Standard Switch exists only on the host on which it is configured.

- VMware vSphere Distributed Switch

  Available only with the Enterprise Plus Edition of VMware vSphere. The vSphere Distributed Switch acts as a single switch across all associated hosts on a datacenter and helps simplify manageability of the software virtual switch.

- Cisco Nexus 1000V Switch

  Cisco has a software switch called the Nexus 1000 Virtual (1000V) Switch. The Cisco Nexus 1000V requires the Enterprise Plus Edition of VMware vSphere. It is a distributed virtual switch visible to multiple VMware hosts and virtual machines. The Cisco Nexus 1000V Series provides policy-based virtual machine connectivity, mobile virtual machine security, enhanced QoS, and network policy.

From the virtual connectivity point of view, each virtual machine can connect to any one of the above virtual switches residing on a blade server. The blade servers physically connect to the rest of the network via a Fabric Extender in the UCS chassis to a UCS Fabric Interconnect Switch (for example, Cisco UCS 6100 or 6200 Series). The UCS Fabric Interconnect Switch is where the physical wiring connects to a customer's 1 Gb or 10 Gb Ethernet LAN and FC SAN.

From the traffic flow point of view, traffic from the virtual machines first goes to the software virtual switch (for example, vSphere Standard Switch, vSphere Distributed Switch, or Cisco Nexus 1000V Switch). The virtual switch then sends the traffic to the physical UCS Fabric Interconnect Switch (UCS 6100 or 6200 Series) through its blade server's Network Adapter and Fabric Extender. The UCS Fabric Interconnect Switch carries both the IP and fibre channel SAN traffic via Fibre Channel over Ethernet (FCoE) on a single wire. The UCS Fabric Interconnect Switch sends IP traffic to an IP switch (for example, Cisco Catalyst or Nexus Series Switch), and it sends SAN traffic to a Fibre Channel SAN Switch (for example, Cisco MDS Series Switch).

Standard Switching Element QoS Behavior

By default within the UCS 6100 or 6200 Series Fabric Interconnect Switch, a priority QoS class is automatically created for all fibre channel (FC) traffic destined to the SAN switch. This FC QoS class has no drop policy, and all the FC traffic is marked with Layer 2 CoS value of 3. By default all other traffic (Ethernet and IP), including voice signaling and media traffic, falls into Best Effort QoS class.

The vSphere Standard Switch, vSphere Distributed Switch, and UCS 6100 or 6200 Series switches cannot map L3 DSCP values to L2 CoS values. Traffic can be prioritized or de-prioritize inside the UCS 6100 and 6200 Series Switches based on L2 CoS only.

Note Unified Communications applications mark the L3 DSCP values only (for instance, CS3 for voice signaling). It is possible to mark traffic with an L2 CoS value through UCS Manager, but all traffic originating from a virtual machine network adapter would be marked with the same L2 CoS value if the Nexus 1000V is not used.

The Nexus 1000V software switch has the ability to map L3 DSCP values to L2 CoS values, and vice versa, like traditional Cisco physical switches such as the Catalyst Series Switches. Therefore, when Unified Communications traffic leaves a virtual machine and enters the Nexus 1000V switch, its L3 DSCP values can be mapped to corresponding L2 CoS values. This traffic can then be prioritized or de-prioritized based on the L2 CoS value inside the UCS 6100 Switch.

For instance, voice signaling traffic with L3 DSCP value of CS3 is mapped to L2 CoS value of 3 by Nexus 1000V. By default, all Fibre Channel over Ethernet (FCoE) traffic is marked with L2 CoS value of 3 by Cisco UCS. When voice signaling and FCoE traffic enter the Cisco UCS 6100 Fabric Interconnect Switch, both will carry a CoS value of 3. In this situation voice signaling traffic will share queues and scheduling with the Fibre Channel priority class and will be given lossless behavior. (Fibre Channel priority class for CoS 3 in the UCS Fabric Interconnect Switch does not imply that the class cannot be shared with other types of traffic.)

The L2 CoS value for FCoE traffic can be changed from its default value of 3 to another value, and CoS 3 can be reserved exclusively for the voice signaling traffic. However, Cisco does not suggest or recommend this approach because some Converged Network Adapters (CNAs) cause problems when the FCoE CoS value is not set to a value of 3.

Congestion Scenario

In the physical server design, the hard drives are locally attached to the MCS server, and the SCSI traffic never competes with the Ethernet IP traffic.

Virtual Unified Communications designs with UCS B-Series Systems are different than traditional MCS-based designs. In a virtual Unified Communications design, because the hard drive is remote and accessed via the FC SAN, there is a potential for FC SAN traffic to compete for bandwidth with the Ethernet IP traffic inside the UCS Fabric Interconnect Switch. This could result in voice-related IP traffic (signaling and media) being dropped because FC traffic has a no-drop policy inside the UCS Fabric Interconnect Switch. This congestion or oversubscription scenario is highly unlikely, however, because the UCS Fabric Interconnect Switch provides a high-capacity switching fabric, and the usable bandwidth per server blade far exceeds the maximum traffic requirements of a typical Unified Communications application.

Design Recommendations

The Nexus 1000V provides enhanced QoS and other features (for example, ACLs, DHCP snooping, IP Source Guard, SPAN, and so forth) that are essential for virtualized data centers and are not available in the other virtual switch implementations. With its capability to map L3 DSCP values to L2 CoS values, the Nexus 1000V switch is recommended for large data center implementations where Cisco Unified Communications Applications are deployed with many other virtual machines running on UCS B-Series system. For other Unified Communications deployments, the decision to use the Nexus 1000V will vary on a case-by-case basis, depending on the available bandwidth for Unified Communications Applications within the UCS architecture. If there is a possibility that a congestion scenario will arise, then the Nexus 1000V switch should be deployed.

An example of an alternative solution that can also be deployed on all virtual switches is to configure all physical Network Adapters on the Unified Communications server blades to set a QoS policy of **Platinum** (CoS=5; No Drop Policy) for all traffic. Any other application running on the same UCS system or chassis should set the QoS policy to **best effort**. The downside to this approach is that all traffic types from virtual Unified Communications applications will have their CoS value set to Platinum, including all non-voice traffic (for example, backups, CDRs, logs, Web traffic, and so forth). Although this solution is not optimal, it does raise the priority of Unified Communications application traffic to that of FC SAN-destined traffic, thus reducing the possibility of traffic drops.

Network Services

The deployment of an IP Communications system requires the coordinated design of a well structured, highly available, and resilient network infrastructure as well as an integrated set of network services including Domain Name System (DNS), Dynamic Host Configuration Protocol (DHCP), Trivial File Transfer Protocol (TFTP), and Network Time Protocol (NTP).

Domain Name System (DNS)

DNS enables the mapping of host names and network services to IP addresses within a network or networks. DNS server(s) deployed within a network provide a database that maps network services to hostnames and, in turn, hostnames to IP addresses. Devices on the network can query the DNS server and receive IP addresses for other devices in the network, thereby facilitating communication between network devices.

Complete reliance on a single network service such as DNS can introduce an element of risk when a critical Unified Communications system is deployed. If the DNS server becomes unavailable and a network device is relying on that server to provide a hostname-to-IP-address mapping, communication can and will fail. For this reason, in networks requiring high availability, Cisco recommends that you do not rely on DNS name resolution for any communications between Unified CM and the Unified Communications endpoints.

For standard deployments, Cisco recommends that you configure Unified CM(s), gateways, and endpoint devices to use IP addresses rather than hostnames. For endpoint devices, Cisco does not recommend configuration of DNS parameters such as DNS server addresses, hostnames, and domain names. During the initial installation of the publisher node in a Unified CM cluster, the publisher will be referenced in the server table by the hostname you provided for the system. Before installation and configuration of any subsequent subscribers or the definition of any endpoints, you should change this server entry to the IP address of the publisher rather than the hostname. Each subscriber added to the cluster should be defined in this same server table via IP address and not by hostname. Each subscriber should be added to this server table one device at a time, and there should be no definitions for non-existent subscribers at any time other than for the new subscriber being installed.

During installation of the publisher and subscriber, Cisco recommend that you do not select the option to enable DNS unless DNS is specifically required for system management purposes. If DNS is enabled, Cisco still highly recommend that you do not use DNS names in the configuration of the IP Communications endpoints, gateways, and Unified CM servers. Even if DNS is enabled on the servers in the cluster, it is never used for any intra-cluster server-to-server communications and is used only for communications to devices external to the cluster itself.

Deploying Unified CM with DNS

There are some situations in which configuring and using DNS might be unavoidable. For example, if Network Address Translation (NAT) is required for communications between the IP phones and Unified CM in the IP Communications network, DNS is required to ensure proper mapping of NAT translated addresses to network host devices. Likewise, some IP telephony disaster recovery network configurations rely on DNS to ensure proper failover of the network during failure scenarios by mapping hostnames to secondary backup site IP addresses.

If either of these two situations exists and DNS must be configured, you must deploy DNS servers in a geographically redundant fashion so that a single DNS server failure will not prevent network communications between IP telephony devices. By providing DNS server redundancy in the event of a single DNS server failure, you ensure that devices relying on DNS to communicate on the network can still receive hostname-to-IP-address mappings from a backup or secondary DNS server.

Unified CM can use DNS to:

- Provide simplified system management
- Resolve fully qualified domain names to IP addresses for trunk destinations
- Resolve fully qualified domain names to IP addresses for SIP route patterns based on domain name
- Resolve service (SRV) records to host names and then to IP addresses for SIP trunk destinations

When DNS is used, Cisco recommends defining each Unified CM cluster as a member of a valid sub-domain within the larger organizational DNS domain, defining the DNS domain on each Cisco MCS server, and defining the primary and secondary DNS server addresses on each MCS server.

Table 3-4 shows an example of how DNS server could use A records (Hostname-to-IP-address resolution), Cname records (aliases), and SRV records (service records for redundancy and load balancing) in a Unified CM environment.

Table 3-4 Example Use of DNS with Unified CM

Host Name	Type	TTL	Data
CUCM-Admin.cluster1.cisco.com	Host (A)	12 Hours	182.10.10.1
CUCM1.cluster1.cisco.com	Host (A)	Default	182.10.10.1
CUCM2.cluster1.cisco.com	Host (A)	Default	182.10.10.2
CUCM3.cluster1.cisco.com	Host (A)	Default	182.10.10.3
CUCM4.cluster1.cisco.com	Host (A)	Default	182.10.10.4
TFTP-server1.cluster1.cisco.com	Host (A)	12 Hours	182.10.10.11
TFTP-server2.cluster1.cisco.com	Host (A)	12 Hours	182.10.10.12
www.CUCM-Admin.cisco.com	Alias (CNAME)	Default	CUCM-Admin.cluster1.cisco.com
_sip._tcp.cluster1.cisco.com.	Service (SRV)	Default	CUCM1.cluster1.cisco.com
_sip._tcp.cluster1.cisco.com.	Service (SRV)	Default	CUCM2.cluster1.cisco.com
_sip._tcp.cluster1.cisco.com.	Service (SRV)	Default	CUCM3.cluster1.cisco.com
_sip._tcp.cluster1.cisco.com.	Service (SRV)	Default	CUCM4.cluster1.cisco.com

Dynamic Host Configuration Protocol (DHCP)

DHCP is used by hosts on the network to obtain initial configuration information, including IP address, subnet mask, default gateway, and TFTP server address. DHCP eases the administrative burden of manually configuring each host with an IP address and other configuration information. DHCP also provides automatic reconfiguration of network configuration when devices are moved between subnets. The configuration information is provided by a DHCP server located in the network, which responds to DHCP requests from DHCP-capable clients.

You should configure IP Communications endpoints to use DHCP to simplify deployment of these devices. Any RFC 2131 compliant DHCP server can be used to provide configuration information to IP Communications network devices. When deploying IP telephony devices in an existing data-only network, all you have to do is add DHCP voice scopes to an existing DHCP server for these new voice devices. Because IP telephony devices are configured to use and rely on a DHCP server for IP configuration information, you must deploy DHCP servers in a redundant fashion. At least two DHCP servers should be deployed within the telephony network such that, if one of the servers fails, the other can continue to answer DHCP client requests. You should also ensure that DHCP server(s) are configured with enough IP subnet addresses to handle all DHCP-reliant clients within the network.

DHCP Option 150

IP telephony endpoints can be configured to rely on DHCP Option 150 to identify the source of telephony configuration information, available from a server running the Trivial File Transfer Protocol (TFTP).

In the simplest configuration, where a single TFTP server is offering service to all deployed endpoints, Option 150 is delivered as a single IP address pointing to the system's designated TFTP server. The DHCP scope can also deliver two IP addresses under Option 150, for deployments where there are two TFTP servers within the same cluster. The phone would use the second address if it fails to contact the primary TFTP server, thus providing redundancy. To achieve both redundancy and load sharing between the TFTP servers, you can configure Option 150 to provide the two TFTP server addresses in reverse order for half of the DHCP scopes.

Note If the primary TFTP server is available but is not able to grant the requested file to the phone (for example, because the requesting phone is not configured on that cluster), the phone will not attempt to contact the secondary TFTP server.

Cisco highly recommends using a direct IP address (that is, not relying on a DNS service) for Option 150 because doing so eliminates dependencies on DNS service availability during the phone boot-up and registration process.

Note Even though IP phones support a maximum of two TFTP servers under Option 150, you could configure a Unified CM cluster with more than two TFTP servers. For instance, if a Unified CM system is clustered over a WAN at three separate sites, three TFTP servers could be deployed (one at each site). Phones within each site could then be granted a DHCP scope containing that site's TFTP server within Option 150. This configuration would bring the TFTP service closer to the endpoints, thus reducing latency and ensuring failure isolation between the sites (one site's failure would not affect TFTP service at another site).

Phone DHCP Operation Following a Power Recycle

If a phone is powered down and comes back up while the DHCP server is still offline, it will attempt to use DHCP to obtain IP addressing information (as normal). In the absence of a response from a DHCP server, the phone will re-use the previously received DHCP information to register with Unified CM.

DHCP Lease Times

Configure DHCP lease times as appropriate for the network environment. Given a fairly static network in which PCs and telephony devices remain in the same place for long periods of time, Cisco recommends longer DHCP lease times (for example, one week). Shorter lease times require more frequent renewal of the DHCP configuration and increase the amount of DHCP traffic on the network. Conversely, networks that incorporate large numbers of mobile devices, such as laptops and wireless telephony devices, should be configured with shorter DHCP lease times (for example, one day) to prevent depletion of DHCP-managed subnet addresses. Mobile devices typically use IP addresses for short increments of time and then might not request a DHCP renewal or new address for a long period of time. Longer lease times will tie up these IP addresses and prevent them from being reassigned even when they are no longer being used.

Cisco Unified IP Phones adhere to the conditions of the DHCP lease duration as specified in the DHCP server's scope configuration. Once half the lease time has expired since the last successful DHCP server acknowledgment, the IP phone will request a lease renewal. This DHCP client Request, once acknowledged by the DHCP server, will allow the IP phone to retain use of the IP scope (that is, the IP address, default gateway, subnet mask, DNS server (optional), and TFTP server (optional)) for another lease period. If the DHCP server becomes unavailable, an IP phone will not be able to renew its DHCP lease, and as soon as the lease expires, it will relinquish its IP configuration and will thus become unregistered from Unified CM until a DHCP server can grant it another valid scope.

In centralized call processing deployments, if a remote site is configured to use a centralized DHCP server (through the use of a DHCP relay agent such as the IP Helper Address in Cisco IOS) and if connectivity to the central site is severed, IP phones within the branch will not be able to renew their DHCP scope leases. In this situation, branch IP phones are at risk of seeing their DHCP lease expire, thus losing the use of their IP address, which would lead to service interruption. Given the fact that phones attempt to renew their leases at half the lease time, DHCP lease expiration can occur as soon as half the lease time since the DHCP server became unreachable. For example, if the lease time of a DHCP scope is set to 4 days and a WAN failure causes the DHCP server to be unavailable to the phones in a branch, those phones will be unable to renew their leases at half the lease time (in this case, 2 days). The IP phones could stop functioning as early as 2 days after the WAN failure, unless the WAN comes back up and the DHCP server is available before that time. If the WAN connectivity failure persists, all phones see their DHCP scope expire after a maximum of 4 days from the WAN failure.

This situation can be mitigated by one of the following methods:

- Set the DHCP scope lease to a long duration (for example, 8 days or more).

  This method would give the system administrator a minimum of half the lease time to remedy any DHCP reachability problem. Long lease durations also have the effect of reducing the frequency of network traffic associated with lease renewals.

- Configure co-located DHCP server functionality (for example, run a DHCP server function on the branch's Cisco IOS router).

  This approach is immune to WAN connectivity interruption. One effect of such an approach is to decentralize the management of IP addresses, requiring incremental configuration efforts in each branch. (See DHCP Network Deployments, page 3-25, for more information.)

Note The term *co-located* refers to two or more devices in the same physical location, with no WAN or MAN connection between them.

DHCP Network Deployments

There are two options for deploying DHCP functionality within an IP telephony network:

- Centralized DHCP Server

  Typically, for a single-site campus IP telephony deployment, the DHCP server should be installed at a central location within the campus. As mentioned previously, redundant DHCP servers should be deployed. If the IP telephony deployment also incorporates remote branch telephony sites, as in a centralized multisite Unified CM deployment, a centralized server can be used to provide DHCP service to devices in the remote sites. This type of deployment requires that you configure the **ip helper-address** on the branch router interface. Keep in mind that, if redundant DHCP servers are deployed at the central site, both servers' IP addresses must be configured as **ip helper-address**. Also note that, if branch-side telephony devices rely on a centralized DHCP server and the WAN link between the two sites fails, devices at the branch site will be unable to send DHCP requests or receive DHCP responses.

Note By default, **service dhcp** is enabled on the Cisco IOS device and does not appear in the configuration. Do not disable this service on the branch router because doing so will disable the DHCP relay agent on the device, and the **ip helper-address** configuration command will not work.

- Centralized DHCP Server and Remote Site Cisco IOS DHCP Server

  When configuring DHCP for use in a centralized multisite Unified CM deployment, you can use a centralized DHCP server to provide DHCP service to centrally located devices. Remote devices could receive DHCP service from a locally installed server or from the Cisco IOS router at the remote site. This type of deployment ensures that DHCP services are available to remote telephony devices even during WAN failures. Example 3-1 lists the basic Cisco IOS DHCP server configuration commands.

Example 3-1 Cisco IOS DHCP Server Configuration Commands

```
! Activate DHCP Service on the IOS Device

service dhcp

! Specify any IP Address or IP Address Range to be excluded from the DHCP pool

ip dhcp excluded-address <ip-address>|<ip-address-low> <ip-address-high>

! Specify the name of this specific DHCP pool, the subnet and mask for this
! pool, the default gateway and up to four TFTP

ip dhcp pool <dhcp-pool name>
   network <ip-subnet> <mask>
   default-router <default-gateway-ip>
   option 150 ip <tftp-server-ip-1> ...

! Note: IP phones use only the first two addresses supplied in the option 150
! field even if more than two are configured.
```

Unified CM DHCP Sever (Standalone versus Co-Resident DHCP)

Typically DHCP servers are dedicated machine(s) in most network infrastructures, and they run in conjunction with the DNS and/or the Windows Internet Naming Service (WINS) services used by that network. In some instances, given a small Unified CM deployment with no more than 1000 devices registering to the cluster, you may run the DHCP server on a Unified CM server to support those devices. However, to avoid possible resource contention such as CPU contention with other critical services running on Unified CM, Cisco recommends moving the DHCP Server functionality to a dedicated server. If more than 1000 devices are registered to the cluster, DHCP must *not* be run on a Unified CM server but instead must be run on a dedicated or standalone server(s).

Note The term *co-resident* refers to two or more services or applications running on the same server.

Trivial File Transfer Protocol (TFTP)

Within a Cisco Unified CM system, endpoints such as IP phones rely on a TFTP-based process to acquire configuration files, software images, and other endpoint-specific information. The Cisco TFTP service is a file serving system that can run on one or more Unified CM servers. It builds configuration files and serves firmware files, ringer files, device configuration files, and so forth, to endpoints.

The TFTP file systems can hold several file types, such as the following:

- Phone configuration files
- Phone firmware files
- Certificate Trust List (CTL) files
- Identity Trust List (ITL) files
- Tone localization files
- User interface (UI) localization and dictionary files
- Ringer files
- Softkey files
- Dial plan files for SIP phones

The TFTP server manages and serves two types of files, those that are not modifiable (for example, firmware files for phones) and those that can be modified (for example, configuration files).

A typical configuration file contains a prioritized list of Unified CMs for a device (for example, an SCCP or SIP phone), the TCP ports on which the device connects to those Unified CMs, and an executable load identifier. Configuration files for selected devices contain locale information and URLs for the messages, directories, services, and information buttons on the phone.

When a device's configuration changes, the TFTP server rebuilds the configuration files by pulling the relevant information from the Unified CM database. The new file(s) is then downloaded to the phone once the phone has been reset. As an example, if a single phone's configuration file is modified (for example, during Extension Mobility login or logout), only that file is rebuilt and downloaded to the phone. However, if the configuration details of a device pool are changed (for example, if the primary Unified CM server is changed), then all devices in that device pool need to have their configuration files rebuilt and downloaded. For device pools that contain large numbers of devices, this file rebuilding process can impact server performance.

Note Prior to Cisco Unified CM 6.1, to rebuild modified files, the TFTP server pulled information from the publisher's database. With Unified CM 6.1 and later releases, the TFTP server can perform a local database read from the database on its co-resident subscriber server. Local database read not only provides benefits such as the preservation of user-facing features when the publisher in unavailable, but also allows multiple TFTP servers to be distributed by means of clustering over the WAN. (The same latency rules for clustering over the WAN apply to TFTP servers as to servers with registered phones.) This configuration brings the TFTP service closer to the endpoints, thus reducing latency and ensuring failure isolation between the sites.

When a device requests a configuration file from the TFTP server, the TFTP server searches for the configuration file in its internal caches, the disk, and then alternate Cisco file servers (if specified). If the TFTP server finds the configuration file, it sends it to the device. If the configuration file provides Unified CM names, the device resolves the name by using DNS and opens a connection to the

Unified CM. If the device does not receive an IP address or name, it uses the TFTP server name or IP address to attempt a registration connection. If the TFTP server cannot find the configuration file, it sends a "file not found" message to the device.

A device that requests a configuration file while the TFTP server is rebuilding configuration files or while it is processing the maximum number of requests, will receive a message from the TFTP server that causes the device to request the configuration file later. The Maximum Serving Count service parameter, which can be configured, specifies the maximum number of requests that can be concurrently handled by the TFTP server. (Default value = 500 requests.) Use the default value if the TFTP service is run along with other Cisco CallManager services on the same server. For a dedicated TFTP server, use the following suggested values for the Maximum Serving Count: 1500 for a single-processor system or 3000 for a dual-processor system.

The Cisco Unified IP Phones 8900 Series and 9900 Series request their TFTP configuration files over the HTTP protocol (port 6970), which is much faster than TFTP.

An Example of TFTP in Operation

Every time an endpoint reboots, the endpoint will request a configuration file (via TFTP) whose name is based on the requesting endpoint's MAC address. (For a Cisco Unified IP Phone 7961 with MAC address ABCDEF123456, the file name would be SEPABCDEF123456.cnf.xml.) The received configuration file includes the version of software that the phone must run and a list of Cisco Unified CM servers with which the phone should register. The endpoint might also download, via TFTP, ringer files, softkey templates, and other miscellaneous files to acquire the necessary configuration information before becoming operational.

If the configuration file includes software file(s) version numbers that are different than those the phone is currently using, the phone will also download the new software file(s) from the TFTP server to upgrade itself. The number of files an endpoint must download to upgrade its software varies based on the type of endpoint and the differences between the phone's current software and the new software. For example, Cisco Unified IP Phones 7961, 7970, and 7971 download five software files under the worst-case software upgrade.

TFTP File Transfer Times

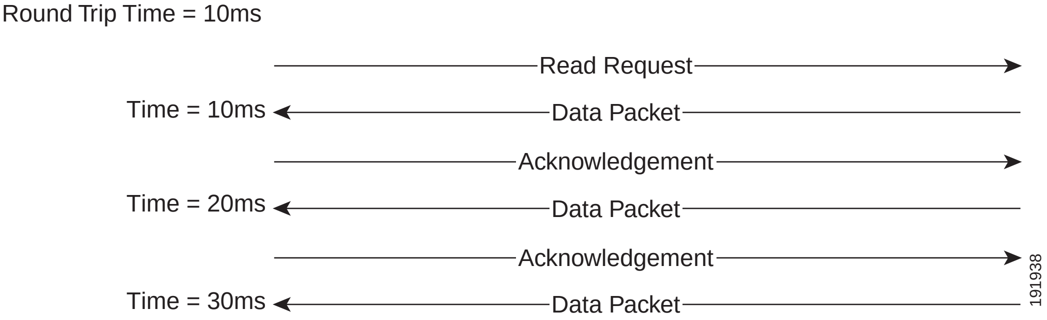
Each time an endpoint requests a file, there is a new TFTP transfer session. For centralized call processing deployments, the time to complete each of these transfers will affect the time it takes for an endpoint to start and become operational as well as the time it takes for an endpoint to upgrade during a scheduled maintenance. While TFTP transfer times are not the only factor that can affect these end states, they are a significant component.

The time to complete each file transfer via TFTP is predictable as a function of the file size, the percentage of TFTP packets that must be retransmitted, and the network latency or round-trip time.

At first glance, network bandwidth might seem to be missing from the previous statement, but it is actually included via the percentage of TFTP packets that must be retransmitted. This is because, if there is not enough network bandwidth to support the file transfer(s), then packets will be dropped by the network interface queuing algorithms and will have to be retransmitted.

TFTP operates on top of the User Datagram Protocol (UDP). Unlike Transmission Control Protocol (TCP), UDP is not a reliable protocol, which means that UDP does not inherently have the ability to detect packet loss. Obviously, detecting packet loss in a file transfer is important, so RFC 1350 defines TFTP as a lock-step protocol. In other words, a TFTP sender will send one packet and wait for a response before sending the next packet (see Figure 3-8).

Figure 3-8 Example of TFTP Packet Transmission Sequence

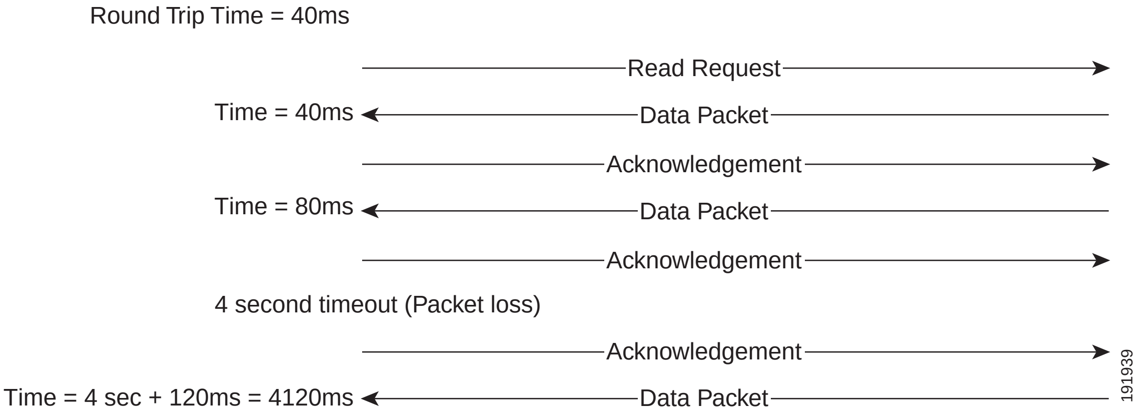


If a response is not received in the timeout period (4 seconds by default), the sender will resend the data packet or acknowledgment. When a packet has been sent five times without a response, the TFTP session fails. Because the timeout period is always the same and not adaptive like a TCP timeout, packet loss can significantly increase the amount of time a transfer session takes to complete.

Because the delay between each data packet is, at a minimum, equal to the network round-trip time, network latency also is a factor in the maximum throughput that a TFTP session can achieve.

In Figure 3-9, the round-trip time has been increased to 40 ms and one packet has been lost in transit. While the error rate is high at 12%, it is easy to see the effect of latency and packet loss on TFTP because the time to complete the session increased from 30 ms (in Figure 3-8) to 4160 ms (in Figure 3-9).

Figure 3-9 Effect of Packet Loss on TFTP Session Completion Time



Use the following formula to calculate how long a TFTP file transfer will take to complete:

FileTransferTime = FileSize ∗ [(RTT + ERR ∗ Timeout) / 512000]

Where:

FileTransferTime is in seconds.

FileSize is in bytes.

RTT is the round-trip time in milliseconds.

ERR is the error rate, or percentage of packets that are lost.

Timeout is in milliseconds.

512000 = (TFTP packet size) * (1000 millisecond per seconds) =
(512 bytes) * (1000 millisecond per seconds)

Table 3-5 and Table 3-6 illustrate the use of this equation to calculate transfer times for the software files for various endpoint device types, protocols, and network latencies.

Table 3-5 TFTP File Transfer Times for SCCP Devices

Device Type (Cisco Unified IP Phone)	Firmware Size (bytes, rounded up to next 100k)	Time to Complete Transfer (1% error rate)				
		40 ms RTT	80 ms RTT	120 ms RTT	160 ms RTT	200 ms RTT
7985	15,000,000	39 min 3 sec	58 min 35 sec	78 min 7 sec	97 min 39 sec	117 min 11 sec
7921	9,700,000	25 min 15 sec	37 min 53 sec	50 min 31 sec	63 min 9 sec	75 min 46 sec
7975	6,300,000	16 min 24 sec	24 min 36 sec	32 min 48 sec	41 min 0 sec	49 min 13 sec
7970 or 7971	6,300,000	16 min 24 sec	24 min 36 sec	32 min 48 sec	41 min 0 sec	49 min 13 sec
7965 or 7945	6,300,000	16 min 24 sec	24 min 36 sec	32 min 48 sec	41 min 0 sec	49 min 13 sec
7962 or 7942	6,200,000	16 min 8 sec	24 min 13 sec	32 min 17 sec	40 min 21 sec	48 min 26 sec
7941 or 7961	6,100,000	15 min 53 sec	23 min 49 sec	31 min 46 sec	39 min 42 sec	47 min 39 sec
7931	6,100,000	15 min 53 sec	23 min 49 sec	31 min 46 sec	39 min 42 sec	47 min 39 sec
7911 or 7906	6,100,000	15 min 53 sec	23 min 49 sec	31 min 46 sec	39 min 42 sec	47 min 39 sec
7935	2,100,000	5 min 28 sec	8 min 12 sec	10 min 56 sec	13 min 40 sec	16 min 24 sec
7920	1,200,000	3 min 7 sec	4 min 41 sec	6 min 15 sec	7 min 48 sec	9 min 22 sec
7936	1,800,000	4 min 41 sec	7 min 1 sec	9 min 22 sec	11 min 43 sec	14 min 3 sec
7940 or 7960	900,000	2 min 20 sec	3 min 30 sec	4 min 41 sec	5 min 51 sec	7 min 1 sec
7910	400,000	1 min 2 sec	1 min 33 sec	2 min 5 sec	2 min 36 sec	3 min 7 sec
7912	400,000	1 min 2 sec	1 min 33 sec	2 min 5 sec	2 min 36 sec	3 min 7 sec
7905	400,000	1 min 2 sec	1 min 33 sec	2 min 5 sec	2 min 36 sec	3 min 7 sec
7902	400,000	1 min 2 sec	1 min 33 sec	2 min 5 sec	2 min 36 sec	3 min 7 sec

Table 3-6 TFTP File Transfer Times for SIP Devices

Device Type (Cisco Unified IP Phone)	Firmware Size (bytes, rounded up to next 100k)	Time to Complete Transfer (1% error rate)				
		40 ms RTT	80 ms RTT	120 ms RTT	160 ms RTT	200 ms RTT
7975	6,600,000	17 min 11 sec	25 min 46 sec	34 min 22 sec	42 min 58 sec	51 min 33 sec
7970 or 7971	6,700,000	17 min 26 sec	26 min 10 sec	34 min 53 sec	43 min 37 sec	52 min 20 sec
7965 or 7945	6,600,000	17 min 11 sec	25 min 46 sec	34 min 22 sec	42 min 58 sec	51 min 33 sec
7962 or 7942	6,500,000	16 min 55 sec	25 min 23 sec	33 min 51 sec	42 min 19 sec	50 min 46 sec
7941 or 7961	6,500,000	16 min 55 sec	25 min 23 sec	33 min 51 sec	42 min 19 sec	50 min 46 sec
7911 or 7906	6,400,000	16 min 40 sec	25 min 0 sec	33 min 20 sec	41 min 40 sec	50 min 0 sec
7940 or 7960	900,000	2 min 20 sec	3 min 30 sec	4 min 41 sec	5 min 51 sec	7 min 1 sec
7912	400,000	1 min 2 sec	1 min 33 sec	2 min 5 sec	2 min 36 sec	3 min 7 sec
7905	400,000	1 min 2 sec	1 min 33 sec	2 min 5 sec	2 min 36 sec	3 min 7 sec

The values in Table 3-5 and Table 3-6 are the approximate times to download the necessary firmware files to the phone. This is *not* an estimate of the time that it will take for a phone to upgrade to the new firmware and become operational.

Cisco Unified IP Phone Firmware Releases 7.*x* have a 10-minute timeout when downloading new files. If the transfer is not completed within this time, the phone will discard the download even if the transfer completes successfully later. If you experience this problem, Cisco recommends that you use a local TFTP server to upgrade phones to the 8.*x* firmware releases, which have a timeout value of 61 minutes.

Because network latency and packet loss have such an effect on TFTP transfer times, a local TFTP Server can be advantageous. This local TFTP server may be a Unified CM subscriber in a deployment with cluster over the WAN or an alternative local TFTP "Load Server" running on a Cisco Integrated Services Router (ISR), for example. Newer endpoints (which have larger firmware files) can be configured with a Load Server address, which allows the endpoint to download the relatively small configuration files from the central TFTP server but use a local TFTP Server (which is not part of the Unified CM cluster) to download the larger software files. For details on which Cisco Unified IP Phones support an alternative local TFTP Load Server, refer to the product documentation for your particular phone models (available at http://www.cisco.com).

Note The exact process each phone goes through on startup and the size of the files downloaded will depend on the phone model, the signaling type configured for the phone (SCCP, MGCP, or SIP) and the previous state of the phone. While there are differences in which files are requested, the general process each phone follows is the same, and in all cases a TFTP server is used to request and deliver the appropriate files. The general recommendations for TFTP server deployment do not change based on the protocol and/or phone models deployed.

TFTP Server Redundancy

Option 150 allows up to two IP addresses to be returned to phones as part of the DHCP scope. The phone tries the first address in the list, and it tries the subsequent address only if it cannot establish communications with the first TFTP server. This address list provides a redundancy mechanism that enables phones to obtain TFTP services from another server even if their primary TFTP server has failed.

TFTP Load Sharing

Cisco recommends that you grant different ordered lists of TFTP servers to different subnets to allow for load balancing. For example:

- In subnet 10.1.1.0/24: Option 150: TFTP1_Primary, TFTP1_Secondary
- In subnet 10.1.2.0/24: Option 150: TFTP1_Secondary, TFTP1_Primary

Under normal operations, a phone in subnet 10.1.1.0/24 will request TFTP services from TFTP1_Primary, while a phone in subnet 10.1.2.0/24 will request TFTP services from TFTP1_Secondary. If TFTP1_Primary fails, then phones from both subnets will request TFTP services from TFTP1_Secondary.

Load balancing avoids having a single TFTP server hot-spot, where all phones from multiple clusters rely on the same server for service. TFTP load balancing is especially important when phone software loads are transferred, such as during a Unified CM upgrade, because more files of larger size are being transferred, thus imposing a bigger load on the TFTP server.

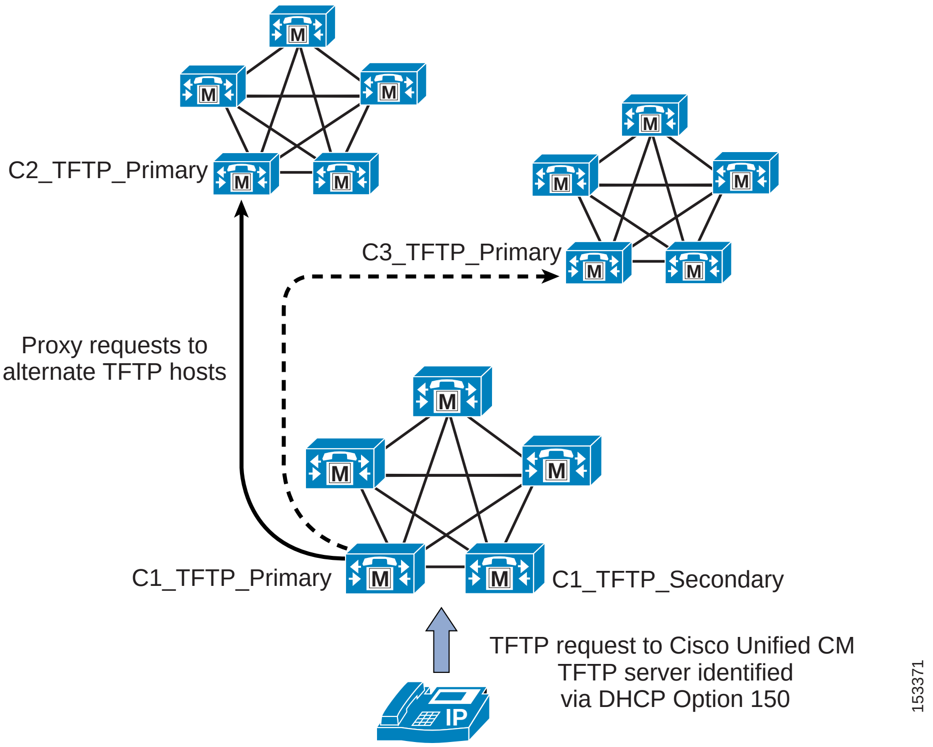
Centralized TFTP and Proxy TFTP Services

In multi-cluster systems, it is possible to have a single subnet or VLAN containing phones from multiple clusters. In this situation, the TFTP servers whose addresses are provided to all phones in the subnet or VLAN must answer the file transfer requests made by each phone, regardless of which cluster contains the phone. In a centralized TFTP deployment, a set of TFTP servers associated with one of the clusters must provide TFTP services to all the phones in the multi-cluster system.

In order to provide this single point of file access, each cluster's TFTP server must be able to serve files via the central proxy TFTP server. With Cisco Unified CM 5.0 and later releases, this proxy arrangement is accomplished by configuring a set of possible redirect locations in the central TFTP server, pointing to each of the other clusters' TFTP servers. This configuration uses a HOST redirect statement in the Alternate File Locations on the centralized TFTP server, one for each of the other clusters. Each of the redundant TFTP servers in the centralized cluster should point to one of the redundant servers in each of the child clusters. It is not necessary to point the centralized server to both redundant servers in the child clusters because the redistribution of files within each individual cluster and the failover mechanisms of the phones between the redundant servers in the central cluster provide for a very high degree of fault tolerance.

Figure 3-10 shows an example of the operation of this process. A request from a phone registered to Cluster 3 is directed to the centralized TFTP server configured in Cluster 1 (C1_TFTP_Primary). This server will in turn query each of the configured alternate TFTP servers until one responds with a copy of the file initially requested by the phone. Requests to the centralized secondary TFTP server (C1_TFTP_Secondary) will be sent by proxy to the other clusters' secondary TFTP servers until either the requested file is found or all servers report that the requested file does not exist.

Figure 3-10 Centralized TFTP Servers



Centralized TFTP in a Mixed Environment, with Servers Running Different Releases of Cisco Unified CM

With the introduction of the Security by Default feature in Cisco Unified CM 8.*x* versions, the endpoints registered to an 8.*x* or later version cluster require the initial trust list (ITL) file in addition to the other configuration files. The endpoints registered to clusters running Unified CM versions prior to 8.0 do not recognize this file.

In a centralized TFTP implementation, all IP phones request configuration files from the same TFTP cluster. This requires that the centralized TFTP function run in an environment where all clusters (including the TFTP cluster) either support ITL files homogeneously (that is, they are all on Unified CM versions 8.*x* or later) or do not work with ITL files (that is, they are on versions 7.*x* or earlier).

If the centralized TFTP implementation has a mix of pre-8.*x* and 8.*x* or later versions of Unified CM, then the ITL functions will have to be disabled temporarily on the clusters that support ITL files. For more information, refer to the Cisco Proxy TFTP Server configuration in the *Cisco Unified Communication Manager Features and Services* guide and the Cisco TFTP section in the *Cisco Unified Communication Manager System Guide*, both available at

http://www.cisco.com/en/US/products/sw/voicesw/ps556/prod_maintenance_guides_list.html

Network Time Protocol (NTP)

NTP allows network devices to synchronize their clocks to a network time server or network-capable clock. NTP is critical for ensuring that all devices in a network have the same time. When troubleshooting or managing a telephony network, it is crucial to synchronize the time stamps within all error and security logs, traces, and system reports on devices throughout the network. This synchronization enables administrators to recreate network activities and behaviors based on a common timeline. Billing records and call detail records (CDRs) also require accurate synchronized time.

Unified CM NTP Time Synchronization

Time synchronization is especially critical on Unified CM servers. In addition to ensuring that CDR records are accurate and that log files are synchronized, having an accurate time source is necessary for any future IPSec features to be enabled within the cluster and for communications with any external entity.

Unified CM automatically synchronizes the NTP time of all subscribers in the cluster to the publisher. During installation, each subscriber is automatically configured to point to an NTP server running on the publisher. The publisher considers itself to be a master server and provides time for the cluster based on its internal hardware clock unless it is configured to synchronize from an external server. Cisco highly recommends configuring the publisher to point to a Stratum-1, Stratum-2, or Stratum-3 NTP server to ensure that the cluster time is synchronized with an external time source.

Cisco recommends synchronizing Unified CM with a Cisco IOS or Linux-based NTP server. Using Windows Time Services as an NTP server is not recommended or supported because Windows Time Services often use Simple Network Time Protocol (SNTP), and Linux-based Unified CM cannot successfully synchronize with SNTP.

The external NTP server specified for the primary node should be NTP v4 (version 4) to avoid potential compatibility, accuracy, and network jitter problems. External NTP servers *must* be NTP v4 if IPv6 addressing is used.

For additional information about NTP time synchronization in a Cisco Unified Communications environment, refer to the *Cisco IP Telephony Clock Synchronization: Best Practices* white paper, available at

http://www.cisco.com/en/US/products/sw/voicesw/ps556/products_white_paper0900aecd8037fdb5.shtml

Cisco IOS and CatOS NTP Time Synchronization

Time synchronization is also important for other devices within the network. Cisco IOS routers and Catalyst switches should be configured to synchronize their time with the rest of the network devices via NTP. This is critical for ensuring that debug, syslog, and console log messages are time-stamped appropriately. Troubleshooting telephony network issues is simplified when a clear timeline can be drawn for events that occur on devices throughout the network.

WAN Infrastructure

Proper WAN infrastructure design is also extremely important for normal Unified Communications operation on a converged network. Proper infrastructure design requires following basic configuration and design best practices for deploying a WAN that is as highly available as possible and that provides guaranteed throughput. Furthermore, proper WAN infrastructure design requires deploying end-to-end QoS on all WAN links. The following sections discuss these requirements:

- WAN Design and Configuration, page 3-34
- WAN Quality of Service (QoS), page 3-37
- Resource Reservation Protocol (RSVP), page 11-42
- Bandwidth Provisioning, page 3-45

WAN Design and Configuration

Properly designing a WAN requires building fault-tolerant network links and planning for the possibility that these links might become unavailable. By carefully choosing WAN topologies, provisioning the required bandwidth, and approaching the WAN infrastructure as another layer in the network topology, you can build a fault-tolerant and redundant network. The following sections examine the required infrastructure layers and network services:

- Deployment Considerations, page 3-34
- Guaranteed Bandwidth, page 3-36
- Best-Effort Bandwidth, page 3-37

Deployment Considerations

WAN deployments for voice networks may use a hub-and-spoke, fully meshed, or partially meshed topology. A hub-and-spoke topology consists of a central hub site and multiple remote spoke sites connected into the central hub site. In this scenario, each remote or spoke site is one WAN-link hop away from the central or hub site and two WAN-link hops away from all other spoke sites. A meshed topology may contain multiple WAN links and any number of hops between the sites. In this scenario there may be many different paths to the same site or there may be different links used for communication with some sites compared to other sites. The simplest example is three sites, each with a WAN link to the other two sites, forming a triangle. In that case there are two potential paths between each site to each other site.

Topology-unaware call admission control requires the WAN to be hub-and-spoke, or a spoke-less hub in the case of MPLS VPN. This topology ensures that call admission control, provided by Unified CM's locations or a gatekeeper, works properly in keeping track of the bandwidth available between any two sites in the WAN. In addition, multiple hub-and-spoke deployments can be interconnected via WAN links.

Topology-aware call admission control may be used with either hub-and-spoke or an arbitrary WAN topology. This form of call admission control requires parts of the WAN infrastructure to support Resource Reservation Protocol (RSVP). For details, see Resource Reservation Protocol (RSVP), page 11-42, and Call Admission Control, page 11-1.

For more information about centralized and distributed multisite deployment models as well as Multiprotocol Label Switching (MPLS) implications for these deployment models, see the chapter on Unified Communications Deployment Models, page 5-1.

WAN links should, when possible, be made redundant to provide higher levels of fault tolerance. Redundant WAN links provided by different service providers or located in different physical ingress/egress points within the network can ensure backup bandwidth and connectivity in the event that a single link fails. In non-failure scenarios, these redundant links may be used to provide additional bandwidth and offer load balancing of traffic on a per-flow basis over multiple paths and equipment within the WAN. Topology-unaware call admission control normally requires redundant paths to be over-provisioned and under-subscribed to allow for failures that reduce the available bandwidth between sites without the call admission control mechanism being aware of those failures or the reduction in bandwidth. Topology-aware call admission control is able to adjust dynamically to many of the topology changes and allows for efficient use of the total available bandwidth.

Voice and data should remain converged at the WAN, just as they are converged at the LAN. QoS provisioning and queuing mechanisms are typically available in a WAN environment to ensure that voice and data can interoperate on the same WAN links. Attempts to separate and forward voice and data over different links can be problematic in many instances because the failure of one link typically forces all traffic over a single link, thus diminishing throughput for each type of traffic and in most cases reducing the quality of voice. Furthermore, maintaining separate network links or devices makes troubleshooting and management difficult at best.

Because of the potential for WAN links to fail or to become oversubscribed, Cisco recommends deploying non-centralized resources as appropriate at sites on the other side of the WAN. Specifically, media resources, DHCP servers, voice gateways, and call processing applications such as Survivable Remote Site Telephony (SRST) and Cisco Unified Communications Manager Express (Unified CME) should be deployed at non-central sites when and if appropriate, depending on the site size and how critical these functions are to that site. Keep in mind that de-centralizing voice applications and devices can increase the complexity of network deployments, the complexity of managing these resources throughout the enterprise, and the overall cost of a the network solution; however, these factors can be mitigated by the fact that the resources will be available during a WAN link failure.

When deploying voice in a WAN environment, Cisco recommends that you use the lower-bandwidth G.729 codec for any voice calls that will traverse WAN links because this practice will provide bandwidth savings on these lower-speed links. Furthermore, media resources such as MoH should be configured to use multicast transport mechanism when possible because this practice will provide additional bandwidth savings.

Where calls are made over best-effort networks with no QoS guarantees for voice, consider using Internet Low Bit Rate Codec (iLBC), which enables graceful speech quality degradation and good error resilience characteristics in networks where frames can get lost. See Table 3-9 for details of bandwidth consumption based on codec type and sample size.

Delay in IP Voice Networks

Recommendation G.114 of the International Telecommunication Union (ITU) states that the one-way delay in a voice network should be less than or equal to 150 milliseconds. It is important to keep this in mind when implementing low-speed WAN links within a network. Topologies, technologies, and physical distance should be considered for WAN links so that one-way delay is kept at or below this 150-millisecond recommendation. Implementing a VoIP network where the one-way delay exceeds 150 milliseconds introduces issues not only with the quality of the voice call but also with call setup and media cut-through times because several call signaling messages need to be exchanged between each device and the call processing application in order to establish the call.

Guaranteed Bandwidth

Because voice is typically deemed a critical network application, it is imperative that bearer and signaling voice traffic always reaches its destination. For this reason, it is important to choose a WAN topology and link type that can provide guaranteed dedicated bandwidth. The following WAN link technologies can provide guaranteed dedicated bandwidth:

- Leased Lines
- Frame Relay
- Asynchronous Transfer Mode (ATM)
- ATM/Frame-Relay Service Interworking
- Multiprotocol Label Switching (MPLS)
- Cisco Voice and Video Enabled IP Security VPN (IPSec V3PN)

These link technologies, when deployed in a dedicated fashion or when deployed in a private network, can provide guaranteed traffic throughput. All of these WAN link technologies can be provisioned at specific speeds or bandwidth sizes. In addition, these link technologies have built-in mechanisms that help guarantee throughput of network traffic even at low link speeds. Features such as traffic shaping, fragmentation and packet interleaving, and committed information rates (CIR) can help ensure that packets are not dropped in the WAN, that all packets are given access at regular intervals to the WAN link, and that enough bandwidth is available for all network traffic attempting to traverse these links.

Dynamic Multipoint VPN (DMVPN)

Spoke-to-spoke DMVPN networks can provide benefits for Cisco Unified Communications compared with hub-and-spoke topologies. Spoke-to-spoke tunnels can provide a reduction in end-to-end latency by reducing the number of WAN hops and decryption/encryption stages. In addition, DMVPN offers a simplified means of configuring the equivalent of a full mesh of point-to-point tunnels without the associated administrative and operational overhead. The use of spoke-to-spoke tunnels also reduces traffic at the hub, thus providing bandwidth and router processing capacity savings. Spoke-to-spoke DMVPN networks, however, are sensitive to the delay variation (jitter) caused during the transition of RTP packets routing from the spoke-hub-spoke path to the spoke-to-spoke path. This variation in delay during the DMVPN path transition occurs very early in the call and is generally unnoticeable, although a single momentary audio distortion might be heard if the latency difference is above 100 ms.

For information on the deployment of multisite DMVPN WANs with centralized call processing, refer to the *Cisco Unified Communications Voice over Spoke-to-Spoke DMVPN Test Results and Recommendations*, available at http://www.cisco.com/go/designzone.

Best-Effort Bandwidth

There are some WAN topologies that are unable to provide guaranteed dedicated bandwidth to ensure that network traffic will reach its destination, even when that traffic is critical. These topologies are extremely problematic for voice traffic, not only because they provide no mechanisms to provision guaranteed network throughput, but also because they provide no traffic shaping, packet fragmentation and interleaving, queuing mechanisms, or end-to-end QoS to ensure that critical traffic such as voice will be given preferential treatment.

The following WAN network topologies and link types are examples of this kind of best-effort bandwidth technology:

- The Internet
- DSL
- Cable
- Satellite
- Wireless

In most cases, none of these link types can provide the guaranteed network connectivity and bandwidth required for critical voice and voice applications. However, these technologies might be suitable for personal or telecommuter-type network deployments. At times, these topologies can provide highly available network connectivity and adequate network throughput; but at other times, these topologies can become unavailable for extended periods of time, can be throttled to speeds that render network throughput unacceptable for real-time applications such as voice, or can cause extensive packet losses and require repeated retransmissions. In other words, these links and topologies are unable to provide guaranteed bandwidth, and when traffic is sent on these links, it is sent best-effort with no guarantee that it will reach its destination. For this reason, Cisco recommends that you do *not* use best-effort WAN topologies for voice-enabled networks that require enterprise-class voice services and quality.

Note There are some new QoS mechanisms for DSL and cable technologies that can provide guaranteed bandwidth; however, these mechanisms are not typically deployed by many service providers. For any service that offers QoS guarantees over networks that are typically based on best-effort, it is important to review and understand the bandwidth and QoS guarantees offered in the service provider's service level agreement (SLA).

Note Upstream and downstream QoS mechanisms are now supported for wireless networks. For more information on QoS for Voice over Wireless LANs, refer to the *Voice over Wireless LAN Design Guide*, available at http://www.cisco.com/en/US/solutions/ns340/ns414/ns742/ns820/landing_voice_wireless.html.

WAN Quality of Service (QoS)

Before placing voice and video traffic on a network, it is important to ensure that there is adequate bandwidth for all required applications. Once this bandwidth has been provisioned, voice priority queuing must be performed on all interfaces. This queuing is required to reduce jitter and possible packet loss if a burst of traffic oversubscribes a buffer. This queuing requirement is similar to the one for the LAN infrastructure.

Next, the WAN typically requires additional mechanisms such as traffic shaping to ensure that WAN links are not sent more traffic than they can handle, which could cause dropped packets.

Finally, link efficiency techniques can be applied to WAN paths. For example, link fragmentation and interleaving (LFI) can be used to prevent small voice packets from being queued behind large data packets, which could lead to unacceptable delays on low-speed links.

The goal of these QoS mechanisms is to ensure reliable, high-quality voice by reducing delay, packet loss, and jitter for the voice traffic. Table 3-7 lists the QoS features and tools required for the WAN infrastructure to achieve this goal.

Table 3-7 QoS Features and Tools Required to Support Unified Communications for Each WAN Technology and Link Speed

WAN Technology	Link Speed: 56 kbps to 768 kbps	Link Speed: Greater than 768 kbps
Leased Lines	• Multilink Point-to-Point Protocol (MLP) • MLP Link Fragmentation and Interleaving (LFI) • Low Latency Queuing (LLQ) • Optional: Compressed Real-Time Transport Protocol (cRTP)	• LLQ
Frame Relay (FR)	• Traffic Shaping • LFI (FRF.12) • LLQ • Optional: cRTP • Optional: Voice-Adaptive Traffic Shaping (VATS) • Optional: Voice-Adaptive Fragmentation (VAF)	• Traffic Shaping • LLQ • Optional: VATS
Asynchronous Transfer Mode (ATM)	• TX-ring buffer changes • MLP over ATM • MLP LFI • LLQ • Optional: cRTP (requires MLP)	• TX-ring buffer changes • LLQ
Frame Relay and ATM Service Inter-Working (SIW)	• TX-ring buffer changes • MLP over ATM and FR • MLP LFI • LLQ • Optional: cRTP (requires MLP)	• TX-ring buffer changes • MLP over ATM and FR • LLQ
Multiprotocol Label Switching (MPLS)	• Same as above, according to the interface technology • Class-based marking is generally required to remark flows according to service provider specifications	• Same as above, according to the interface technology • Class-based marking is generally required to remark flows according to service provider specifications

The following sections highlight some of the most important features and techniques to consider when designing a WAN to support both voice and data traffic:

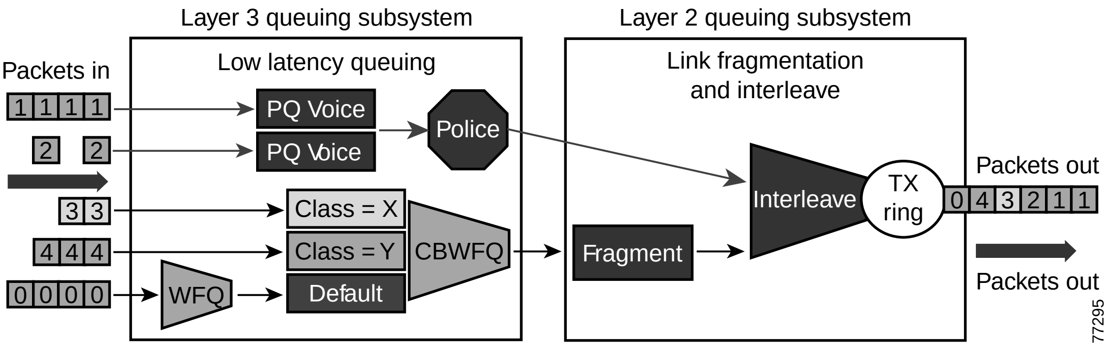
- Traffic Prioritization, page 3-39
- Link Efficiency Techniques, page 3-40
- Traffic Shaping, page 3-42

Traffic Prioritization

In choosing from among the many available prioritization schemes, the major factors to consider include the type of traffic involved and the type of media on the WAN. For multi-service traffic over an IP WAN, Cisco recommends low-latency queuing (LLQ) for all links. This method supports up to 64 traffic classes, with the ability to specify, for example, priority queuing behavior for voice and interactive video, minimum bandwidth class-based weighted fair queuing for voice control traffic, additional minimum bandwidth weighted fair queues for mission critical data, and a default best-effort queue for all other traffic types.

Figure 3-11 shows an example prioritization scheme.

Figure 3-11 Optimized Queuing for VoIP over the WAN



Cisco recommends the following prioritization criteria for LLQ:

- The criterion for *voice* to be placed into a priority queue is the differentiated services code point (DSCP) value of 46, or a per-hop behavior (PHB) value of EF.
- The criterion for *video conferencing* traffic to be placed into a priority queue is a DSCP value of 34, or a PHB value of AF41. However, due to the larger packet sizes of video traffic, these packets should be placed in the priority queue only on WAN links that are faster than 768 Kbps. Link speeds below this value require packet fragmentation, but packets placed in the priority queue are not fragmented, thus smaller voice packets could be queued behind larger video packets. For links speeds of 768 Kbps or lower, video conferencing traffic should be placed in a separate class-based weighted fair queue (CBWFQ).

Note One-way video traffic, such as the traffic generated by streaming video applications for services such as video-on-demand or live video feeds, should always use a CBWFQ scheme because that type of traffic has a much higher delay tolerance than two-way video conferencing traffic

- As the WAN links become congested, it is possible to starve the *voice control* signaling protocols, thereby eliminating the ability of the IP phones to complete calls across the IP WAN. Therefore, voice control protocols, such as H.323, MGCP, and Skinny Client Control Protocol (SCCP), require their own class-based weighted fair queue. The entrance criterion for this queue is a DSCP value of 24 or a PHB value of CS3.

Note Cisco has transitioned the marking of voice control protocols from DSCP 26 (PHB AF31) to DSCP 24 (PHB CS3). However, some products still mark signaling traffic as DSCP 26 (PHB AF31); therefore, Cisco recommends that you reserve both AF31 and CS3 for call signaling.

- In some cases, certain data traffic might require better than best-effort treatment. This traffic is referred to as *mission-critical data*, and it is placed into one or more queues that have the required amount of bandwidth. The queuing scheme within this class is first-in-first-out (FIFO) with a minimum allocated bandwidth. Traffic in this class that exceeds the configured bandwidth limit is placed in the default queue. The entrance criterion for this queue could be a Transmission Control Protocol (TCP) port number, a Layer 3 address, or a DSCP/PHB value.
- All remaining enterprise traffic can be placed in a default queue for best-effort treatment. If you specify the keyword **fair**, the queuing algorithm will be weighted fair queuing (WFQ).

Scavenger Class

The Scavenger class is intended to provide less than best-effort services to certain applications. Applications assigned to this class have little or no contribution to the organizational objectives of the enterprise and are typically entertainment oriented in nature. Assigning Scavenger traffic to a minimal bandwidth queue forces it to be squelched to virtually nothing during periods of congestion, but it allows it to be available if bandwidth is not being used for business purposes, such as might occur during off-peak hours.

- Scavenger traffic should be marked as DSCP CS1.
- Scavenger traffic should be assigned the lowest configurable queuing service. For instance, in Cisco IOS, this means assigning a CBWFQ of 1% to Scavenger class.

Link Efficiency Techniques

The following link efficiency techniques improve the quality and efficiency of low-speed WAN links.

Compressed Real-Time Transport Protocol (cRTP)

You can increase link efficiency by using Compressed Real-Time Transport Protocol (cRTP). This protocol compresses a 40-byte IP, User Datagram Protocol (UDP), and RTP header into approximately two to four bytes. cRTP operates on a per-hop basis. Use cRTP on a particular link only if that link meets *all* of the following conditions:

- Voice traffic represents more than 33% of the load on the specific link.
- The link uses a low bit-rate codec (such as G.729).
- No other real-time application (such as video conferencing) is using the same link.

If the link fails to meet any one of the preceding conditions, then cRTP is not effective and you should not use it on that link. Another important parameter to consider before using cRTP is router CPU utilization, which is adversely affected by compression and decompression operations.

cRTP on ATM and Frame Relay Service Inter-Working (SIW) links requires the use of Multilink Point-to-Point Protocol (MLP).

Note that cRTP compression occurs as the final step before a packet leaves the egress interface; that is, after LLQ class-based queueing has occurred. Beginning in Cisco IOS Release 12.(2)2T and later, cRTP provides a feedback mechanism to the LLQ class-based queueing mechanism that allows the bandwidth in the *voice* class to be configured based on the compressed packet value. With Cisco IOS releases prior to 12.(2)2T, this mechanism is not in place, so the LLQ is unaware of the compressed bandwidth and, therefore, the *voice* class bandwidth has to be provisioned as if no compression is taking place. Table 3-8 shows an example of the difference in *voice* class bandwidth configuration given a 512-kbps link with G.729 codec and a requirement for 10 calls.

Note that Table 3-8 assumes 24 kbps for non-cRTP G.729 calls and 10 kbps for cRTP G.729 calls. These bandwidth numbers are based on voice payload and IP/UDP/RTP headers only. They do not take into consideration Layer 2 header bandwidth. However, actual bandwidth provisioning should also include Layer 2 header bandwidth based on the type WAN link used.

Table 3-8 LLQ Voice Class Bandwidth Requirements for 10 Calls with 512 kbps Link Bandwidth and G.729 Codec

Cisco IOS Release	With cRTP Not Configured	With cRTP Configured
Prior to 12.2(2)T	240 kbps	240 kbps[1]
12.2(2)T or later	240 kbps	100 kbps

[1] 140 kbps of unnecessary bandwidth must be configured in the LLQ *voice* class.

It should also be noted that, beginning in Cisco IOS Release 12.2(13)T, cRTP can be configured as part of the voice class with the Class-Based cRTP feature. This option allows cRTP to be specified within a class, attached to an interface via a service policy. This new feature provides compression statistics and bandwidth status via the **show policy interface** command, which can be very helpful in determining the offered rate on an interface service policy class given the fact that cRTP is compressing the IP/RTP headers.

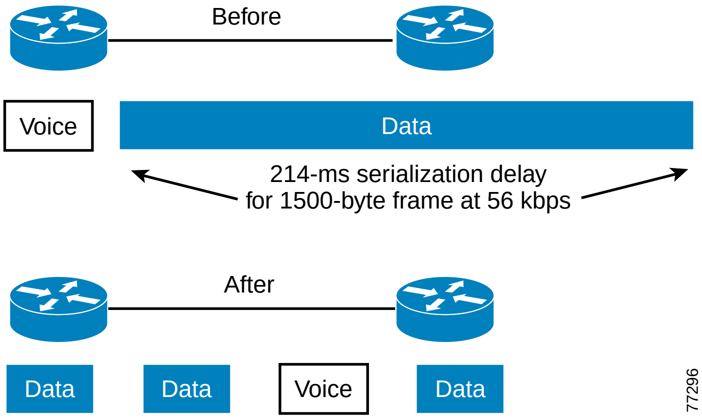
For additional recommendations about using cRTP with a Voice and Video Enabled IPSec VPN (V3PN), refer to the V3PN documentation available at

http://www.cisco.com/en/US/solutions/ns340/ns414/ns742/ns817/landing_voice_video.html

Link Fragmentation and Interleaving (LFI)

For low-speed links (less than 768 kbps), use of link fragmentation and interleaving (LFI) mechanisms is required for acceptable voice quality. This technique limits jitter by preventing voice traffic from being delayed behind large data frames, as illustrated in Figure 3-12. The two techniques that exist for this purpose are Multilink Point-to-Point Protocol (MLP) LFI (for Leased Lines, ATM, and SIW) and FRF.12 for Frame Relay.

Figure 3-12 Link Fragmentation and Interleaving (LFI)



Voice-Adaptive Fragmentation (VAF)

In addition to the LFI mechanisms mentioned above, voice-adaptive fragmentation (VAF) is another LFI mechanism for Frame Relay links. VAF uses FRF.12 Frame Relay LFI; however, once configured, fragmentation occurs only when traffic is present in the LLQ priority queue or when H.323 signaling packets are detected on the interface. This method ensures that, when voice traffic is being sent on the WAN interface, large packets are fragmented and interleaved. However, when voice traffic is not present on the WAN link, traffic is forwarded across the link unfragmented, thus reducing the overhead required for fragmentation.

VAF is typically used in combination with voice-adaptive traffic shaping (see Voice-Adaptive Traffic Shaping (VATS), page 3-44). VAF is an optional LFI tool, and you should exercise care when enabling it because there is a slight delay between the time when voice activity is detected and the time when the LFI mechanism engages. In addition, a configurable deactivation timer (default of 30 seconds) must expire after the last voice packet is detected and before VAF is deactivated, so during that time LFI will occur unnecessarily. VAF is available in Cisco IOS Release 12.2(15)T and later.

Traffic Shaping

Traffic shaping is required for multiple-access, non-broadcast media such as ATM and Frame Relay, where the physical access speed varies between two endpoints and several branch sites are typically aggregated to a single router interface at the central site.

Figure 3-13 illustrates the main reasons why traffic shaping is needed when transporting voice and data on the same IP WAN.

Figure 3-13 Traffic Shaping with Frame Relay and ATM

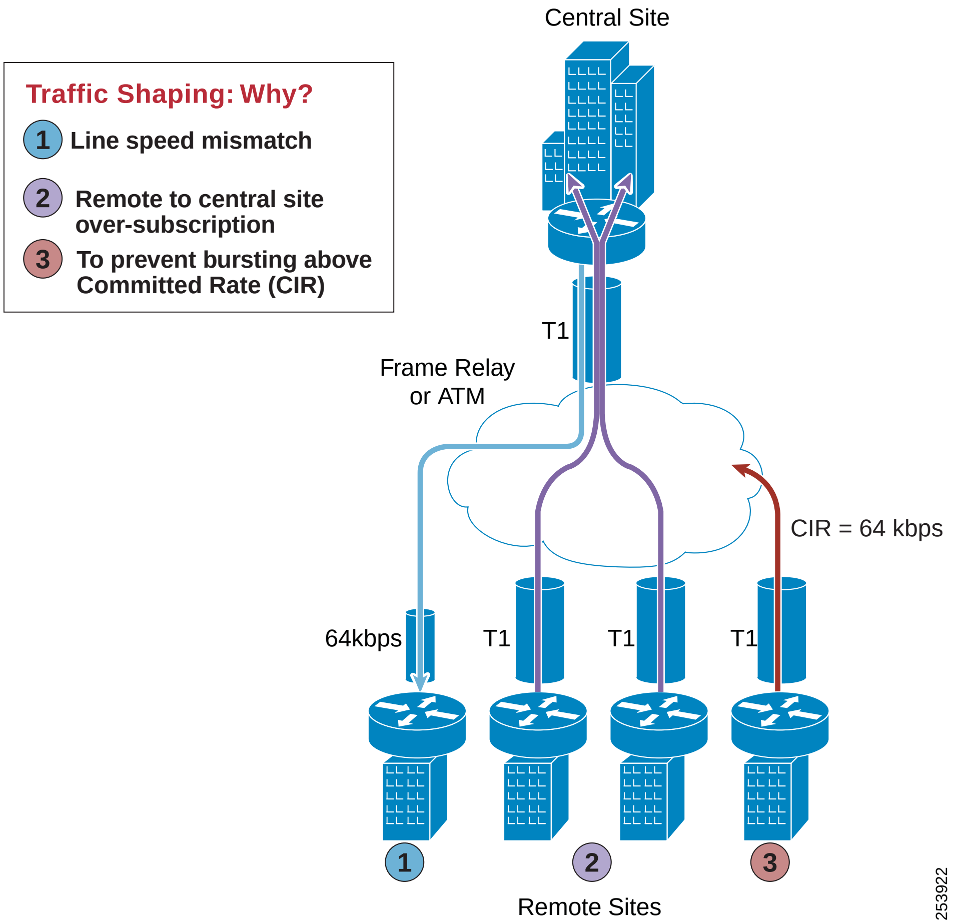


Figure 3-13 shows three different scenarios:

1. Line speed mismatch

   While the central-site interface is typically a high-speed one (such as T1 or higher), smaller remote branch interfaces may have significantly lower line speeds, such as 64 kbps. If data is sent at full rate from the central site to a slow-speed remote site, the interface at the remote site might become congested, resulting in dropped packets which causes a degradation in voice quality.

2. Oversubscription of the link between the central site and the remote sites

   It is common practice in Frame Relay or ATM networks to oversubscribe bandwidth when aggregating many remote sites to a single central site. For example, there may be multiple remote sites that connect to the WAN with a T1 interface, yet the central site has only a single T1 interface. While this configuration allows the deployment to benefit from statistical multiplexing, the router interface at the central site can become congested during traffic bursts, thus degrading voice quality.

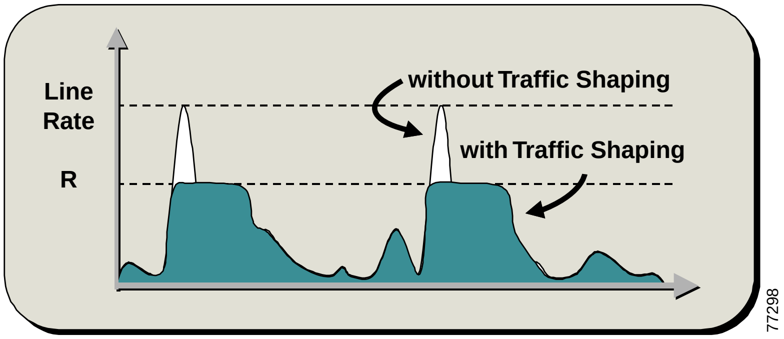
3. Bursting above Committed Information Rate (CIR)

   Another common configuration is to allow traffic bursts above the CIR, which represents the rate that the service provider has guaranteed to transport across its network with no loss and low delay. For example, a remote site with a T1 interface might have a CIR of only 64 kbps. When more than

64 kbps worth of traffic is sent across the WAN, the provider marks the additional traffic as "discard eligible." If congestion occurs in the provider network, this traffic will be dropped with no regard to traffic classification, possibly having a negative effect on voice quality.

Traffic shaping provides a solution to these issues by limiting the traffic sent out an interface to a rate lower than the line rate, thus ensuring that no congestion occurs on either end of the WAN. Figure 3-14 illustrates this mechanism with a generic example, where R is the rate with traffic shaping applied.

Figure 3-14 Traffic Shaping Mechanism



Voice-Adaptive Traffic Shaping (VATS)

VATS is an optional dynamic mechanism that shapes traffic on Frame Relay permanent virtual circuits (PVCs) at different rates based on whether voice is being sent across the WAN. The presence of traffic in the LLQ voice priority queue or the detection of H.323 signaling on the link causes VATS to engage. Typically, Frame Relay shapes traffic to the guaranteed bandwidth or CIR of the PVC at all times. However, because these PVCs are typically allowed to burst above the CIR (up to line speed), traffic shaping keeps traffic from using the additional bandwidth that might be present in the WAN. With VATS enabled on Frame Relay PVCs, WAN interfaces are able to send at CIR when voice traffic is present on the link. However, when voice is not present, non-voice traffic is able to burst up to line speed and take advantage of the additional bandwidth that might be present in the WAN.

When VATS is used in combination with voice-adaptive fragmentation (VAF) (see Link Fragmentation and Interleaving (LFI), page 3-41), all non-voice traffic is fragmented and all traffic is shaped to the CIR of the WAN link when voice activity is detected on the interface.

As with VAF, exercise care when enabling VATS because activation can have an adverse effect on non-voice traffic. When voice is present on the link, data applications will experience decreased throughput because they are throttled back to well below CIR. This behavior will likely result in packet drops and delays for non-voice traffic. Furthermore, after voice traffic is no longer detected, the deactivation timer (default of 30 seconds) must expire before traffic can burst back to line speed. It is important, when using VATS, to set end-user expectations and make them aware that data applications will experience slowdowns on a regular basis due to the presence of voice calls across the WAN. VATS is available in Cisco IOS Release 12.2(15)T and later.

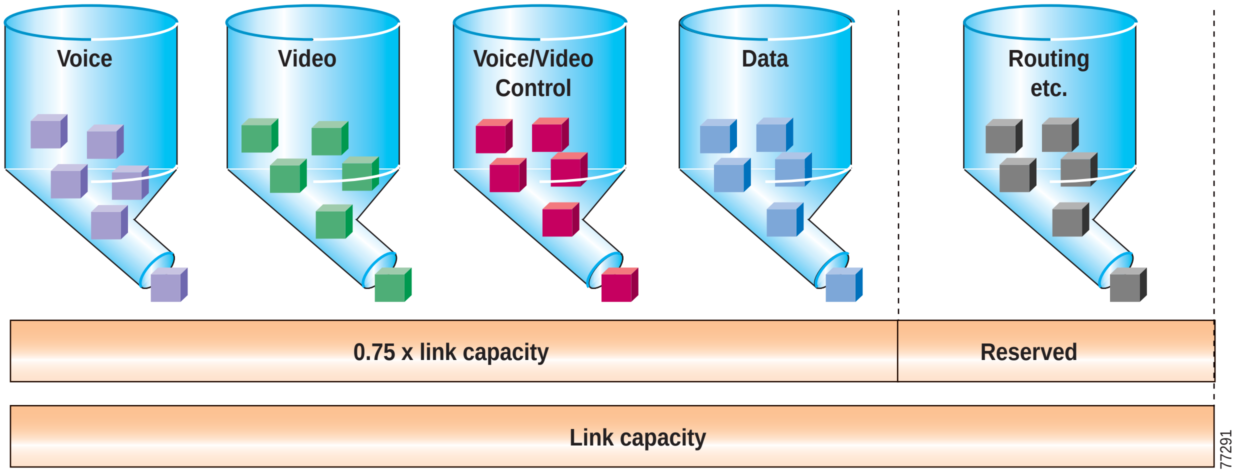
For more information on the Voice-Adaptive Traffic Shaping and Fragmentation features and how to configure them, refer to the documentation at

http://www.cisco.com/en/US/docs/ios/12_2t/12_2t15/feature/guide/ft_vats.html

Bandwidth Provisioning

Properly provisioning the network bandwidth is a major component of designing a successful IP network. You can calculate the required bandwidth by adding the bandwidth requirements for each major application (for example, voice, video, and data). This sum then represents the minimum bandwidth requirement for any given link, and it should not exceed approximately 75% of the total available bandwidth for the link. This 75% rule assumes that some bandwidth is required for overhead traffic, such as routing and Layer 2 keep-alives. Figure 3-15 illustrates this bandwidth provisioning process.

Figure 3-15 Link Bandwidth Provisioning



In addition to using no more than 75% of the total available bandwidth for data, voice, and video, the total bandwidth configured for all LLQ priority queues should typically not exceed 33% of the total link bandwidth. Provisioning more than 33% of the available bandwidth for the priority queue can be problematic for a number of reasons. First, provisioning more than 33% of the bandwidth for voice can result in increased CPU usage. Because each voice call will send 50 packets per second (with 20 ms samples), provisioning for large numbers of calls in the priority queue can lead to high CPU levels due to high packet rates. In addition, if more than one type of traffic is provisioned in the priority queue (for example, voice and video), this configuration defeats the purpose of enabling QoS because the priority queue essentially becomes a first-in, first-out (FIFO) queue. A larger percentage of reserved priority bandwidth effectively dampens the QoS effects by making more of the link bandwidth FIFO. Finally, allocating more than 33% of the available bandwidth can effectively starve any data queues that are provisioned. Obviously, for very slow links (less than 192 kbps), the recommendation to provision no more than 33% of the link bandwidth for the priority queue(s) might be unrealistic because a single call could require more than 33% of the link bandwidth. In these situations, and in situations where specific business needs cannot be met while holding to this recommendation, it may be necessary to exceed the 33% rule.

From a traffic standpoint, an IP telephony call consists of two parts:

- The voice and video bearer streams, which consists of Real-Time Transport Protocol (RTP) packets that contain the actual voice samples.
- The call control signaling, which consists of packets belonging to one of several protocols, according to the endpoints involved in the call (for example, H.323, MGCP, SCCP, or (J)TAPI). Call control functions are, for instance, those used to set up, maintain, tear down, or redirect a call.

Bandwidth provisioning should include not only the bearer traffic but also the call control traffic. In fact, in multisite WAN deployments, the call control traffic (as well as the bearer traffic) must traverse the WAN, and failure to allocate sufficient bandwidth for it can adversely affect the user experience.

The next three sub-sections describe the bandwidth provisioning recommendations for the following types of traffic:

- Voice and video bearer traffic in all multisite WAN deployments (see Provisioning for Bearer Traffic, page 3-46)
- Call control traffic in multisite WAN deployments with centralized call processing (see .Provisioning for Call Control Traffic with Centralized Call Processing, page 3-49)
- Call control traffic in multisite WAN deployments with distributed call processing (see Provisioning for Call Control Traffic with Distributed Call Processing, page 3-53)

Provisioning for Bearer Traffic

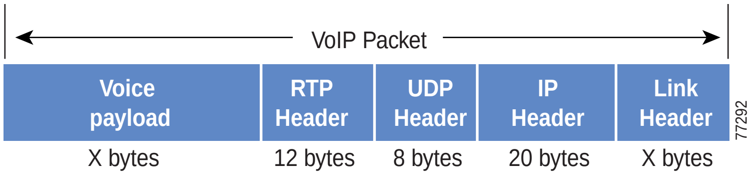
The section describes bandwidth provisioning for the following types of traffic:

- Voice Bearer Traffic, page 3-46
- Video Bearer Traffic, page 3-49

Voice Bearer Traffic

As illustrated in Figure 3-16, a voice-over-IP (VoIP) packet consists of the voice payload, IP header, User Datagram Protocol (UDP) header, Real-Time Transport Protocol (RTP) header, and Layer 2 Link header. When Secure Real-Time Transport Protocol (SRTP) encryption is used, the voice payload for each packet is increased by 4 bytes. The link header varies in size according to the Layer 2 media used.

Figure 3-16 Typical VoIP Packet



The bandwidth consumed by VoIP streams is calculated by adding the packet payload and all headers (in bits), then multiplying by the packet rate per second, as follows:

Layer 2 bandwidth in kbps = [(Packets per second) ∗ (*X* bytes for voice payload + 40 bytes for RTP/UDP/IP headers + *Y* bytes for Layer 2 overhead) ∗ 8 bits] / 1000

Layer 3 bandwidth in kbps = [(Packets per second) ∗ (*X* bytes for voice payload + 40 bytes for RTP/UDP/IP headers) ∗ 8 bits] / 1000

Packets per second = [1/(sampling rate in msec)] ∗ 1000

Voice payload in bytes = [(codec bit rate in kbps) ∗ (sampling rate in msec)] / 8

Table 3-9 details the Layer 3 bandwidth per VoIP flow. Table 3-9 lists the bandwidth consumed by the voice payload and IP header only, at a default packet rate of 50 packets per second (pps) and at a rate of 33.3 pps for both non-encrypted and encrypted payloads. Table 3-9 does not include Layer 2 header overhead and does not take into account any possible compression schemes, such as compressed Real-Time Transport Protocol (cRTP). You can use the Service Parameters menu in Unified CM Administration to adjust the codec sampling rate.

Table 3-9 Bandwidth Consumption for Voice Payload and IP Header Only

CODEC	Sampling Rate	Voice Payload in Bytes	Packets per Second	Bandwidth per Conversation
G.711 and G.722-64k	20 ms	160	50.0	80.0 kbps
G.711 and G.722-64k (SRTP)	20 ms	164	50.0	81.6 kbps
G.711 and G.722-64k	30 ms	240	33.3	74.7 kbps
G.711 and G.722-64k (SRTP)	30 ms	244	33.3	75.8 kbps
iLBC	20 ms	38	50.0	31.2 kbps
iLBC (SRTP)	20 ms	42	50.0	32.8 kbps
iLBC	30 ms	50	33.3	24.0 kbps
iLBC (SRTP)	30 ms	54	33.3	25.1 kbps
G.729A	20 ms	20	50.0	24.0 kbps
G.729A (SRTP)	20 ms	24	50.0	25.6 kbps
G.729A	30 ms	30	33.3	18.7 kbps
G.729A (SRTP)	30 ms	34	33.3	19.8 kbps

A more accurate method for provisioning is to include the Layer 2 headers in the bandwidth calculations. Table 3-10 lists the amount of bandwidth consumed by voice traffic when the Layer 2 headers are included in the calculations.

Table 3-10 Bandwidth Consumption with Layer 2 Headers Included

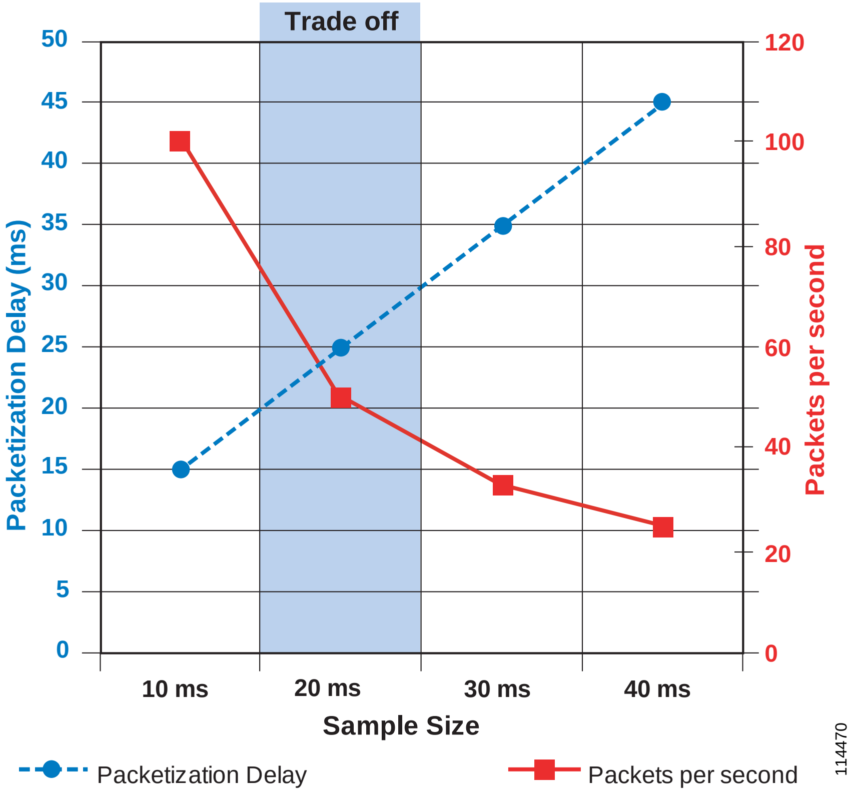
	Header Type and Size						
CODEC	Ethernet 14 Bytes	PPP 6 Bytes	ATM 53-Byte Cells with a 48-Byte Payload	Frame Relay 4 Bytes	MLPPP 10 Bytes	MPLS 4 Bytes	WLAN 24 Bytes
G.711 and G.722-64k at 50.0 pps	85.6 kbps	82.4 kbps	106.0 kbps	81.6 kbps	84.0 kbps	81.6 kbps	89.6 kbps
G.711 and G.722-64k (SRTP) at 50.0 pps	87.2 kbps	84.0 kbps	106.0 kbps	83.2 kbps	85.6 kbps	83.2 kbps	N/A
G.711 and G.722-64k at 33.3 pps	78.4 kbps	76.3 kbps	84.8 kbps	75.7 kbps	77.3 kbps	75.7 kbps	81.1 kbps
G.711 and G.722-64k (SRTP) at 33.3 pps	79.5 kbps	77.4 kbps	84.8 kbps	76.8 kbps	78.4 kbps	76.8 kbps	N/A
iLBC at 50.0 pps	36.8 kbps	33.6 kbps	42.4 kbps	32.8 kbps	35.2 kbps	32.8 kbps	40.8 kbps
iLBC (SRTP) at 50.0 pps	38.4 kbps	35.2 kbps	42.4 kbps	34.4 kbps	36.8 kbps	34.4 kbps	42.4 kbps
iLBC at 33.3 pps	27.7 kbps	25.6 kbps	28.3 kbps	25.0 kbps	26.6 kbps	25.0 kbps	30.4 kbps
iLBC (SRTP) at 33.3 pps	28.8 kbps	26.6 kbps	42.4 kbps	26.1 kbps	27.7 kbps	26.1 kbps	31.5 kbps
G.729A at 50.0 pps	29.6 kbps	26.4 kbps	42.4 kbps	25.6 kbps	28.0 kbps	25.6 kbps	33.6 kbps

Table 3-10 Bandwidth Consumption with Layer 2 Headers Included (continued)

	Header Type and Size						
CODEC	Ethernet 14 Bytes	PPP 6 Bytes	ATM 53-Byte Cells with a 48-Byte Payload	Frame Relay 4 Bytes	MLPPP 10 Bytes	MPLS 4 Bytes	WLAN 24 Bytes
G.729A (SRTP) at 50.0 pps	31.2 kbps	28.0 kbps	42.4 kbps	27.2 kbps	29.6 kbps	27.2 kbps	35.2 kbps
G.729A at 33.3 pps	22.4 kbps	20.3 kbps	28.3 kbps	19.7 kbps	21.3 kbps	19.8 kbps	25.1 kbps
G729A (SRTP) at 33.3 pps	23.5 kbps	21.4 kbps	28.3 kbps	20.8 kbps	22.4 kbps	20.8 kbps	26.2 kbps

While it is possible to configure the sampling rate above 30 ms, doing so usually results in very poor voice quality. As illustrated in Figure 3-17, as sampling size increases, the number of packets per second decreases, resulting in a smaller impact to the CPU of the device. Likewise, as the sample size increases, IP header overhead is lower because the payload per packet is larger. However, as sample size increases, so does packetization delay, resulting in higher end-to-end delay for voice traffic. The trade-off between packetization delay and packets per second must be considered when configuring sample size. While this trade-off is optimized at 20 ms, 30 ms sample sizes still provide a reasonable ratio of delay to packets per second; however, with 40 ms sample sizes, the packetization delay becomes too high.

Figure 3-17 Voice Sample Size: Packets per Second vs. Packetization Delay



Video Bearer Traffic

For audio, it is relatively easy to calculate a percentage of overhead per packet given the sample size of each packet. For video, however, it is nearly impossible to calculate an exact percentage of overhead because the payload varies depending upon how much motion is present in the video (that is, how many pixels changed since the last frame).

To resolve this inability to calculate the exact overhead ratio for video, Cisco recommends that you add 20% to the call speed regardless of which type of Layer-2 medium the packets are traversing. The additional 20% gives plenty of headroom to allow for the differences between Ethernet, ATM, Frame Relay, PPP, HDLC, and other transport protocols, as well as some cushion for the bursty nature of video traffic.

Note that the call speed requested by the endpoint (for example, 128 kbps, 256 kbps, and so forth) represents the maximum burst speed of the call, with some additional amount for a cushion. The average speed of the call is typically much less than these values.

Provisioning for Call Control Traffic

When Unified Communications endpoints are separated from their call control application by a WAN, or when two interconnected Unified Communications systems are separated by a WAN, consideration must be given to the amount of bandwidth that must be provisioned for call control and signaling traffic between these endpoints and systems. This section discusses WAN bandwidth provisioning for call signaling traffic where centralized or distributed call processing models are deployed. For more information on Unified Communications centralized and distributed call processing deployment models, see Unified Communications Deployment Models, page 5-1.

.Provisioning for Call Control Traffic with Centralized Call Processing

In a centralized call processing deployment, the Unified CM cluster and the applications (such as voicemail) are located at the central site, while several remote sites are connected through an IP WAN. The remote sites rely on the centralized Unified CMs to handle their call processing.

The following considerations apply to this deployment model:

- Each time a remote branch phone places a call, the control traffic traverses the IP WAN to reach the Unified CM at the central site, even if the call is local to the branch.
- The signaling protocols that may traverse the IP WAN in this deployment model are SCCP (encrypted and non-encrypted), SIP (encrypted and non-encrypted), H.323, MGCP, and CTI-QBE. All the control traffic is exchanged between a Unified CM at the central site and endpoints or gateways at the remote branches.
- If RSVP is deployed within the cluster, the control traffic between the Unified CM cluster at the central site and the Cisco RSVP Agents at the remote sites uses the SCCP protocol.

As a consequence, you must provision bandwidth for control traffic that traverses the WAN between the branch routers and the WAN aggregation router at the central site.

The control traffic that traverses the WAN in this scenario can be split into two categories:

- Quiescent traffic, which consists of keep-alive messages periodically exchanged between the branch endpoints (phones, gateways, and Cisco RSVP Agents) and Unified CM, regardless of call activity. This traffic is a function of the quantity of endpoints.
- Call-related traffic, which consists of signaling messages exchanged between the branch endpoints and the Unified CM at the central site when a call needs to be set up, torn down, forwarded, and so forth. This traffic is a function of the quantity of endpoints and their associated call volume.

To obtain an estimate of the generated call control traffic, it is necessary to make some assumptions regarding the average number of calls per hour made by each branch IP phone. In the interest of simplicity, the calculations in this section assume an average of 10 calls per hour per phone.

Note If this average number does not satisfy the needs of your specific deployment, you can calculate the recommended bandwidth by using the advanced formulas provided in Advanced Formulas, page 3-51.

Given the assumptions made, and initially considering the case of a remote branch with no signaling encryption configured, the recommended bandwidth needed for call control traffic can be obtained from the following formula:

Equation 1A: Recommended Bandwidth Needed for SCCP Control Traffic without Signaling Encryption.

Bandwidth (bps) = 265 ∗ (Number of IP phones and gateways in the branch)

Equation 1B: Recommended Bandwidth Needed for SIP Control Traffic without Signaling Encryption.

Bandwidth (bps) = 538 ∗ (Number of IP phones and gateways in the branch)

If a site features a mix of SCCP and SIP endpoints, the two equations above should be employed separately for the quantity of each type of phone used, and the results added.

Equation 1 and all other formulas within this section include a 25% over-provisioning factor. Control traffic has a bursty nature, with peaks of high activity followed by periods of low activity. For this reason, assigning just the minimum bandwidth required to a control traffic queue can result in undesired effects such as buffering delays and, potentially, packet drops during periods of high activity. The default queue depth for a Class-Based Weighted Fair Queuing (CBWFQ) queue in Cisco IOS equals 64 packets. The bandwidth assigned to this queue determines its servicing rate. Assuming that the bandwidth configured is the average bandwidth consumed by this type of traffic, it is clear that, during the periods of high activity, the servicing rate will not be sufficient to "drain" all the incoming packets out of the queue, thus causing them to be buffered. Note that, if the 64-packet limit is reached, any subsequent packets are either assigned to the best-effort queue or are dropped. It is therefore advisable to introduce this 25% over-provisioning factor to absorb and smooth the variations in the traffic pattern and to minimize the risk of a temporary buffer overrun. This is equivalent to increasing the servicing rate of the queue.

If encryption is configured, the recommended bandwidth is affected because encryption increases the size of signaling packets exchanged between Unified CM and the endpoints. The following formula takes into account the impact of signaling encryption:

Equation 2A: Recommended Bandwidth Needed for SCCP Control Traffic with Signaling Encryption.

Bandwidth with signaling encryption (bps) = 415 ∗ (Number of IP phones and gateways in the branch)

Equation 2B: Recommended Bandwidth Needed for SIP Control Traffic with Signaling Encryption.

Bandwidth with signaling encryption (bps) = 619 ∗ (Number of IP phones and gateways in the branch)

If we now take into account the fact that the smallest bandwidth that can be assigned to a queue on a Cisco IOS router is 8 kbps, we can summarize the values of minimum and recommended bandwidth for various branch office sizes, as shown in Table 3-11.

Table 3-11 Recommended Layer 3 Bandwidth for Call Control Traffic With and Without Signaling Encryption

Branch Office Size (Number of IP Phones and Gateways)	Recommended Bandwidth for SCCP Control Traffic (no encryption)	Recommended Bandwidth for SCCP Control Traffic (with encryption)	Recommended Bandwidth for SIP Control Traffic (no encryption)	Recommended Bandwidth for SIP Control Traffic (with encryption)
1 to 10	8 kbps	8 kbps	8 kbps	8 kbps
20	8 kbps	9 kbps	11 kbps	12 kbps
30	8 kbps	13 kbps	16 kbps	19 kbps
40	11 kbps	17 kbps	22 kbps	25 kbps
50	14 kbps	21 kbps	27 kbps	31 kbps
100	27 kbps	42 kbps	54 kbps	62 kbps
150	40 kbps	62 kbps	81 kbps	93 kbps

Note Table 3-11 assumes 10 calls per hour per phone, and it does not include RSVP control traffic. To determine the RSVP-related bandwidth to add to the values in this table, see Considerations for Calls Using RSVP, page 11-62.

Note If an RSVP-based locations policy is used for inter-site calls, the values of Table 3-11 must be increased to compensate for the control traffic of the Cisco RSVP Agent. For example, if 10% of the calls go over the WAN, multiply the value from Table 3-11 by 1.1.

Advanced Formulas

The previous formulas presented in this section assume an average call rate per phone of 10 calls per hour. However, this rate might not correspond to your deployment if the call patterns are significantly different (for example, with call center agents at the branches). To calculate call control bandwidth requirements in these cases, use the following formulas, which contain an additional variable (CH) that represents the average calls per hour per phone:

Equation 3A: Recommended Bandwidth Needed for SCCP Control Traffic for a Branch with No Signaling Encryption.

Bandwidth (bps) = (53 + 21 ∗ CH) ∗ (Number of IP phones and gateways in the branch)

Equation 3B: Recommended Bandwidth Needed for SIP Control Traffic for a Branch with No Signaling Encryption.

Bandwidth (bps) = (138 + 40 ∗ CH) ∗ (Number of IP phones and gateways in the branch)

Equation 4A: Recommended Bandwidth Needed for SCCP Control Traffic for a Remote Branch with Signaling Encryption.

Bandwidth with signaling encryption (bps) = (73.5 + 33.9 ∗ CH) ∗ (Number of IP phones and gateways in the branch)

Equation 4B: Recommended Bandwidth Needed for SIP Control Traffic for a Remote Branch with Signaling Encryption.

Bandwidth with signaling encryption (bps) = (159 + 46 ∗ CH) ∗ (Number of IP phones and gateways in the branch)

Note Equations 3A and 4A are based on the default SCCP keep-alive period of 30 seconds, while equations 3B and 4B are based on the default SIP keep-alive period of 120 seconds.

Considerations for Shared Line Appearances

Calls placed to shared line appearances, or calls sent to line groups using the Broadcast distribution algorithm, have two net effects on the bandwidth consumed by the system:

- Because all the phones on which the line is configured ring simultaneously, they represent a load on the system corresponding to a much higher calls-per-hour (CH) value than the CH of the line. The corresponding bandwidth consumption is therefore increased. The network infrastructure's bandwidth provisioning requires adjustments when WAN-connected shared line functionality is deployed. The CH value employed for Equations 3 and 4 must be increased according to the following formula:

  CHS = CHL ∗ (Number line appearances) / (Number of lines)

  Where CHS is the shared-line calls per hour to be used in Equations 3 and 4, and CHL is the calls-per-hour rating of the line. For example, if a site is configured with 5 lines making an average of 6 calls per hour but 2 of those lines are shared across 4 different phones, then:

  Number of lines = 5

  Number of line appearances = (2 lines appear on 4 phones, and 3 lines appear on only one phone) = (2∗4) + 3 = 11 line appearances

  CHL = 6

  CHS = 6 ∗ (11 / 5) = 13.2

- Because each of the ringing phones requires a separate signaling control stream, the quantity of packets sent from Unified CM to the same branch is increased in linear proportion to the quantity of phones ringing. Because Unified CM is attached to the network through a 100 Mbps or larger interface, it can instantaneously generate a very large quantity of packets that must be buffered while the queuing mechanism is servicing the signaling traffic. The servicing speed is limited by the WAN interface's effective information transfer speed, which is typically two orders of magnitude smaller than 100 Mbps.

  This traffic may overwhelm the queue depth of the central site's WAN router. By default, the queue depth available for each of the classes of traffic in Cisco IOS is 64. In order to prevent any packets from being dropped before they are queued for the WAN interface, you must ensure that the signaling queue's depth is sized to hold all the packets from at least one full shared-line event for each shared-line phone. Avoiding drops is paramount in ensuring that the call does not create a race condition where dropped packets are retransmitted, causing system response times to suffer.

  Therefore, the quantity of packets required to operate shared-line phones is as follows:

  – SCCP protocol: 13 packets per shared-line phone

  – SIP protocol: 11 packets per shared-line phone

  For example, with SCCP and with 6 phones sharing the same line, the queue depth for the signaling class of traffic must be adjusted to a minimum of 78. Table 3-12 provides recommended queue depths based on the quantity of shared line appearances within a branch site.

Table 3-12 Recommended Queue Depth per Branch Site

Number of Shared Line Appearances	Queue Depth (Packets)	
	SCCP	SIP
5	65	55
10	130	110
15	195	165
20	260	220
25	325	275

When using a Layer 2 WAN technology such as Frame Relay, this adjustment must be made on the circuit corresponding to the branch where the shared-line phones are located.

When using a Layer 3 WAN technology such as MPLS, there may be a single signaling queue servicing multiple branches. In this case, adjustment must be made for the total of all branches serviced.

Provisioning for Call Control Traffic with Distributed Call Processing

In distributed call processing deployments, several sites are connected through an IP WAN. Each site contains a Unified CM cluster and can follow either the single-site model or the centralized call processing model. A gatekeeper may be used for call admission control between sites.

The following considerations apply to this deployment model:

- The signaling protocol used to place a call across the WAN is H.323 or SIP.
- Control traffic is exchanged between the Cisco IOS gatekeeper and the Unified CM clusters at each site, as well as between the Unified CM clusters themselves.

Therefore, bandwidth for control traffic must be provisioned on the WAN links between Unified CMs as well as between each Unified CM and the gatekeeper. Because the topology is limited to hub-and-spoke, with the gatekeeper typically located at the hub, the WAN link that connects each site to the other sites usually coincides with the link that connects the site to the gatekeeper.

The control traffic that traverses the WAN belongs to one of the following categories:

- Quiescent traffic, which consists of registration messages periodically exchanged between each Unified CM and the gatekeeper
- Call-related traffic, which in turn consists of two types of traffic:
  - Call admission control traffic, exchanged between the Unified CMs and the call admission control device (such as a gatekeeper or Cisco RSVP Agent) before a call can be set up and after it has been torn down.
  - Signaling traffic associated with a media stream, exchanged over an intercluster trunk when a call needs to be set up, torn down, forwarded, and so on.

Because the total amount of control traffic depends on the number of calls that are set up and torn down at any given time, it is necessary to make some assumptions about the call patterns and the link utilization. The WAN links that connect each of the spoke sites to the hub site are normally provisioned to accommodate different types of traffic (for example, data, voice, and video). Using a traditional telephony analogy, we can view the portion of the WAN link that has been provisioned for voice as a number of *virtual tie lines*.

Assuming an average call duration of 2 minutes and 100 percent utilization of each virtual tie line, we can derive that each tie line carries a volume of 30 calls per hour. This assumption allows us to obtain the following formula that expresses the recommended bandwidth for call control traffic as a function of the number of virtual tie lines.

Equation 6: Recommended Bandwidth Based on Number of Virtual Tie Lines.

Recommended Bandwidth (bps) = 116 $*$ (Number of virtual tie lines)

If we take into account the fact that 8 kbps is the smallest bandwidth that can be assigned to a queue on a Cisco IOS router, we can deduce that a minimum queue size of 8 kbps can accommodate the call control traffic generated by *up to 70 virtual tie lines*. This amount should be sufficient for most large enterprise deployments.

Wireless LAN Infrastructure

Wireless LAN infrastructure design becomes important when Unified Communications is added to the wireless LAN (WLAN) portions of a converged network. With the introduction of Cisco Unified Wireless endpoints, voice and video traffic has moved onto the WLAN and is now converged with the existing data traffic there. Just as with wired LAN and wired WAN infrastructure, the addition of voice and video in the WLAN requires following basic configuration and design best-practices for deploying a highly available network. In addition, proper WLAN infrastructure design requires understanding and deploying QoS on the wireless network to ensure end-to-end voice and video quality on the entire network. The following sections discuss these requirements:

- Architecture for Voice and Video over WLAN, page 3-54
- High Availability for Voice and Video over WLAN, page 3-58
- Capacity Planning for Voice and Video over WLAN, page 3-60
- Design Considerations for Voice and Video over WLAN, page 3-60

For more information about Voice over Wireless LANs, refer to the latest version of the *Voice over Wireless LAN Design Guide*, available at

http://www.cisco.com/en/US/solutions/ns340/ns414/ns742/ns820/landing_voice_wireless.html

Architecture for Voice and Video over WLAN

IP telephony architecture has used wired devices since its inception, but enterprise users have long sought the ability to communicate while moving through the company premises. Wireless IP networks have enabled IP telephony to deliver enterprise mobility by providing on-premises roaming communications to the users with wireless IP telephony devices.

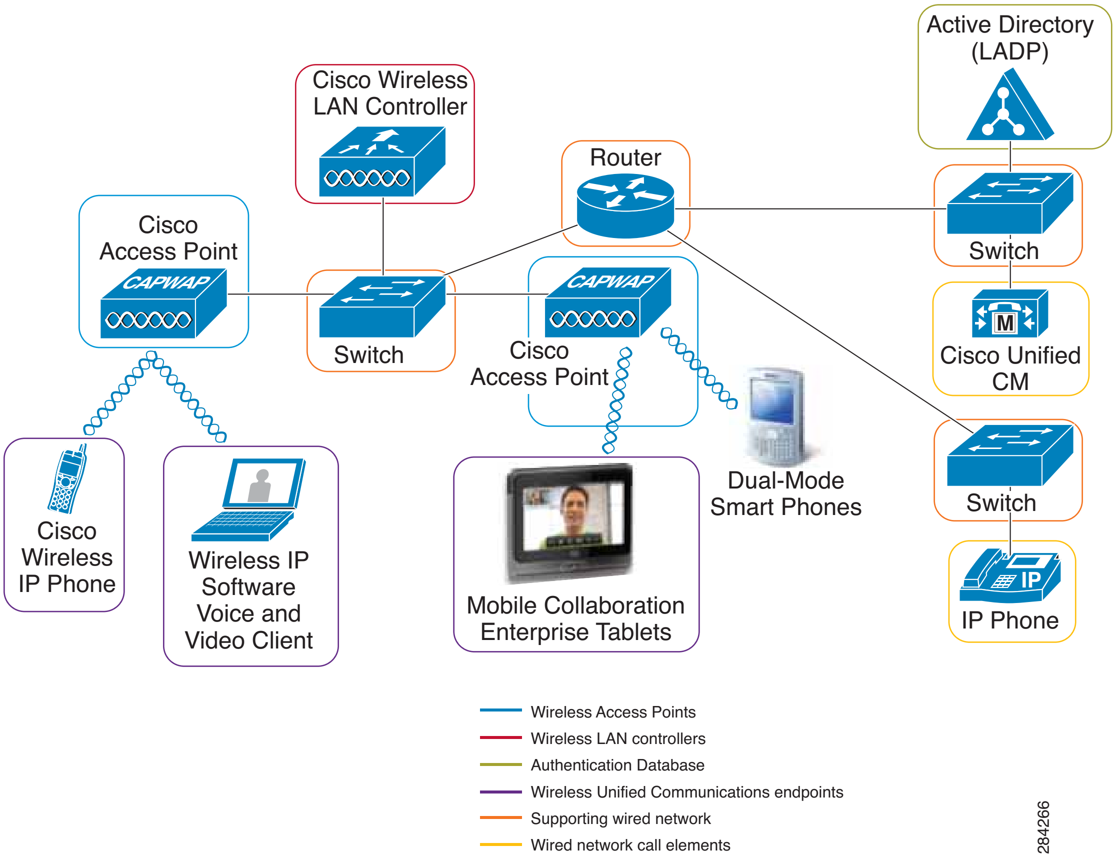
Wireless IP telephony and wireless IP video telephony are extensions of their wired counterparts, which leverage the same call elements. Additionally, wireless IP telephony and IP video telephony take advantage of wireless 802.11-enabled media, thus providing a cordless IP voice and video experience. The cordless experience is achieved by leveraging the wireless network infrastructure elements for the transmission and reception of the control and media packets.

The architecture for voice and video over wireless LAN includes the following basic elements, illustrated in Figure 3-18:

- Wireless Access Points, page 3-55
- Wireless LAN Controllers, page 3-56

- Authentication Database, page 3-56
- Supporting Wired Network, page 3-57
- Wireless Unified Communications Endpoints, page 3-57
- Wired Call Elements, page 3-57

Figure 3-18 Basic Layout for a Voice and Video Wireless Network

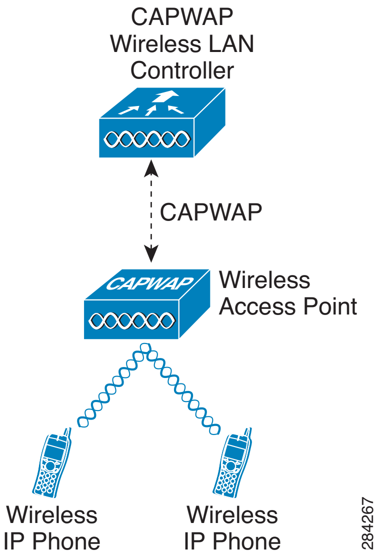


Wireless Access Points

The wireless access points enable wireless devices (Unified Communications endpoints in the case of voice and video over WLAN) to communicate with wired network elements. Access points function as adapters between the wired and wireless world, creating an entry-way between these two media. Cisco access points can be managed by a wireless LAN controller (WLC) or they can function in autonomous mode. When the access points are managed by a WLC they are referred as Lightweight Access Points, and in this mode they use the Lightweight Access Point Protocol (LWAPP) or Control and Provisioning of Wireless Access Points (CAPWAP) protocol, depending on the controller version, when communicating with the WLC.

Figure 3-19 illustrates the basic relationship between lightweight access points and WLCs. Although the example depicted in Figure 3-19 is for a CAPWAP WLC, from the traffic flow and relationship perspective there are no discernible differences between CAPWAP and LWAPP, so the example also applies to wireless LWAPP networks. Some advantages of leveraging WLCs and lightweight access points for the wireless infrastructure include ease of management, dynamic network tuning, and high availability. However, if you are using the managed mode instead of the autonomous mode in the access points, you need to consider the network tunneling effect of the LWAP-WLC communication architecture when designing your solution. This network tunneling effect is discussed in more depth in the section on Wireless LAN Controller Design Considerations, page 3-65.

Figure 3-19 Lightweight Access Point



Wireless LAN Controllers

Many corporate environments require deployment of wireless networks on a large scale. The wireless LAN controller (WLC) is a device that assumes a central role in the wireless network and helps to make it easier to manage such large-scale deployments. Traditional roles of access points, such as association or authentication of wireless clients, are done by the WLC. Access points, called Lightweight Access Points (LWAPs) in the Unified Communications environment, register themselves with a WLC and tunnel all the management and data packets to the WLCs, which then switch the packets between wireless clients and the wired portion of the network. All the configurations are done on the WLC. LWAPs download the entire configuration from WLCs and act as a wireless interface to the clients.

Authentication Database

The authentication database is a core component of the wireless networks, and it holds the credentials of the users to be authenticated while the wireless association is in progress. The authentication database provides the network administrators with a centralized repository to validate the credentials. Network administrators simply add the wireless network users to the authentication database instead of having to add the users to all the wireless access points with which the wireless devices might associate.

In a typical wireless authentication scenario, the WLC couples with the authentication database to allow the wireless association to proceed or fail. Authentication databases commonly used are LDAP and RADIUS, although under some scenarios the WLC can also store a small user database locally that can be used for authentication purposes.

Supporting Wired Network

The supporting wired network is the portion of the system that serves as a path between WLCs, APs, and wired call elements. Because the APs need or might need to communicate to the wired world, part of the wired network has to enable those communications. The supporting wired network consists of the switches, routers, and wired medium (WAN links and optical links) that work together to communicate with the various components that form the architecture for voice and video over WLAN.

Wireless Unified Communications Endpoints

The wireless Unified Communications endpoints are the components of the architecture for voice and video over WLAN that users employ to communicate with each other. These endpoints can be voice-only or enabled for both voice and video. When end users employ the wireless communications endpoints to call a desired destination, the endpoints in turn forward the request to their associated call processing server. If the call is allowed, the endpoints process the voice or video, encode it, and send it to the receiving device or the next hop of processing. Typical Cisco wireless Unified Communications endpoints are wireless IP phones, voice and video software clients running on desktop computers, mobile smart phones connected through wireless media, and mobile collaboration enterprise tablets.

Wired Call Elements

Whether the wireless Unified Communications endpoints initiate a session between each other or with wired endpoints, wired call elements are involved in some way. Wired call elements are the supporting Unified Communications infrastructure (gateways and call processing entities), with voice and video endpoints coupled to that infrastructure.

Wired call elements are needed typically to address two requirements:

- Call Control, page 3-57
- Media Termination, page 3-57

Call Control

Cisco wireless Unified Communications endpoints require a call control or call processing server to route calls efficiently and to provide a feature-rich experience for the end users. The call processing entity resides somewhere in the wired network, either in the LAN or across a WAN.

Call control for the Cisco wireless Unified Communications endpoints is achieved through a call control protocol, either SIP or SCCP.

Media Termination

Media termination on wired Unified Communications endpoints occurs when the end users of the wireless Unified Communications endpoints communicate with IP phones, PSTN users, or video endpoints. Voice gateways, IP phones, video terminals, PBX trunks, and transcoders all serve as termination points for media when a user communicates through them. This media termination occurs by means of coding and decoding of the voice or video session for the user communication.

High Availability for Voice and Video over WLAN

Providing high availability in Unified Communications solutions is a critical requirement for meeting the modern demands of continuous connectivity. Unified Communications deployments designed for high availability increase reliability and up time. Using real-time applications such as voice or video over WLAN without high availability could have very adverse effects on the end user experience, including an inability to make voice or video calls.

Designing a solution for voice and video over WLAN with high availability requires focusing of the following main areas:

- Supporting Wired Network High Availability, page 3-58
- WLAN High Availability, page 3-58
- Call Processing High Availability, page 3-60

Supporting Wired Network High Availability

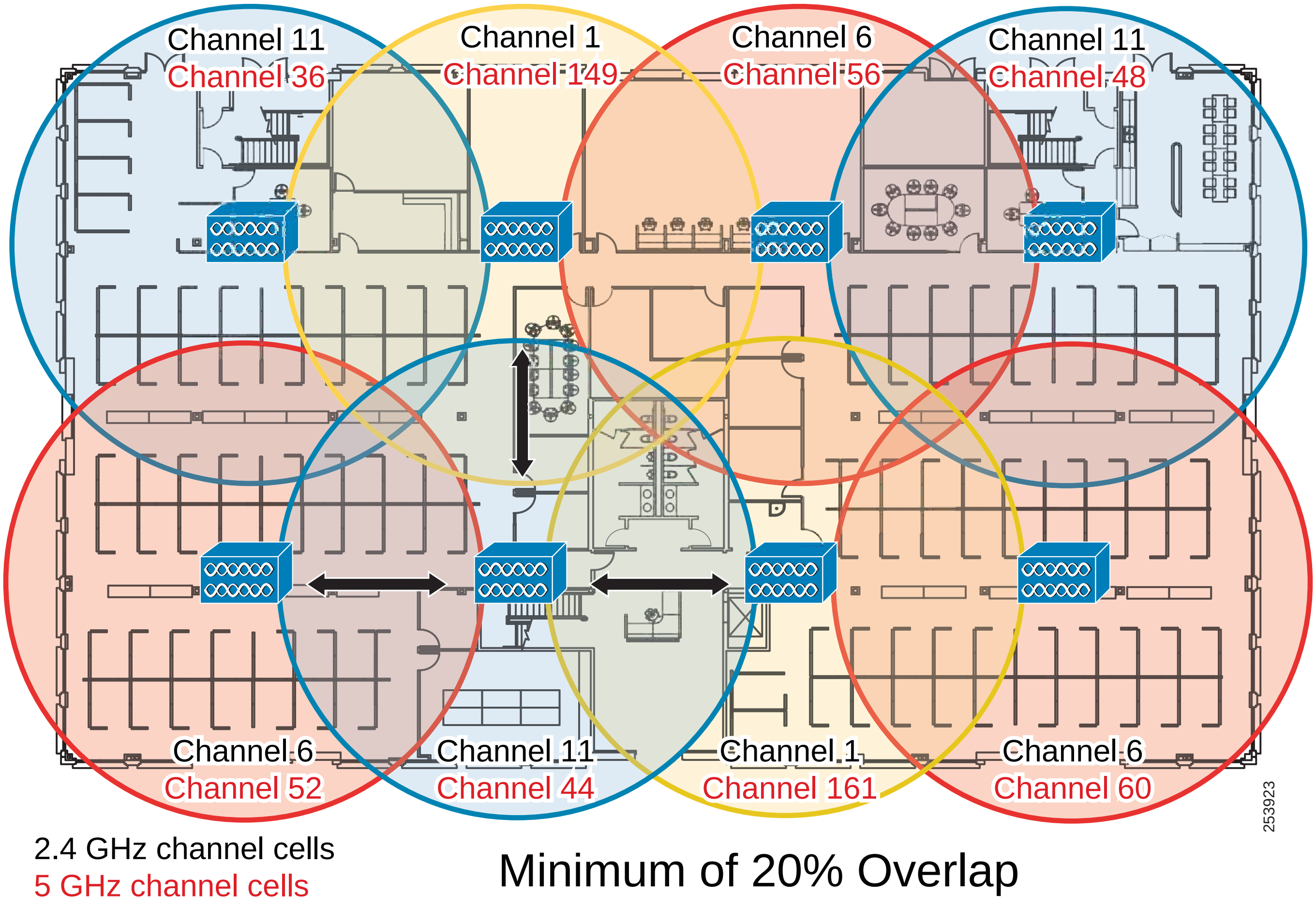
When deploying voice and video over WLAN, the same high-availability strategies used in wired networks can be applied to the wired components of the solution for voice and video over WLAN. For example, you can optimize layer convergence in the network to minimize disruption and take advantage of equal-cost redundant paths.

See LAN Design for High Availability, page 3-4, for further information about how to design highly available wired networks.

WLAN High Availability

A unique aspect of high availability for voice and video over WLAN is high availability of radio frequency (RF) coverage to provide Wi-Fi channel coverage that is not dependent upon a single WLAN radio. The Wi-Fi channel coverage is provided by the AP radios in the 2.4 GHz and 5 GHz frequency bands. The primary mechanism for providing RF high availability is cell boundary overlap. In general, a cell boundary overlap of 20% to 30% on non-adjacent channels is recommended to provide high availability in the wireless network. For mission-critical environments there should be at least two APs visible at the required signal level (-67 dBm or better). An overlap of 20% means that the RF cells of APs using non-adjacent channels overlap each other on 20% of their coverage area, while the remaining 80% of the coverage area is handled by a single AP. Figure 3-20 depicts a 20% overlap of AP non-adjacent channel cells to provide high availability. Furthermore, when determining the locations for installing the APs, avoid mounting them on reflective surfaces (such as metal, glass, and so forth), which could cause multi-path effects that result in signal distortion.

Figure 3-20 Non-Adjacent Channel Access Point Overlap



Careful deployment of APs and channel configuration within the wireless infrastructure are imperative for proper wireless network operation. For this reason, Cisco requires customers to conduct a complete and thorough site survey before deploying wireless networks in a production environment. The survey should include verifying non-overlapping channel configurations, Wi-Fi channel coverage, and required data and traffic rates; eliminating rogue APs; and identifying and mitigating the impact of potential interference sources.

Additionally, evaluate utilizing a 5 GHz frequency band, which is generally less crowded and thus usually less prone to interference. If Bluetooth is used then 5 GHz 802.11a is highly recommended. Similarly, the usage of Cisco CleanAir technology will increase the WLAN reliability by detecting radio frequency interference in real time and providing a self-healing and self-optimizing wireless network. For further information about Cisco CleanAir technology, refer to the product documentation available at

http://www.cisco.com/en/US/netsol/ns1070/index.html

For further information on how to provide high availability in a WLAN that supports rich media, refer to the *Voice over Wireless LAN Design Guide*, available at

http://www.cisco.com/en/US/solutions/ns340/ns414/ns742/ns820/landing_voice_wireless.html

Call Processing High Availability

For information regarding call processing resiliency, see High Availability for Call Processing, page 8-14.

Capacity Planning for Voice and Video over WLAN

A crucial piece in planning for voice and video over WLAN is adequately sizing the solution for the desired call capacity. Capacity is defined as the number of simultaneous voice and video sessions over WLAN that can be supported in a given area. Capacity can vary depending upon the RF environment, the Unified Communications endpoint features, and the WLAN system features. For instance, a solution using Cisco Unified Wireless IP Phones 7925G on a WLAN that provides optimized WLAN services (such as the Cisco Unified Wireless Network) would have a maximum call capacity of 27 simultaneous sessions per channel at a data rate of 24 Mbps or higher for both 802.11a and 802.11g. On the other hand, a similar solution using only Cisco Cius making video calls at 720p and a video rate of 2,500 kbps on a WLAN, where access points are configured as 802.11a/n with a data rate index of Modulation and Coding Scheme 7 in 40 MHz channels, would have a maximum capacity of 7 video calls (two bidirectional voice and video streams) per channel.

To achieve these capacities, there must be minimal wireless LAN background traffic and radio frequency (RF) utilization, and Bluetooth must be disabled in the devices. It is also important to understand that call capacities are established per non-overlapping channel because the limiting factor is the channel capacity and not the number of access points (APs).

The call capacity specified by the actual wireless Unified Communications endpoint should be used for deployment purposes because it is the supported capacity of that endpoint. For capacity information about the wireless endpoints, refer to the following documentation:

- Cisco Unified IP Phones 7900 Series Design Guides

  http://www.cisco.com/en/US/products/hw/phones/ps379/products_implementation_design_guides_list.html

- Cisco Unified IP Phones 9900 Series Deployment Guide

  http://www.cisco.com/en/US/products/ps10453/products_implementation_design_guides_list.html

- Cisco Cius Deployment Guide

  http://www.cisco.com/en/US/products/ps11156/products_implementation_design_guides_list.html

For further information about calculating call capacity in a WLAN, refer to the *Voice over Wireless LAN Design Guide*, available at

http://www.cisco.com/en/US/solutions/ns340/ns414/ns742/ns820/landing_voice_wireless.html

Design Considerations for Voice and Video over WLAN

This section provides additional design considerations for deploying Unified Communications endpoints over WLAN solutions. WLAN configuration specifics can vary depending on the voice or video WLAN devices being used and the WLAN design. The following sections provide general guidelines and best practices for designing the WLAN infrastructure:

- VLANs, page 3-61
- Roaming, page 3-61
- Wireless Channels, page 3-62

- Wireless Interference and Multipath Distortion, page 3-63
- Multicast on the WLAN, page 3-63
- Wireless AP Configuration and Design, page 3-64
- Wireless LAN Controller Design Considerations, page 3-65
- WAN Quality of Service (QoS), page 3-37

VLANs

Just as with a wired LAN infrastructure, when deploying voice or video in a wireless LAN, you should enable at least two virtual LANs (VLANs) at the Access Layer. The Access Layer in a wireless LAN environment includes the access point (AP) and the first-hop access switch. On the AP and access switch, you should configure both a native VLAN for data traffic and a voice VLAN (under Cisco IOS) or Auxiliary VLAN (under CatOS) for voice traffic. This auxiliary voice VLAN should be separate from all the other wired voice VLANs in the network. However, when the wireless clients (for example, smart phones or software rich-media clients) do not support the concept of an auxiliary VLAN, alternative packet marking strategies (for example, packet classification per port) must be applied to segregate the important traffic such as voice and video and treat it with priority. When deploying a wireless infrastructure, Cisco also recommends configuring a separate management VLAN for the management of WLAN APs. This management VLAN should not have a WLAN appearance; that is, it should not have an associated service set identifier (SSID) and it should not be directly accessible from the WLAN.

Roaming

To improve the user experience, Cisco recommends designing the cell boundary distribution with a 20% to 30% overlap of non-adjacent channels to facilitate seamless roaming of the wireless client between access points. Furthermore, when devices roam at Layer 3, they move from one AP to another AP across native VLAN boundaries. When the WLAN infrastructure consists of autonomous APs, a Cisco Wireless LAN Controller allows the Cisco Unified Wireless endpoints to keep their IP addresses and roam at Layer 3 while still maintaining an active call. Seamless Layer 3 roaming occurs only when the client is roaming within the same mobility group. For details about the Cisco Wireless LAN Controller and Layer 3 roaming, refer to the product documentation available at

http://www.cisco.com/en/US/products/hw/wireless/index.html

Seamless Layer 3 roaming for clients across a lightweight access point infrastructure is accomplished by WLAN controllers that use dynamic interface tunneling. Cisco Wireless Unified Communications endpoints that roam across WLAN controllers and VLANs can keep their IP address when using the same SSID and therefore can maintain an active call.

Note In dual-band WLANs (those with 2.4 GHz and 5 GHz bands), it is possible to roam between 802.11b/g and 802.11a with the same SSID, provided the client is capable of supporting both bands. However, this can cause gaps in the voice path. If Cisco Unified Wireless IP Phones 7921 or 7925 are used, make sure that firmware version 1.3(4) or higher is installed on the phones to avoid these gaps; otherwise use only one band for voice. (The Cisco Unified Wireless IP Phone 7926 provides seamless inter-band roaming from its first firmware version.)

Wireless Channels

Wireless endpoints and APs communicate by means of radios on particular channels. When communicating on one channel, wireless endpoints typically are unaware of traffic and communication occurring on other non-overlapping channels.

Optimal channel configuration for 2.4 GHz 802.11b/g/n requires a minimum of five-channel separation between configured channels to prevent interference or overlap between channels. Non-overlapping channels have 22 MHz of separation. Channel 1 is 2.412 GHz, channel 6 is 2.437 GHz, and channel 11 is 2.462 GHz. In North America, with allowable channels of 1 to 11, channels 1, 6, and 11 are the three usable non-overlapping channels for APs and wireless endpoint devices. However, in Europe where the allowable channels are 1 to 13, multiple combinations of five-channel separation are possible. Multiple combinations of five-channel separation are also possible in Japan, where the allowable channels are 1 to 14.

Optimal channel configuration for 5 GHz 802.11a and 802.11n requires a minimum of one-channel separation to prevent interference or overlap between channels. In North America, there are 20 possible non-overlapping channels: 36, 40, 44, 48, 52, 56, 60, 64, 100, 104, 108, 112, 116, 132, 136, 140, 149, 153, 157, and 161. Europe and Japan allow 16 possible non-overlapping channels: 36, 40, 44, 48, 52, 56, 60, 64, 100, 104, 108, 112, 116, 132, 136, and 140. Because of the larger set of non-overlapping channels, 802.11a and 5 Ghz 802.11n allow for more densely deployed WLANs; however, Cisco recommends not enabling all channels but using a 12-channel design instead.

Note that the 802.11a and 802.11n bands (when using channels operating at 5.25 to 5.725 GHz, which are 15 of the 24 possible channels) do require support for Dynamic Frequency Selection (DFS) and Transmit Power Control (TPC) on some channels in order to avoid interference with radar (military, satellite, and weather). Regulations require that channels 52 to 64, 100 to 116, and 132 to 140 support DFS and TPC. TPC ensures that transmissions on these channels are not powerful enough to cause interference. DFC monitors channels for radar pulses and, when it detects a radar pulse, DFC stops transmission on the channel and switches to a new channel.

AP coverage should be deployed so that no (or minimal) overlap occurs between APs configured with the same channel. Same- channel overlap should typically occur at 19 dBm of separation. However, proper AP deployment and coverage on non-overlapping channels requires a minimum overlap of 20%. This amount of overlap ensures smooth roaming for wireless endpoints as they move between AP coverage cells. Overlap of less than 20% can result in slower roaming times and poor voice quality.

Deploying wireless devices in a multi-story building such as an office high-rise or hospital introduces a third dimension to wireless AP and channel coverage planning. Both the 2.4 GHz and 5.0 GHz wave forms of 802.11 can pass through floors and ceilings as well as walls. For this reason, not only is it important to consider overlapping cells or channels on the same floor, but it is also necessary to consider channel overlap between adjacent floors. With the 2.4 GHz wireless spectrum limited to only three usable non-overlapping channels, proper overlap design can be achieved only through careful three-dimensional planning.

Note Careful deployment of APs and channel configuration within the wireless infrastructure are imperative for proper wireless network operation. For this reason, Cisco requires that a complete and thorough site survey be conducted before deploying wireless networks in a production environment. The survey should include verifying non-overlapping channel configurations, AP coverage, and required data and traffic rates; eliminating rogue APs; and identifying and mitigating the impact of potential interference sources.

Wireless Interference and Multipath Distortion

Interference sources within a wireless environment can severely limit endpoint connectivity and channel coverage. In addition, objects and obstructions can cause signal reflection and multipath distortion. Multipath distortion occurs when traffic or signaling travels in more than one direction from the source to the destination. Typically, some of the traffic arrives at the destination before the rest of the traffic, which can result in delay and bit errors in some cases. You can reduce the effects of multipath distortion by eliminating or reducing interference sources and obstructions, and by using diversity antennas so that only a single antenna is receiving traffic at any one time. Interference sources should be identified during the site survey and, if possible, eliminated. At the very least, interference impact should be alleviated by proper AP placement and the use of location-appropriate directional or omni-directional diversity radio antennas.

Possible interference and multipath distortion sources include:

- Other APs on overlapping channels
- Other 2.4 GHz and 5 Ghz devices, such as 2.4 GHz cordless phones, personal wireless network devices, sulphur plasma lighting systems, microwave ovens, rogue APs, and other WLAN equipment that takes advantage of the license-free operation of the 2.4 GHz and 5 Ghz bands
- Metal equipment, structures, and other metal or reflective surfaces such as metal I-beams, filing cabinets, equipment racks, wire mesh or metallic walls, fire doors and fire walls, concrete, and heating and air conditioning ducts
- High-power electrical devices such as transformers, heavy-duty electric motors, refrigerators, elevators, and elevator equipment
- High-power electrical devices such as transformers, heavy-duty electric motors, refrigerators, elevators and elevator equipment, and any other power devices that could cause electromagnetic interference (EMI)

Because Bluetooth-enabled devices use the same 2.4 GHz radio band as 802.11b/g/n devices, it is possible that Bluetooth and 802.11b/g/n devices can interfere with each other, thus resulting in connectivity issues. Due to the potential for Bluetooth devices to interfere with and disrupt 802.11b/g/n WLAN voice and video devices (resulting in poor voice quality, de-registration, call setup delays, and/or reduce per-channel-cell call capacity), Cisco recommends, when possible, that you deploy all WLAN voice and video devices on the 5 GHz Wi-Fi band using 802.11a and/or 802.11n protocols. By deploying wireless clients on the 5 Ghz radio band, you can avoid interference caused by Bluetooth devices. Additionally, Cisco CleanAir technology is recommended within the wireless infrastructure because it enables real-time interference detection. For more information about Cisco CleanAir technology, refer to the product documentation available at

http://www.cisco.com/en/US/netsol/ns1070/index.html

Note 802.11n can operate on both the 2.4 GHz and 5 GHz bands; however, Cisco recommends using 5 GHz for Unified Communications.

Multicast on the WLAN

By design, multicast does not have the acknowledgement level of unicast. According to 802.11 specifications, the access point must buffer all multicast packets until the next Delivery Traffic Indicator Message (DTIM) period is met. The DTIM period is a multiple of the beacon period. If the beacon period is 100 ms (typical default) and the DTIM value is 2, then the access point must wait up to 200 ms before transmitting a single buffered multicast packet. The time period between beacons (as a product of the DTIM setting) is used by battery-powered devices to go into power save mode temporarily. This power save mode helps the device conserve battery power.

Multicast on WLAN presents a twofold problem in which administrators must weigh multicast traffic quality requirements against battery life requirements. First, delaying multicast packets will negatively affect multicast traffic quality, especially for applications that multicast real-time traffic such as voice and video. In order to limit the delay of multicast traffic, DTIM periods should typically be set to a value of 1 so that the amount of time multicast packets are buffered is low enough to eliminate any perceptible delay in multicast traffic delivery. However, when the DTIM period is set to a value of 1, the amount of time that battery-powered WLAN devices are able to go into power save mode is shortened, and therefore battery life is shortened. In order to conserve battery power and lengthen battery life, DTIM periods should typically be set to a value of 2 or more.

For WLAN networks with no multicast applications or traffic, the DTIM period should be set to a value of 2 or higher. For WLAN networks where multicast applications are present, the DTIM period should be set to a value of 2 with a 100 ms beacon period whenever possible; however, if multicast traffic quality suffers or if unacceptable delay occurs, then the DTIM value should be lowered to 1. If the DTIM value is set to 1, administrators must keep in mind that battery life of battery-operated devices will be shortened significantly.

Before enabling multicast applications on the wireless network, Cisco recommends testing these applications to ensure that performance and behavior are acceptable.

For additional considerations with multicast traffic, see the chapter on Media Resources, page 17-1.

Wireless AP Configuration and Design

Proper AP selection, deployment, and configuration are essential to ensure that the wireless network handles voice traffic in a way that provides high-quality voice to the end users.

AP Selection

For recommends on deploying access points for wireless voice, refer to the documentation at http://www.cisco.com/en/US/products/ps5678/Products_Sub_Category_Home.html.

AP Deployment

The number of devices active with an AP affects the amount of time each device has access to the transport medium, the Wi-Fi channel. As the number of devices increases, the traffic contention increases. Associating more devices to the AP and the bandwidth of the medium can result in poor performance and slower response times for all the endpoint devices associated to the AP.

While there is no specific mechanism prior to Cisco Wireless LAN Controller release 7.2 to ensure that only a limited number of devices are associated to a single AP, system administrators can manage device-to-AP ratios by conducting periodic site surveys and analyzing user and device traffic patterns. If additional devices and users are added to the network in a particular area, additional site surveys should be conducted to determine whether additional APs are required to handle the number of endpoints that need to access the network.

Additionally, APs that support Cisco CleanAir technology should be considered because they provide the additional function of remote monitoring of the Wi-Fi channel.

AP Configuration

When deploying wireless voice, observe the following specific AP configuration requirements:

- Enable Address Resolution Protocol (ARP) caching.

  ARP caching is required on the AP because it enables the AP to answer ARP requests for the wireless endpoint devices without requiring the endpoint to leave power-save or idle mode. This feature results in extended battery life for the wireless endpoint devices.

- Enable Dynamic Transmit Power Control (DTPC) on the AP.

  This ensures that the transmit power of the AP matches the transmit power of the voice endpoints. Matching transmit power helps eliminate the possibility of one-way audio traffic. Voice endpoints adjust their transmit power based on the Limit Client Power (mW) setting of the AP to which they are associated.

- Assign a Service Set Identifier (SSID) to each VLAN configured on the AP.

  SSIDs enable endpoints to select the wireless VLAN they will use for sending and receiving traffic. These wireless VLANs and SSIDs map to wired VLANs. For voice endpoints, this mapping ensures priority queuing treatment and access to the voice VLAN on the wired network.

- Enable **QoS Element for Wireless Phones** on the AP.

  This feature ensures that the AP will provide QoS Basic Service Set (QBSS) information elements in beacons. The QBSS element provides an estimate of the channel utilization on the AP, and Cisco wireless voice devices use it to help make roaming decisions and to reject call attempts when loads are too high. The APs also provide 802.11e clear channel assessment (CCA) QBSS in beacons. The CCA-based QBSS values reflect true channel utilization.

- Configure two QoS policies on the AP, and apply them to the VLANs and interfaces.

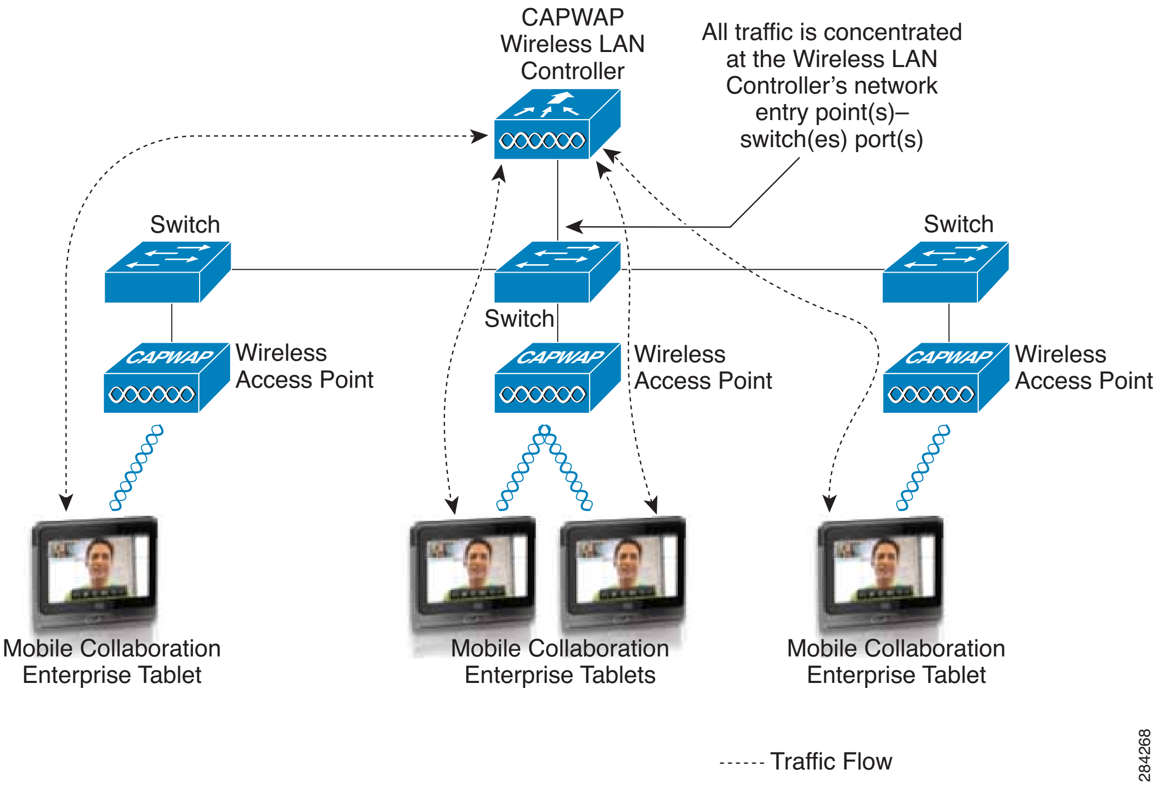
  To ensure that voice traffic is given priority queuing treatment, configure a voice policy and a data policy with default classifications for the respective VLANs. (See Interface Queuing, page 3-67, for more information).

Wireless LAN Controller Design Considerations

When designing a wireless network that will service voice or video, it is important to consider the role that the wireless LAN controller plays with regard to the voice and video media path if the access points used are not autonomous or stand alone. Because all wireless traffic is tunneled to its correspondent wireless LAN controller regardless of its point of origin and destination, it is critical to adequately size the network connectivity entry points of the wireless controllers. Figure 3-21 is a representation of this problem. If any Cisco Cius tries to call another Cius, the traffic has to be hairpinned in the wireless LAN controller and sent to the receiving device. This includes the scenario where both devices are associated to the same AP.

The switch ports where the wireless LAN controllers are connected should provide enough bandwidth coverage for the traffic generated by the Unified Communications devices, whether they are video or voice endpoints and whether their traffic is control or media traffic.

Figure 3-21 Traffic Concentrated at the Wireless LAN Controller Network Entry Point



Additionally, the switch interface and switch platform egress buffer levels should match the maximum combined burst you plan to support in your wireless network.

Failure to select adequate buffer levels could lead to packet drops and severely affect the user experience of video over a wireless LAN, while lack of bandwidth coverage would cause packets to be queued and in extreme cases cause delayed packets

WLAN Quality of Service (QoS)

Just as QoS is necessary for the LAN and WAN wired network infrastructure in order to ensure high voice quality, QoS is also required for the wireless LAN infrastructure. Because of the bursty nature of data traffic and the fact that real-time traffic such as voice and video are sensitive to packet loss and delay, QoS tools are required to manage wireless LAN buffers, limit radio contention, and minimize packet loss, delay, and delay variation.

However, unlike most wired networks, wireless networks are a shared medium, and wireless endpoints do not have dedicated bandwidth for sending and receiving traffic. While wireless endpoints can mark traffic with 802.1p CoS, ToS, DSCP, and PHB, the shared nature of the wireless network means limited admission control and access to the network for these endpoints.

Wireless QoS involves the following main areas of configuration:

- Traffic Classification, page 3-67
- User Priority Mapping, page 3-67
- Interface Queuing, page 3-67
- Wireless Call Admission Control, page 3-68

Traffic Classification

As with the wired network infrastructure, it is important to classify or mark pertinent wireless traffic as close to the edge of the network as possible. Because traffic marking is an entrance criterion for queuing schemes throughout the wired and wireless network, marking should be done at the wireless endpoint device whenever possible. Marking or classification by wireless network devices should be identical to that for wired network devices, as indicated in Table 3-13.

In accordance with traffic classification guidelines for wired networks, the Cisco wireless Unified Communications endpoints mark voice media traffic or voice RTP traffic with DSCP 46 (or PHB EF), video media traffic or video RTP traffic with DSCP 34 (or PHB AF41), and call control signaling traffic (SCCP or SIP) with DSCP 24 (or PHB CS3). Once this traffic is marked, it can be given priority or better than best-effort treatment and queuing throughout the network. All wireless voice and video devices that are capable of marking traffic should do it in this manner. All other traffic on the wireless network should be marked as best-effort or with some intermediary classification as outlined in wired network marking guidelines. If the wireless voice or video devices are unable to do packet marking, alternate methods such as port-based marking should be implemented to provide priority to video and voice traffic.

User Priority Mapping

While 802.1p and DSCP (Differentiated Service Code Point) are the standards to set priorities on wired networks, 802.11e is the standard used for wireless networks. This is commonly referred as User Priority (UP), and it is important to map the UP to its appropriate DSCP value. Table 3-13 lists the values for Unified Communications traffic.

Table 3-13 QoS Traffic Classification

Traffic Type	DSCP (PHB)	802.1p UP	IEEE 802.11e UP
Voice	46 (EF)	5	6
Video	34 (AF41)	4	5
Voice and video control	24 (CS3)	3	4

For further information about 802.11e and its configuration, refer to your corresponding product documentation available at

http://www.cisco.com/en/US/products/ps6302/Products_Sub_Category_Home.html

Interface Queuing

Once traffic marking has occurred, it is necessary to enable the wired network APs and devices to provide QoS queuing so that voice and video traffic types are given separate queues to reduce the chances of this traffic being dropped or delayed as it traverses the wireless LAN. Queuing on the wireless network occurs in two directions, upstream and downstream. Upstream queuing concerns traffic traveling from the wireless endpoint up to the AP, and from the AP up to the wired network. Downstream queuing concerns traffic traveling from the wired network to the AP and down to the wireless endpoint.

For upstream queuing, devices that support Wi-Fi Multimedia (WMM) are able to take advantage of queueing mechanisms, including priority queueing.

As for downstream QoS, Cisco APs currently provide up to eight queues for downstream traffic being sent to wireless clients. The entrance criterion for these queues can be based on a number of factors, including DSCP, access control lists (ACLs), and VLAN. Although eight queues are available, Cisco recommends using only two queues when deploying wireless voice. All voice media and signaling traffic should be placed in the highest-priority queue, and all other traffic should be placed in the best-effort queue. This ensures the best possible queuing treatment for voice traffic.

In order to set up this two-queue configuration for autonomous APs, create two QoS policies on the AP. Name one policy **Voice**, and configure it with the class of service **Voice < 10 ms Latency (6)** as the Default Classification for all packets on the VLAN. Name the other policy **Data**, and configure it with the class of service **Best Effort (0)** as the Default Classification for all packets on the VLAN. Then assign the Data policy to the incoming and outgoing radio interface for the data VLAN(s), and assign the Voice policy to the incoming and outgoing radio interfaces for the voice VLAN(s). With the QoS policies applied at the VLAN level, the AP is not forced to examine every packet coming in or going out to determine the type of queuing the packet should receive.

For lightweight APs, the WLAN controller has built-in QoS profiles that can provide the same queuing policy. Voice VLAN or voice traffic is configured to use the **Platinum** policy, which sets priority queueing for the voice queue. Data VLAN or data traffic is configured to use the **Silver** policy, which sets best-effort queuing for the Data queue. These policies are then assigned to the incoming and outgoing radio interfaces based on the VLAN.

The above configurations ensure that all voice and video media and signaling are given priority queuing treatment in a downstream direction.

Note Because Wi-Fi Multimedia (WMM) access is based on Enhanced Distributed Channel Access (EDCA), it is important to assign the right priorities to the traffic to avoid Arbitration Inter-Frame Space (AIFS) alteration and delivery delay. For further information on Cisco Unified Wireless QoS, refer to the *Enterprise Mobility Design Guide*, available at http://www.cisco.com/en/US/solutions/ns340/ns414/ns742/ns820/landing_ent_mob_design.html.

Wireless Call Admission Control

To avoid exceeding the capacity limit of a given AP channel, some form of call admission control is required. Cisco APs and wireless Unified Communications clients now use Traffic Specification (TSPEC) instead of QoS Basic Service Set (QBSS) for call admission control.

Wi-Fi Multimedia Traffic Specification (WMM TSPEC) is the QoS mechanism that enables WLAN clients to provide an indication of their bandwidth and QoS requirements so that APs can react to those requirements. When a client is preparing to make a call, it sends an Add Traffic Stream (ADDTS) message to the AP with which it is associated, indicating the TSPEC. The AP can then accept or reject the ADDTS request based on whether bandwidth and priority treatment are available. If the call is rejected, the client receives a Network Busy message. If the client is roaming, the TSPEC request is embedded in the re-association request message to the new AP as part of the association process, and the TSPEC response is embedded in the re-association response.

Alternatively, endpoints without WMM TSPEC support, but using SIP as call signaling, can be managed by the AP. Media snooping must be enabled for the service set identifier (SSID). The client's implementation of SIP must match that of the Wireless LAN Controller, including encryption and port numbers. For details about media snooping, refer to the *Cisco Wireless LAN Controller Configuration Guide*, available at

http://www.cisco.com/en/US/docs/wireless/controller/7.0/configuration/guide/c70wlan.html

Note Currently there is no call admission control support for video. The QoS Basic Service Set (QBSS) information element is sent by the AP only if **QoS Element for Wireless Phones** has been enable on the AP. (Refer to Wireless AP Configuration and Design, page 3-64.)

Service Advertisement Framework (SAF)

The Cisco Service Advertisement Framework (SAF) enables networking applications to advertise and discover information about networked services within an IP network. SAF consists of the following functional components and protocols:

- SAF Clients advertise and consume information about services.
- SAF Forwarders distribute and maintain SAF service availability information.
- SAF Client Protocol is used between SAF Clients and SAF Forwarders.
- SAF Forwarder Protocol is used between SAF Forwarders.

The nature of the advertised service is unimportant to the network of SAF Forwarders. The SAF Forwarder protocol is designed to dynamically distribute information about the availability of services to SAF client applications that have registered to the SAF network.

Services that SAF Can Advertise

In theory, any service can be advertised through SAF. The first service to use SAF is Cisco Unified Communications Call Control Discovery (CCD). CCD uses SAF to distribute and maintain information about the availability of internal directory numbers (DNs) hosted by call control agents such as Cisco Unified CM and Unified CME. CCD also distributes the corresponding number prefixes that allow these internal directory numbers to be reached from the PSTN ("To PSTN" prefixes).

The dynamic nature of SAF and the ability for call agents to advertise the availability of their hosted DN ranges and To PSTN prefixes to other call agents in a SAF network, provides distinct advantages over other static and more labor-intensive methods of dial plan distribution. For more information on SAF CCD, see Call Routing and Dial Plan Distribution Using Call Control Discovery for the Service Advertisement Framework, page 5-52.

SAF Networks

SAF networks contain a number of functional components, as described in the following sections.

SAF Forwarders, SAF Clients, and non-SAF Networks

In a Cisco SAF network, service information is distributed though a network of SAF-capable nodes that assume specific functions to efficiently distribute knowledge of services and facilitate their discovery. Cisco SAF network nodes are classified by two functional responsibilities:

- SAF Forwarder
- SAF Client

To configure a Cisco SAF network, you must configure both SAF Forwarders and SAF Clients. The flexibility of Cisco SAF allows you to configure a single edge router to act as a Cisco SAF Forwarder and a Cisco SAF Client, if necessary.

The following platforms support the SAF Forwarder:

- Cisco Integrated Services Routers (ISR), ISR Generation 2 (ISR G2), and 7200 Series Routers with Cisco IOS Release 15.0(1)M (See http://wwwin.cisco.com/ios/release/15mt)
- Cisco 7600 Series Routers with Cisco IOS Release 12.2(33)SRE
- Cisco ASR 1000 Series Aggregation Services Routers with Cisco IOS Release 12.2XE 2.5.0 (RLS5)

The following platforms support the SAF Client:

- Cisco Integrated Services Routers (ISR) and ISR Generation 2 (ISR G2) with Cisco IOS Release 15.0(1)M (See http://wwwin.cisco.com/ios/release/15mt)
- Cisco Unified Communications Manager 8.0(1) and higher versions

Cisco SAF Forwarder

The SAF Forwarder runs on a Cisco IOS router. A Cisco SAF Forwarder receives services advertised by Cisco SAF Clients, distributes the services reliably throughout the network of SAF Forwarders, and makes services available for Cisco SAF Clients to use.

Cisco SAF Forwarders use IP multicast to automatically discover and communicate as peers with other Cisco SAF Forwarders on a LAN. On networks that do not support IP multicast, SAF Forwarders can connect statically as peers by creating unicast point-to-point adjacencies with SAF neighbors.

To enable SAF within a network, you need to configure only a subset of the routers as SAF Forwarders. Once peer relationships have been created between the SAF Forwarders, the TCP/IP-based SAF messages exchanged between SAF Forwarders can traverse any IP network. Networks of non-SAF routers and SAF routers can run any IP routing protocol.

The SAF Forwarder Protocol (SAF-FP) is a "service" routing protocol, not an IP routing protocol. The SAF Forwarder Protocol routes information about services over IP networks. SAF-FP is based on EIGRP technology and takes advantage of many of the features historically developed for EIGRP-based IP routing, applying this functionality to the distribution of service information.

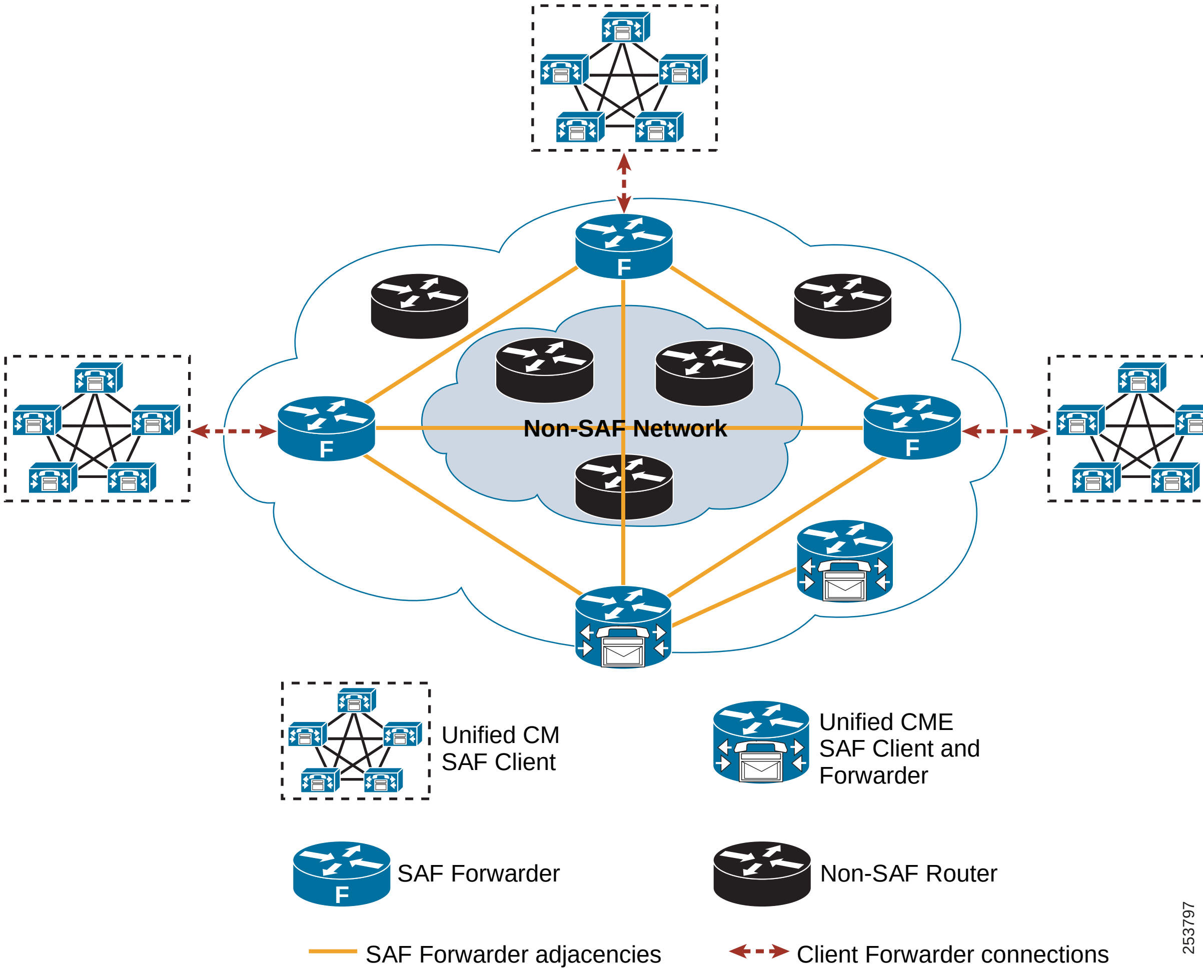
The SAF-Forwarder Protocol has the following characteristics:

- Uses the DUAL algorithm and split horizon rule to prevent routing loops
- Does not send periodic broadcasts, but sends updates only when changes occur
- Uses a keep-alive mechanism to track the availability of peer SAF Forwarders

- Is scalable and provides fast convergence when a SAF Forwarder fails
- Provides methods for SAF peer (neighbor) authentication

A Cisco SAF Forwarder provides the basis of the relationship between a Cisco SAF Client and the SAF network. Cisco SAF Forwarders may be located anywhere within the network but are normally located at the edges, or boundaries, of a network. (See Figure 3-22.) The Client/Forwarder relationship is used to maintain the state of each advertised service. If a Client removes a service or disconnects from the Forwarder node, the node informs the SAF network about the services that are no longer available. When a SAF Forwarder node receives advertisements from other Forwarder nodes, it keeps a copy of the entire advertisement and then forwards it to other SAF peers.

Figure 3-22 SAF Clients, SAF Forwarders, and Adjacencies Across Non-SAF Networks



Cisco SAF Client Overview

A Cisco SAF Client can be a producer of services (advertises services to the SAF network), a consumer of services (requests one or more services from the SAF network), or both. SAF clients perform three basic functions:

- Registering with the SAF network
- Publishing services
- Subscribing to services

SAF Clients take two forms (see Figure 3-23):

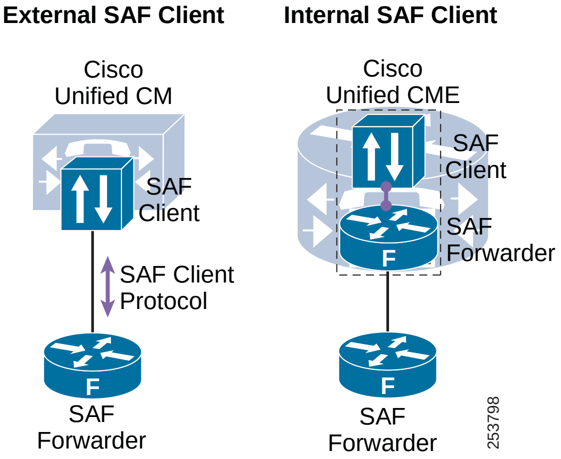
- Internal SAF clients

  An internal SAF client resides on the same Cisco IOS platform as the SAF Forwarder. The Client/Forwarder connection is established through an internal application programming interface (API). Call control applications that reside in Cisco IOS, such as Cisco Unified Communications Manager Express (Unified CME), can use the internal SAF client to connect to a co-resident internal SAF Forwarder.

- External SAF clients

  External SAF clients do not reside within Cisco IOS, and they use the SAF Client Protocol (SAF-CP) to communicate to a Cisco IOS-based SAF Forwarder. An external Cisco SAF client, such as the SAF client used by Cisco Unified CM, initiates a TCP/IP connection to a Cisco SAF Forwarder through a configured IP address and port number.

Figure 3-23 External and Internal SAF Clients and SAF Forwarders



Once the connection between the Client and Forwarder is established, the Cisco SAF Client sends a Register message to the Cisco SAF Forwarder. This register message uses a handle (called a "client label") to uniquely identify the Cisco SAF Client from all other Cisco SAF Clients connected to the Cisco SAF Forwarder. Once the Cisco SAF Client has completed its registration with the SAF Forwarder, it can then advertise (publish) services to, or request (subscribe) services from, the SAF network.

When advertising a service, a Cisco SAF Client publishes (sends) advertisements that contain details of the offered service to the Cisco SAF Forwarder. The Cisco SAF Client can send multiple publish requests, each advertising a distinct service. The Cisco SAF Forwarder advertises all services published by the Cisco SAF Client.

When requesting a service, the Cisco SAF Client sends the Forwarder a subscribe request. The subscribe request contains a filter that describes the set of services in which the Cisco SAF Client is interested. In response to this request, the Cisco SAF Forwarder sends the current set of services that match the filter to the Cisco SAF Client in a series of notify requests. Multiple notify requests are sent in order to provide flow control, and the Cisco SAF Client must respond to each notify request before the Cisco SAF Forwarder sends the next request. As with a publish request, the Cisco SAF Client can generate multiple subscribe requests, each with a different filter. The Cisco SAF Client can also generate an unsubscribe request, which removes one of its existing subscriptions.

Cisco External SAF Client and SAF Forwarder Interaction

Client/Forwarder Authentication

During the establishment of the TCP/IP connection between an external SAF Client and SAF Forwarder, a shared secret consisting of a username and a password is used for authentication. The username is used as an index to determine which password to use as the shared secret. When a Cisco SAF Client sends a request, it sends attributes that include its username, the actual message contents, and the MD5 hash of the password. When a Cisco SAF Forwarder receives a request, it locates the username attribute and uses it to access its local copy of the password. It then computes the MD5 hash of its locally stored password. If the passwords match, the Cisco SAF Client is authenticated and the connection proceeds. A Cisco SAF Forwarder can also elect to reject the request.

Client /Forwarder Keepalive

Once a SAF client has published its services to the SAF network, the Cisco SAF Forwarder uses a keepalive mechanism to track the status of the Cisco SAF Client. A Cisco SAF Forwarder and a Cisco SAF Client exchange a keepalive timer value at the time of registration. A Cisco SAF Forwarder considers a Cisco SAF Client to have failed if it has not seen a request from the Cisco SAF Client in a time period equal to the keepalive timer value. A Cisco SAF Client ensures that the interval between requests never exceeds this value. If a Cisco SAF Client has no data to send, it generates a register message to refresh the timer.

When a Cisco SAF Forwarder detects that the Cisco SAF Client has failed, it withdraws the services advertised on behalf of that Cisco SAF Client from the network and removes any subscriptions that the Cisco SAF Client had established. A Cisco SAF Client can be unregistered manually to cause a Cisco SAF Forwarder to withdraw all services and subscriptions gracefully.

SAF Forwarder Deployment Options

To enable SAF in a Unified Communications network, you must add one or more SAF Forwarders to the Unified Communications network. For Cisco IOS call control applications such as Unified CME, the SAF Client and Forwarder are co-resident on the router and can be used to interconnect to other SAF Forwarders in the SAF network. Non-IOS call control applications that use an External SAF Client, such as Unified CM, must connect to a Cisco IOS SAF Forwarder configured in the Unified Communications network. SAF Forwarders that are not co-resident with call control applications can be placed anywhere in the network. The number and location of these Forwarders largely depend on the degree of resilience and redundancy required within the SAF network. To provide redundancy, a minimum of two SAF Forwarders are required (see Figure 3-24.) Additional SAF Forwarders can be added to the SAF network to provide additional redundancy and local SAF Forwarder resources for each grouping of Unified CM clusters (see Figure 3-25). With the initial version of SAF for Cisco IOS Release 15.0(1) on Cisco ISR and 7200 Series Routers, up to 50 clients can connect to a single SAF Forwarder.

Figure 3-24 SAF Network with Two Dedicated SAF Forwarders and Two Unified CME SAF Forwarders

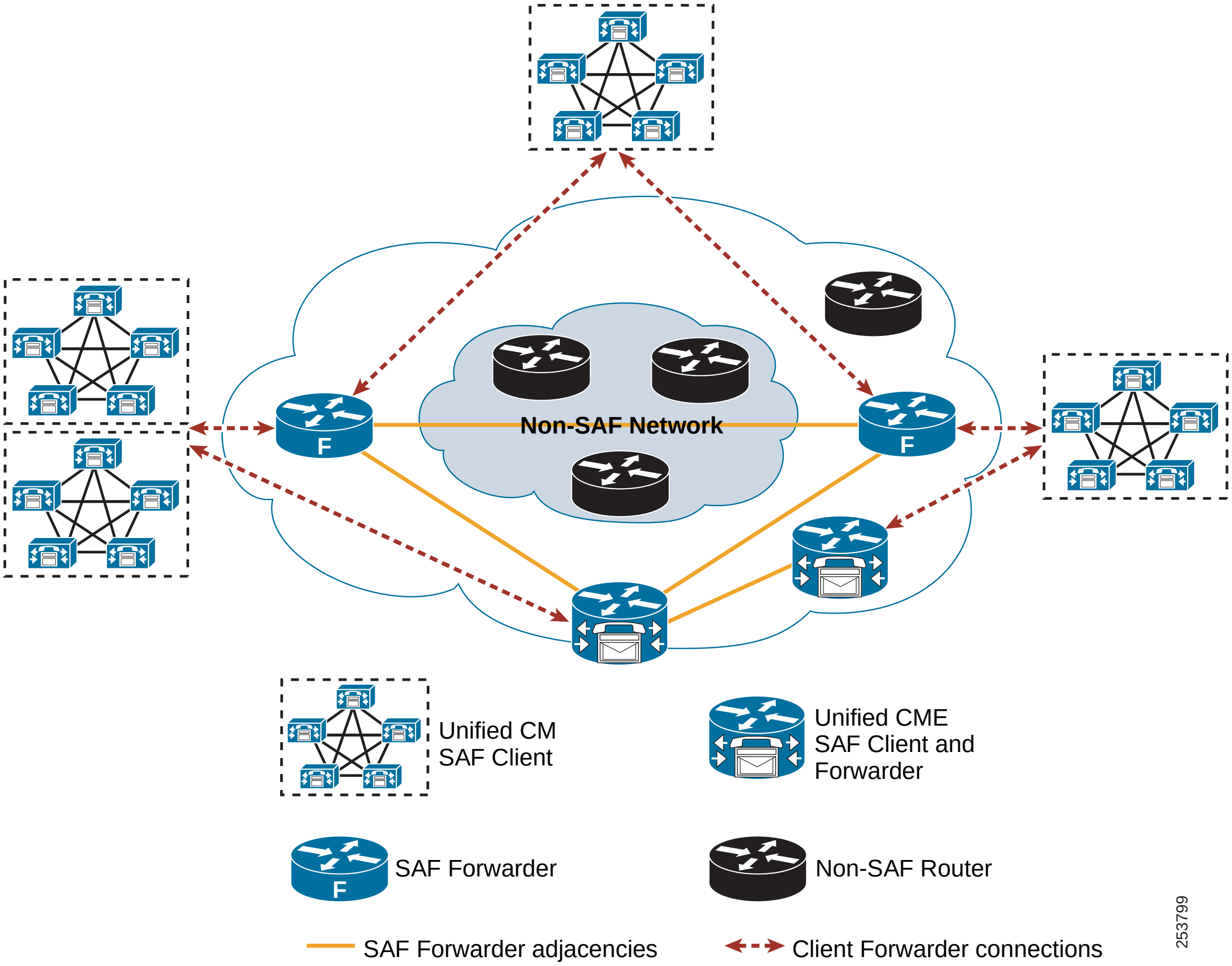
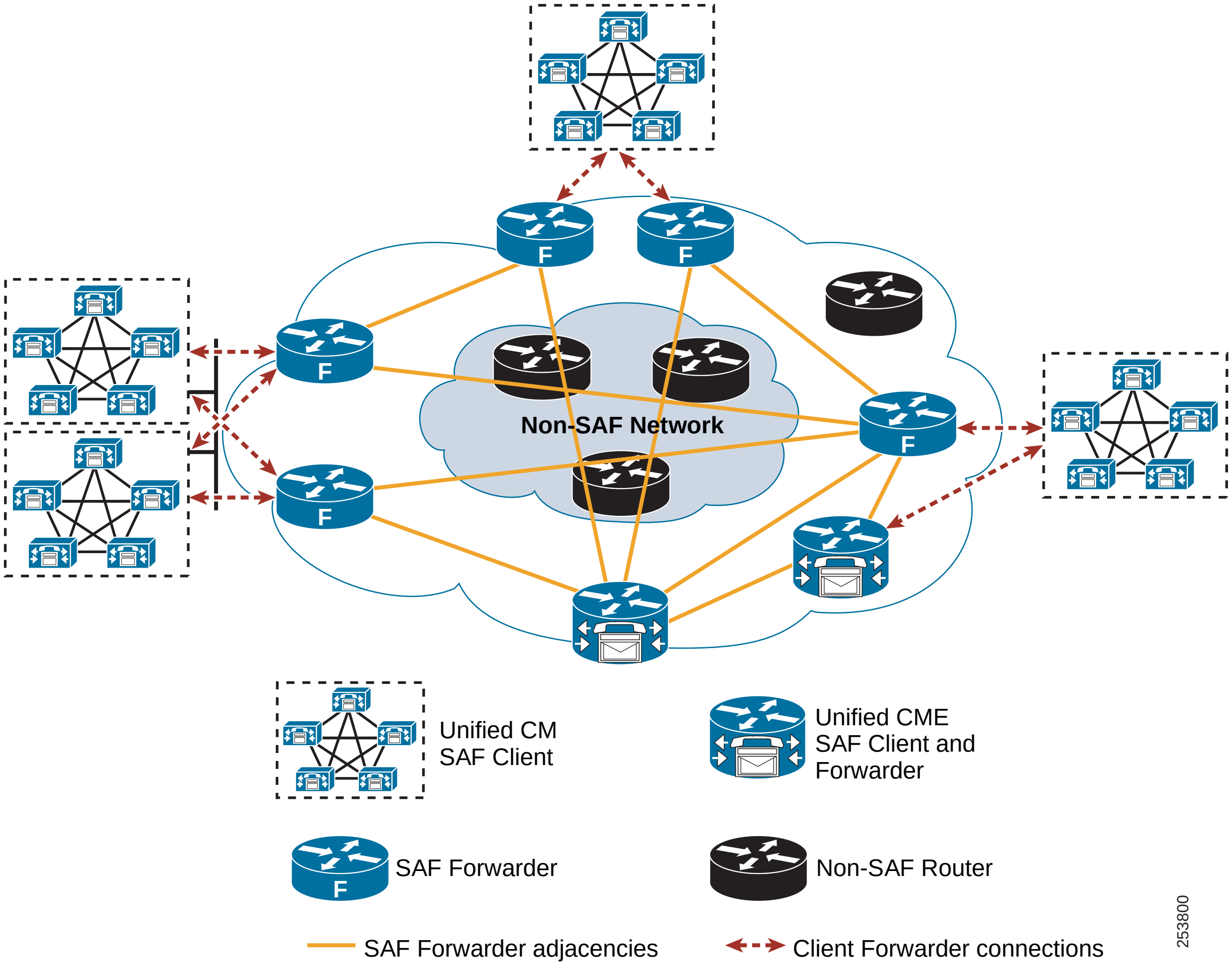


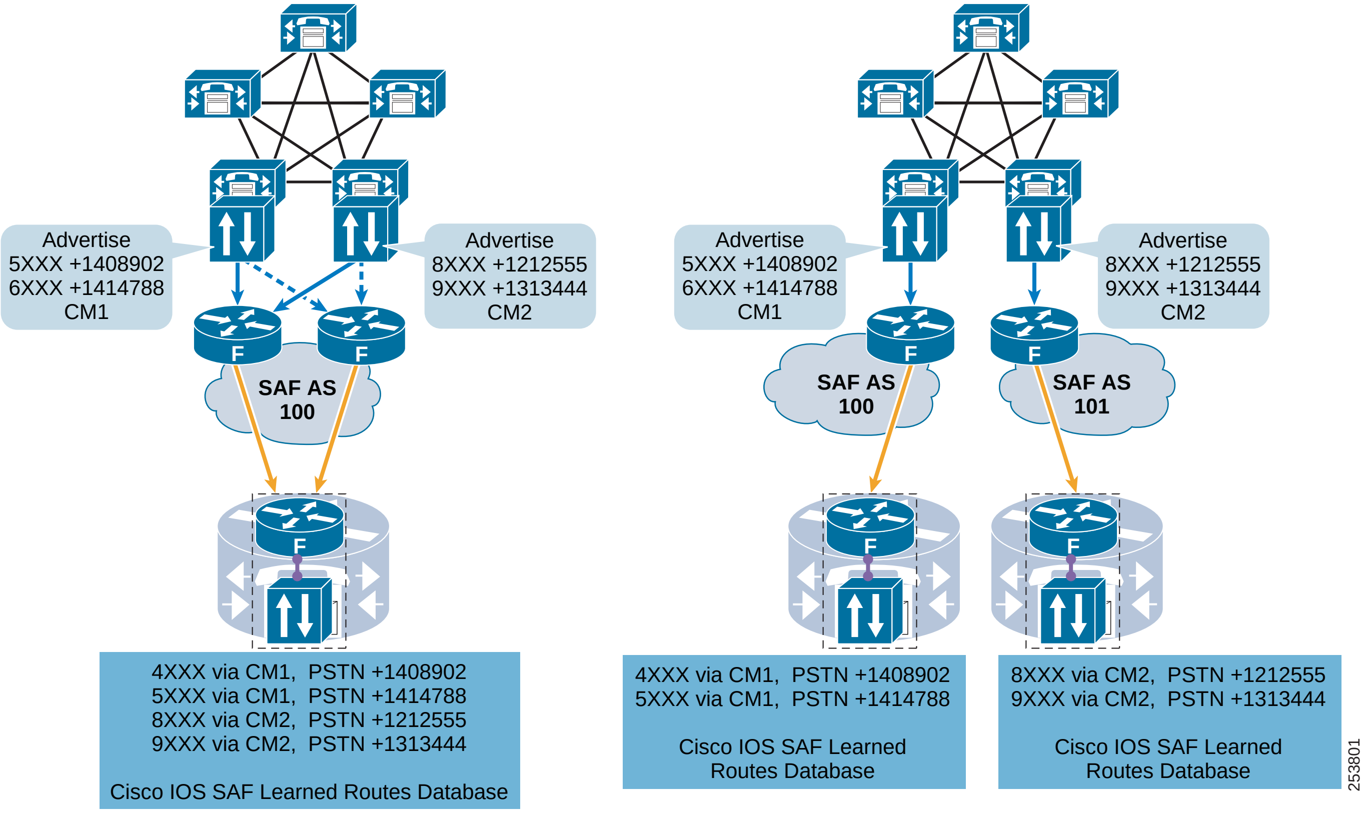
Figure 3-25 SAF Network with Multiple Redundant Dedicated SAF Forwarders and Two Unified CME SAF Forwarders



SAF Autonomous Systems

Similar to IP routing protocols, SAF uses the concept of an autonomous system (AS) to define the boundaries of a SAF network and the common SAF Forwarders within that SAF network. (See Figure 3-26.) The majority of SAF deployments require only a single SAF AS; however, in some cases (for example, where segregation of SAF services is required) multiple SAF ASs may be deployed. Each external SAF client can connect and publish to a single SAF AS. If you deploy multiple External SAF clients in a Unified CM cluster, the cluster can publish services into multiple SAF ASs and receive advertisements from each AS. Internal SAF clients can publish and subscribe to any number of Cisco IOS co-resident SAF ASs. Redistribution of SAF Services between SAF ASs is not available today.

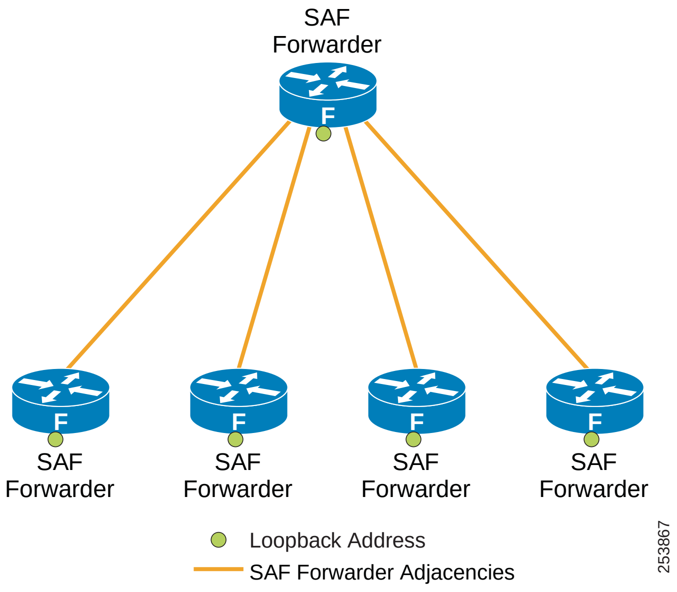
Figure 3-26 SAF Autonomous Systems



SAF Forwarder Loopback Addresses and Split Horizon

In Figure 3-27, if loopback addresses are used in the configuration of the SAF Forwarders, the split horizon rule comes into effect and the central SAF Forwarder does not forward advertisements between spoke Forwarders. To allow the central SAF Forwarder to forward advertisements between spoke Forwarders (and hence avoid the need to configure a full mesh of SAF peers), use the **no split horizon** command under the loopback interface of the central SAF Forwarder.

Figure 3-27 SAF and Split Horizon



For more information on Cisco IOS SAF configuration, refer to the *Cisco IOS Service Advertisement Framework Configuration Guide*, available at

http://www.cisco.com/en/US/docs/ios/saf/configuration/guide/15_0/saf_15_0_book.html

CHAPTER 4

Unified Communications Security

Revised: September 28, 2012; OL-27282-05

Securing the various components in a Cisco Unified Communications System is necessary for protecting the integrity and confidentiality of voice calls.

This chapter presents security guidelines pertaining specifically to Unified Communications technology and the voice network. For more information on data network security, refer to the Cisco SAFE Blueprint documentation available at

http://www.cisco.com/en/US/netsol/ns744/networking_solutions_program_home.html

Following the guidelines in this chapter does not guarantee a secure environment, nor will it prevent all penetration attacks on a network. You can achieve reasonable security by establishing a good security policy, following that security policy, staying up-to-date on the latest developments in the hacker and security communities, and maintaining and monitoring all systems with sound system administration practices.

This chapter addresses centralized and distributed call processing, including clustering over the WAN but not local failover mechanisms such as Survivable Remote Site Telephony (SRST). This chapter assumes that all remote sites have a redundant link to the head-end or local call-processing backup in case of head-end failure. The interaction between Network Address Translation (NAT) and IP Telephony, for the most part, is not addressed here. This chapter also assumes that all networks are privately addressed and do not contain overlapping IP addresses.

What's New in This Chapter

Table 4-1 lists the topics that are new in this chapter or that have changed significantly from previous releases of this document.

Table 4-1 New or Changed Information Since the Previous Release of This Document

New or Revised Topic	Described in:	Revision Date
Cisco Unified Video Advantage has reached end-of-sale (EoS) and has been removed from this chapter.	End-of-sale and end-of-life notice available at http://www.cisco.com/en/US/prod/collateral/video/ps7190/ps5662/end_of_life_notice_c51-704911.html	September 28, 2012
Minor updates for Cisco Unified Communications System Release 9.0	Various sections throughout this chapter	June 28, 2012

General Security

This section covers general security features and practices that can be used to protect the voice data within a network.

Security Policy

Cisco Systems recommends creating a security policy associated with every network technology deployed within your enterprise. The security policy defines which data in your network is sensitive so that it can be protected properly when transported throughout the network. Having this security policy helps you define the security levels required for the types of data traffic that are on your network. Each type of data may or may not require its own security policy.

If no security policy exists for data on the company network, you should create one before enabling any of the security recommendations in this chapter. Without a security policy, it is difficult to ascertain whether the security that is enabled in a network is doing what it is designed to accomplish. Without a security policy, there is also no systematic way of enabling security for all the applications and types of data that run in a network.

Note While it is important to adhere to the security guidelines and recommendations presented in this chapter, they alone are not sufficient to constitute a security policy for your company. You must define a corporate security policy before implementing any security technology.

This chapter details the features and functionality of a Cisco Systems network that are available to protect the Unified Communications data on a network. It is up to the security policy to define which data to protect, how much protection is needed for that type of data, and which security techniques to use to provide that protection.

One of the more difficult issues with a security policy that includes IP Telephony is combining the security policies that usually exist for both the data network and the traditional voice network. Ensure that all aspects of the integration of the voice data onto the network are secured at the correct level for your security policy or corporate environment.

The basis of a good security policy is defining how important your data is within the network. Once you have ranked the data according to its importance, you can decide how the security levels should be established for each type of data. You can then achieve the correct level of security by using both the network and application features.

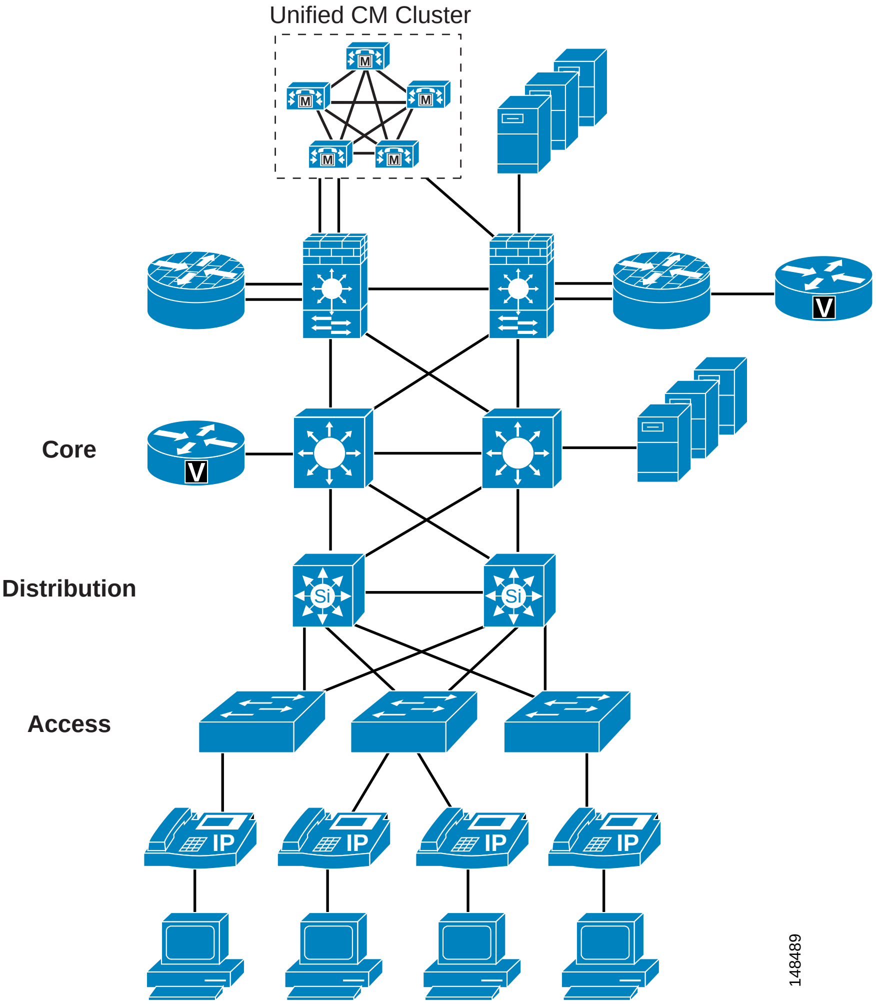
In summary, you can use the following process to define a security policy:

- Define the data that is on the network.
- Define the importance of that data.
- Apply security based on the importance of the data.

Security in Layers

This chapter starts with hardening the IP phone endpoints in a Cisco Unified Communications Solution and works its way through the network from the phone to the access switch, to the distribution layer, into the core, and then into the data center. (See Figure 4-1.) Cisco recommends building layer upon layer of security, starting at the access port into the network itself. This design approach gives a network architect the ability to place the devices where it is both physically and logically easy to deploy Cisco Unified Communications applications. But with this ease of deployment, the security complexity increases because the devices can be placed anywhere in a network as long as they have connectivity.

Figure 4-1 Layers of Security



Secure Infrastructure

As the IP Telephony data crosses a network, that data is only as safe and secure as the devices that are transporting the data. Depending on the security level that is defined in your security policy, the security of the network devices might have to be improved or they might already be secure enough for the transportation of IP Telephony traffic.

There are many best practices within a data network that, if used, will increase the entire security of your network. For example, instead of using Telnet (which sends passwords in clear text) to connect to any of the network devices, use Secure Shell (SSH, the secure form of Telnet) so that an attacker would not be able to see a password in clear text.

Gateways and gatekeepers can be configured with Cisco IOS feature sets that provide the required voice functionality but support only Telnet and not Secure Shell (SSH). Cisco recommends that you use access control lists (ACLs) to control who is permitted to connect to the routers using Telnet. It is more secure to connect to the gatekeeper from a host that is in a secure segment of the network, because user names and passwords are sent over Telnet in clear text.

You should also use firewalls, access control lists, authentication services, and other Cisco security tools to help protect these devices from unauthorized access.

Physical Security

Just as a traditional PBX is usually locked in a secure environment, the IP network should be treated in a similar way. Each of the devices that carries IP Telephony traffic is really part of an IP PBX, and normal general security practices should be used to control access to those devices. Once a user or attacker has physical access to one of the devices in a network, all kinds of problems could occur. Even if you have excellent password security and the user or attacker cannot get into the network device, that does not mean that they cannot cause havoc in a network by simply unplugging the device and stopping all traffic.

For more information on general security practices, refer to the documentation at the following locations:

- http://www.cisco.com/en/US/netsol/ns744/networking_solutions_program_home.html
- http://www.cisco.com/en/US/products/svcs/ps2961/ps2952/serv_group_home.html

IP Addressing

IP addressing can be critical for controlling the data that flows in and out of the logically separated IP Telephony network. The more defined the IP addressing is within a network, the easier it becomes to control the devices on the network.

As stated in other sections of this document (see Campus Access Layer, page 3-5), you should use IP addressing based on RFC 1918. This method of addressing allows deployment of an IP Telephony system into a network without redoing the IP addressing of the network. Using RFC 1918 also allows for better control in the network because the IP addresses of the voice endpoints are well defined and easy to understand. If the voice endpoints are all addressed within a 10.x.x.x. network, access control lists (ACLs) and tracking of data to and from those devices are simplified.

If you have a well defined IP addressing plan for your voice deployments, it becomes easier to write ACLs for controlling the IP Telephony traffic and it also helps with firewall deployments.

Using RFC 1918 enables you easily to deploy one VLAN per switch, which is a best practice for campus design, and also enables you to keep the Voice VLAN free of any Spanning Tree Protocol (STP) loops.

If deployed correctly, route summarization could help to keep the routing table about the same as before the voice deployment, or just slightly larger.

IPv6 Addressing

The introduction of IPv6 addressing has extended the network address space and increased the options for privacy and security of endpoints. Though both IPv4 and IPv6 have similar security concerns, IPv6 provides some advantages. For example, one of the major benefits with IPv6 is the enormous size of the subnets, which discourages automated scanning and reconnaissance attacks.

For a comparison of IPv6 and IPv4 in terms of security, refer to the *IPv6 and IPv4 Threat Comparison and Best-Practice Evaluation*, available at:

http://www.cisco.com/web/about/security/security_services/ciag/documents/v6-v4-threats.pdf

When considering IPv6 as your IP addressing method, adhere to the best practices documented in the following campus and branch office design guides:

- *Deploying IPv6 in Campus Networks*

  http://www.cisco.com/en/US/docs/solutions/Enterprise/Campus/CampIPv6.html
- *Deploying IPv6 in Branch Networks*

  http://www.cisco.com/en/US/docs/solutions/Enterprise/Branch/BrchIPv6.html

Access Security

This section covers security features at the Access level that can be used to protect the voice data within a network.

Voice and Video VLANs

Before the phone has its IP address, the phone determines which VLAN it should be in by means of the Cisco Discovery Protocol (CDP) negotiation that takes place between the phone and the switch. This negotiation allows the phone to send packets with 802.1q tags to the switch in a "voice VLAN" so that the voice data and all other data coming from the PC behind the phone are separated from each other at Layer 2. Voice VLANs are not required for the phones to operate, but they provide additional separation from other data on the network.

Voice VLANs can be assigned automatically from the switch to the phone, thus allowing for Layer 2 and Layer 3 separations between voice data and all other data on a network. A voice VLAN also allows for a different IP addressing scheme because the separate VLAN can have a separate IP scope at the Dynamic Host Configuration Protocol (DHCP) server.

Applications use CDP messaging from the phones to assist in locating phones during an emergency call. The location of the phone will be much more difficult to determine if CDP is not enabled on the access port to which that phone is attached.

There is a possibility that information could be gathered from the CDP messaging that would normally go to the phone, and that information could be used to discover some of the network. Not all devices that can be used for voice or video with Unified CM are able to use CDP to assist in discovering the voice VLAN.

Sony and Tandberg SCCP endpoints do not support Cisco Discovery Protocol (CDP) or 802.1Q VLAN ID tagging. To allow device discovery when third-party devices are involved, use the Link Layer Discovery Protocol (LLDP). LLDP for Media Endpoint Devices (LLDP-MED) is an extension to LLDP that enhances support for voice endpoints. LLDP-MED defines how a switch port transitions from LLDP to LLDP-MED if it detects an LLDP-MED-capable endpoint. Support for both LLDP and LLDP-MED on IP phones and LAN switches depends on the firmware and device models. To determine if LLDP-MED is supported on particular phone or switch models, check the specific product release notes or bulletins available at:

- http://www.cisco.com/en/US/products/hw/phones/ps379/prod_release_notes_list.html
- http://www.cisco.com/en/US/products/sw/iosswrel/ps5012/prod_bulletins_list.html

Note If an IP phone with LLDP-MED capability is connected to a Cisco Catalyst switch running an earlier Cisco IOS release that does not support LLDP, the switch might indicate that an extra device has been connected to the switch port. This can happen if the Cisco Catalyst switch is using Port Security to count the number of devices connected. The appearance of an LLDP packet might cause the port count to increase and cause the switch to disable the port. Verify that your Cisco Catalyst switch supports LLDP, or increase the port count to a minimum of three, before deploying Cisco IP Phones with firmware that supports LLDP-MED Link Layer protocol.

H.323 clients, Multipoint Control Units (MCUs), and gateways communicate with Unified CM using the H.323 protocol. Unified CM H.323 trunks (such as H.225 and intercluster trunk variants as well as the RASAggregator trunk type) use a random port range rather than the well-known TCP port 1720. Therefore, you must permit a wide range of TCP ports between these devices and the Unified CM servers. For port usage details, refer to the latest version of the *Cisco Unified Communications Manager TCP and UDP Port Usage* guide, available at:

http://www.cisco.com/en/US/products/sw/voicesw/ps556/prod_maintenance_guides_list.html

MCUs and gateways are considered infrastructure devices, and they typically reside within the datacenter adjacent to the Unified CM servers. H.323 clients, on the other hand, typically reside in the data VLAN.

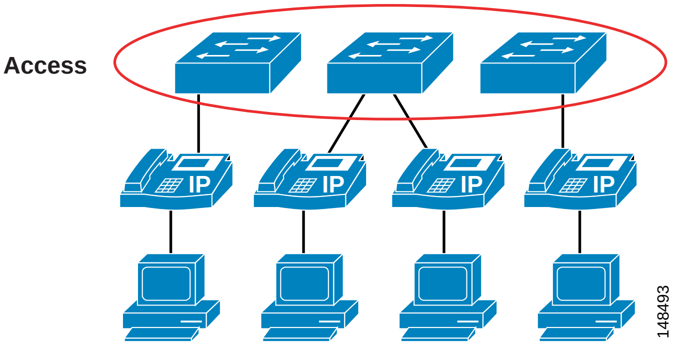
Cisco Unified Videoconferencing MCUs configured to run in SCCP mode communicate with the TFTP server(s) to download their configuration, with the Unified CM servers for signaling, and with other endpoints for RTP media traffic. Therefore, TFTP must be permitted between the MCU and the TFTP server(s), TCP port 2000 must be permitted between the MCUs and the Unified CM server(s), and UDP ports for RTP media must be permitted between the MCUs and the voice, data, and gateway VLANs

Switch Port

There are many security features within a Cisco switch infrastructure that can be used to secure a data network. This section describes some of the features that can be used in Cisco Access Switches to protect the IP Telephony data within a network. (See Figure 4-2.) This section does not cover all of the security features available for all of the current Cisco switches, but it does list the most common security features used across many of the switches that Cisco manufactures. For additional information on the security features available on the particular Cisco gear deployed within your network, refer to the appropriate product documentation available at

http://www.cisco.com

Figure 4-2 A Typical Access Layer Design to Which the Phones Attach



Port Security: MAC CAM Flooding

A classic attack on a switched network is a MAC content-addressable memory (CAM) flooding attack. This type of attack floods the switch with so many MAC addresses that the switch does not know which port an end station or device is attached to. When the switch does not know which port a device is attached to, it broadcasts the traffic destined for that device to the entire VLAN. In this way, the attacker is able to see all traffic that is coming to all the users in a VLAN.

To disallow malicious MAC flooding attacks from hacker tools such as macof, limit the number of MAC addresses allowed to access individual ports based on the connectivity requirements for those ports. Malicious end-user stations can use macof to originate MAC flooding from random-source to random-destination MAC addresses, both directly connected to the switch port or through the IP phone. The macof tool is very aggressive and typically can fill a Cisco Catalyst switch content-addressable memory (CAM) table in less than ten seconds. The flooding of subsequent packets that remain unlearned because the CAM table is filled, is as disruptive and unsecure as packets on a shared Ethernet hub for the VLAN that is being attacked.

Either port security or dynamic port security can be used to inhibit a MAC flooding attack. A customer with no requirement to use port security as an authorization mechanism would want to use dynamic port security with the number of MAC addresses appropriate to the function attached to a particular port. For example, a port with only a workstation attached to it would want to limit the number of learned MAC addresses to one. A port with a Cisco Unified IP Phone and a workstation behind it would want to set the number of learned MAC addresses to two (one for the IP phone itself and one for the workstation behind the phone) if a workstation is going to plug into the PC port on the phone. This setting in the past has been three MAC addresses, used with the older way of configuring the port in trunk mode. If you use the multi-VLAN access mode of configuration for the phone port, this setting will be two MAC addresses, one for the phone and one for the PC plugged into the phone. If there will be no workstation on the PC port, then the number of MAC addresses on that port should be set to one. These configurations are for a multi-VLAN access port on a switch. The configuration could be different if the port is set to trunk mode (not the recommended deployment of an access port with a phone and PC).

Port Security: Gratuitous ARP

Just like any other data device on the network, the phones are vulnerable to traditional data attacks. The phones have features to prevent some of the common data attacks that can occur on a corporate network. One such feature is Gratuitous ARP (Gratuitous Address Resolution Protocol, or GARP). This feature helps to prevent man-in-the-middle (MITM) attacks to the phone. A MITM attack involves an attacker who tricks an end station into believing that he is the router and tricks the router into believing that he is the end station. This scheme makes all the traffic between the router and the end station travel through the attacker, thus enabling the attacker to log all of the traffic or inject new traffic into the data conversation.

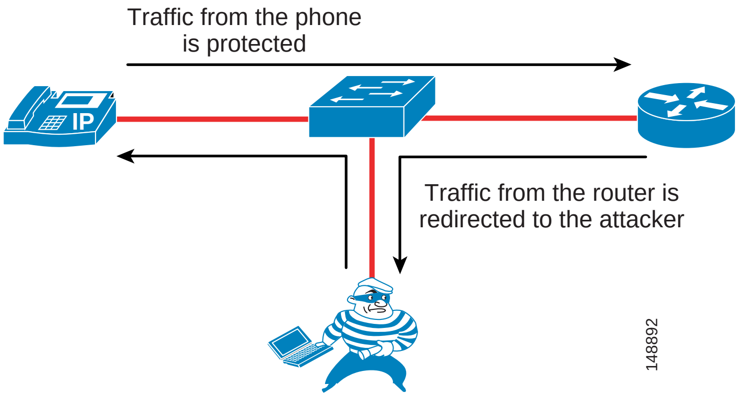
The Gratuitous ARP feature configured on an IP phone protects the phone from a traditional MITM attack on the signaling and RTP voice streams that are sourced from the phone to the network. It helps protect the phones from having an attacker capture the signaling and RTP voice streams from the phone if the attacker was able to get onto the voice segment of the network. This feature protects only the phones; it does not protect the rest of the infrastructure from a Gratuitous ARP attack. This feature is of less importance if you are running a Cisco infrastructure because the switch port provides features that protect both the phones and the network gear. For a description of these switch port features see the section on Switch Port, page 4-6.

Note The Gratuitous ARP feature does not apply to devices configured using IPv6 addressing. IPv6 uses neighbor discovery (ND) and not ARP.

The downstream signaling and RTP voice streams coming from another phone or coming across the network are not protected by this feature in the phone. Only the data coming from the phone that has this feature enabled is protected. (See Figure 4-3.)

If the default gateway is running Hot Standby Router Protocol (HSRP), if the HSRP configuration uses the burned-in MAC address rather than the virtual MAC address for the default gateway, and if the primary router fails-over to a secondary router that has a new MAC address, the phones could maintain the old MAC address of the default gateway. This scenario could cause an outage for up to 40 minutes. Always use the virtual MAC address in an HSRP environment to avoid this potential problem.

Figure 4-3 Gratuitous ARP Protects the Phone that Has It but Not Other Traffic



As shown in Figure 4-3, the traffic from the phone that has Gratuitous ARP is protected, but the attacker could still see the traffic coming from another endpoint because that endpoint might not have the ability to protect the data flow.

Port Security: Prevent Port Access

Prevent all port access except from those devices designated by their MAC addresses to be on the port. This is a form of device-level security authorization. This requirement is used to authorize access to the network by using the single credential of the device's MAC address. By using port security (in its non-dynamic form), a network administrator would be required to associate MAC addresses statically for every port. However, with dynamic port security, network administrators can merely specify the number of MAC addresses they would like the switch to learn and, assuming the correct devices are the first devices to connect to the port, allow only those devices access to that port for some period of time.

The period of time can be determined by either a fixed timer or an inactivity timer (non-persistent access), or it can be permanently assigned. In the latter case, the MAC address learned will remain on the port even in the event of a reload or reboot of the switch.

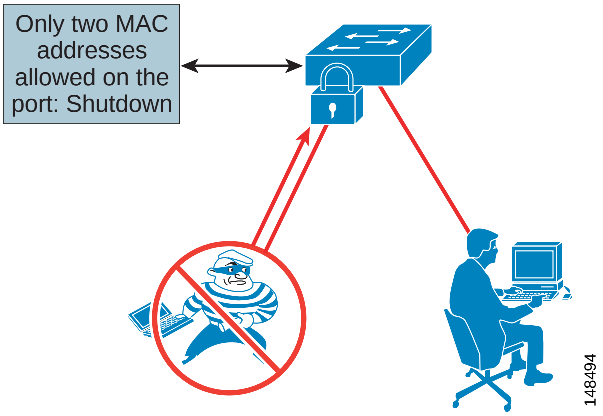
No provision is made for device mobility by static port security or persistent dynamic port security. Although it is not the primary requirement, MAC flooding attacks are implicitly prevented by port security configurations that aim to limit access to certain MAC addresses.

From a security perspective, there are better mechanisms for both authenticating and authorizing port access based on userid and/or password credentials rather than using MAC address authorization. MAC addresses alone can easily be spoofed or falsified by most operating systems.

Port Security: Prevent Rogue Network Extensions

Port security prevents an attacker from flooding the CAM table of a switch and from turning any VLAN into a hub that transmits all received traffic to all ports. It also prevents unapproved extensions of the network by adding hubs or switches into the network. Because it limits the number of MAC addresses to a port, port security can also be used as a mechanism to inhibit user extension to the IT-created network. For example, if a user plugs a wireless access point (AP) into a user-facing port or data port on a phone with port security defined for a single MAC address, the wireless AP itself would occupy that MAC address and not allow any devices behind it to access the network. (See Figure 4-4.) Generally, a configuration appropriate to stop MAC flooding is also appropriate to inhibit rogue access.

Figure 4-4 Limited Number of MAC Addresses Prevents Rogue Network Extensions



If the number of MAC addresses is not defined correctly, there is a possibility of denying access to the network or error-disabling the port and removing all devices from the network.

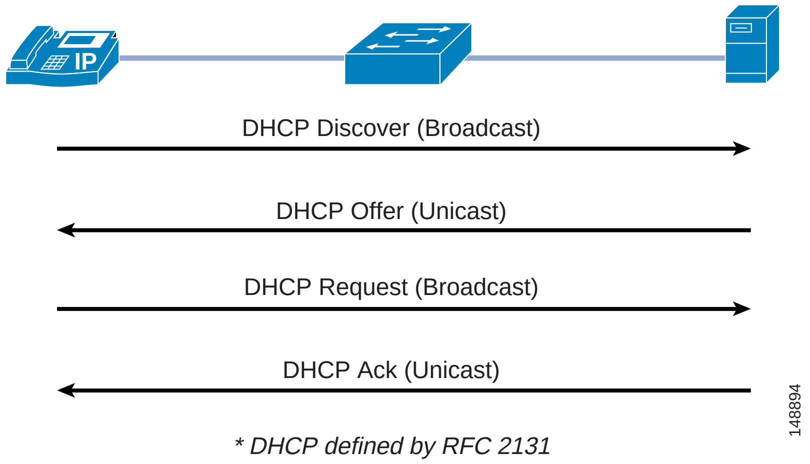
DHCP Snooping: Prevent Rogue DHCP Server Attacks

Dynamic Host Configuration Protocol (DHCP) Snooping prevents a non-approved DHCP or rogue DHCP server from handing out IP addresses on a network by blocking all replies to a DHCP request unless that port is allowed to reply. Because most phone deployments use DHCP to provide IP addresses to the phones, you should use the DHCP Snooping feature in the switches to secure DHCP messaging. Rogue DHCP servers can attempt to respond to the broadcast messages from a client to give out incorrect IP addresses, or they can attempt to confuse the client that is requesting an address.

When enabled, DHCP Snooping treats all ports in a VLAN as untrusted by default. An untrusted port is a user-facing port that should never make any reserved DHCP responses. If an untrusted DHCP-snooping port makes a DHCP server response, it will be blocked from responding. Therefore, rogue DHCP servers will be prevented from responding. However, legitimately attached DHCP servers or uplinks to legitimate servers must be trusted.

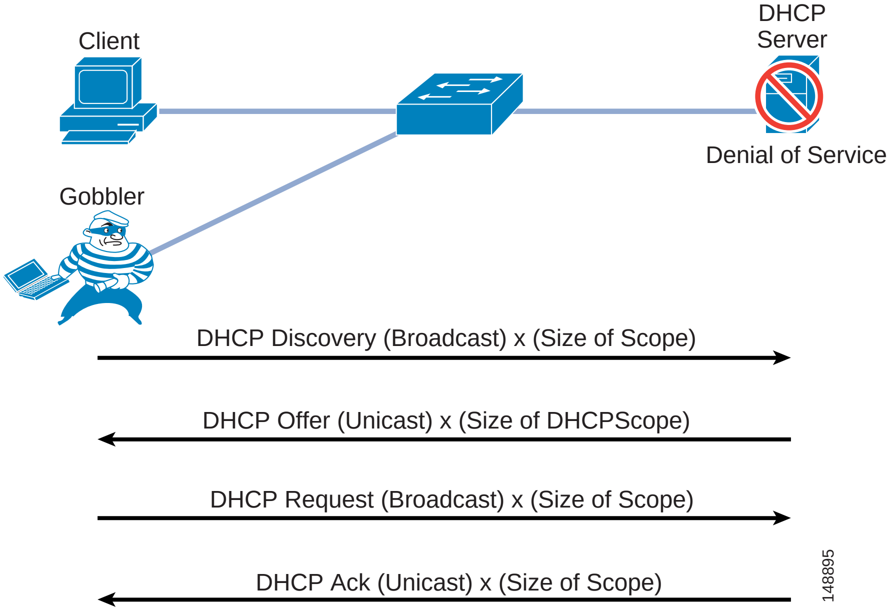
Figure 4-5 illustrates the normal operation of a network-attached device that requests an IP address from the DHCP server.

Figure 4-5 Normal Operation of a DHCP Request



However, an attacker can request not just a single IP address but all of the IP addresses that are available within a VLAN. (See Figure 4-6.) This means that there would be no addresses for a legitimate device trying to get on the network, and without an IP address the phone cannot connect to Unified CM.

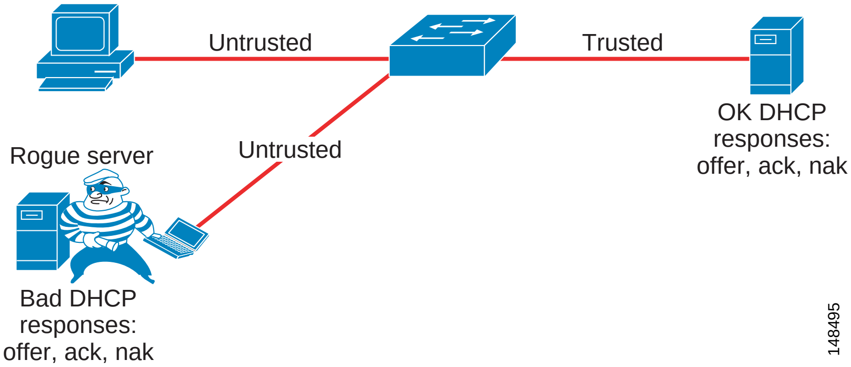
Figure 4-6 An Attacker Can Take All Available IP Addresses on the VLAN


Client
DHCP Server
Denial of Service
Gobbler
DHCP Discovery (Broadcast) x (Size of Scope)
DHCP Offer (Unicast) x (Size of DHCPScope)
DHCP Request (Broadcast) x (Size of Scope)
DHCP Ack (Unicast) x (Size of Scope)
148895

DHCP Snooping: Prevent DHCP Starvation Attacks

DHCP address scope starvation attacks from tools such as Gobbler are used to create a DHCP denial-of-service (DoS) attack. Because the Gobbler tool makes DHCP requests from different random source MAC addresses, you can prevent it from starving a DHCP address space by using port security to limit the number of MAC addresses. (See Figure 4-7.) However, a more sophisticated DHCP starvation tool can make the DHCP requests from a single source MAC address and vary the DHCP payload information. With DHCP Snooping enabled, untrusted ports will make a comparison of the source MAC address to the DHCP payload information and fail the request if they do not match.

Figure 4-7 Using DHCP Snooping to Prevent DHCP Starvation Attacks


Untrusted
Trusted
Untrusted
OK DHCP responses: offer, ack, nak
Rogue server
Bad DHCP responses: offer, ack, nak
148495

DHCP Snooping prevents any single device from capturing all the IP addresses in any given scope, but incorrect configurations of this feature can deny IP addresses to approved users.

DHCP Snooping: Binding Information

Another function of DHCP Snooping is to record the DHCP binding information for untrusted ports that successfully get IP addresses from the DHCP servers. The binding information is recorded in a table on the Cisco Catalyst switch. The DHCP binding table contains the IP address, MAC address, lease length, port, and VLAN information for each binding entry. The binding information from DHCP Snooping remains in effect for the length of the DHCP binding period set by the DHCP server (that is, the DHCP lease time). The DHCP binding information is used to create dynamic entries for Dynamic ARP Inspection (DAI) to limit ARP responses for only those addresses that are DHCP-bound. The DHCP binding information is also used by the IP source guard to limit sourcing of IP packets to only those addresses that are DHCP-bound.

There is a maximum limit to the number of binding table entries that each type of switch can store for DHCP Snooping. (Refer to the product documentation for your switch to determine this limit.) If you are concerned about the number of entries in your switch's binding table, you can reduce the lease time on the DHCP scope so that the entries in the binding table time-out sooner. The entries remain in the DHCP binding table until the lease runs out. In other words, the entries remain in the DHCP Snooping binding table as long at the DHCP server thinks the end station has that address. They are not removed from the port when the workstation or phone is unplugged.

If you have a Cisco Unified IP Phone plugged into a port and then move it to a different port, you might have two entries in the DHCP binding table with the same MAC and IP address on different ports. This behavior is considered normal operation.

Requirement for Dynamic ARP Inspection

Dynamic Address Resolution Protocol (ARP) Inspection (DAI) is a feature used on the switch to prevent Gratuitous ARP attacks on the devices plugged into the switch and on the router. Although it is similar to the Gratuitous ARP feature mentioned previously for the phones, Dynamic ARP protects all the devices on the LAN, and it is not just a phone feature.

In its most basic function, Address Resolution Protocol (ARP) enables a station to bind a MAC address to an IP address in an ARP cache, so that the two stations can communicate on a LAN segment. A station sends out an ARP request as a MAC broadcast. The station that owns the IP address in that request will give an ARP response (with its IP and MAC address) to the requesting station. The requesting station will cache the response in its ARP cache, which has a limited lifetime. The default ARP cache lifetime for Microsoft Windows is 2 minutes; for Linux, the default lifetime is 30 seconds; and for Cisco IP phones, the default lifetime is 40 minutes.

ARP also makes the provision for a function called Gratuitous ARP. Gratuitous ARP (GARP) is an unsolicited ARP reply. In its normal usage, it is sent as a MAC broadcast. All stations on a LAN segment that receive a GARP message will cache this unsolicited ARP reply, which acknowledges the sender as the owner of the IP address contained in the GARP message. Gratuitous ARP has a legitimate use for a station that needs to take over an address for another station on failure.

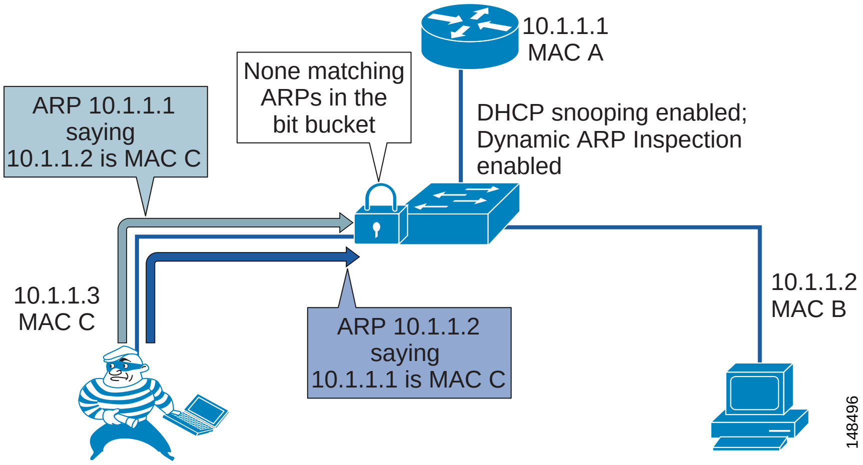
However, Gratuitous ARP can also be exploited by malicious programs that want to illegitimately take on the identity of another station. When a malicious station redirects traffic to itself from two other stations that were talking to each other, the hacker who sent the GARP messages becomes the man-in-the-middle. Hacker programs such as ettercap do this with precision by issuing "private" GARP messages to specific MAC addresses rather than broadcasting them. In this way, the victim of the attack does not see the GARP packet for its own address. Ettercap also keeps its ARP poisoning in effect by repeatedly sending the private GARP messages every 30 seconds.

Dynamic ARP Inspection (DAI) is used to inspect all ARP requests and replies (gratuitous or non-gratuitous) coming from untrusted (or user-facing) ports to ensure that they belong to the ARP owner. The ARP owner is the port that has a DHCP binding which matches the IP address contained in the ARP reply. ARP packets from a DAI trusted port are not inspected and are bridged to their respective VLANs.

Using DAI

Dynamic ARP Inspection (DAI) requires that a DHCP binding be present to legitimize ARP responses or Gratuitous ARP messages. If a host does not use DHCP to obtain its address, it must either be trusted or an ARP inspection access control list (ACL) must be created to map the host's IP and MAC address. (See Figure 4-8.) Like DHCP Snooping, DAI is enabled per VLAN, with all ports defined as untrusted by default. To leverage the binding information from DHCP Snooping, DAI requires that DHCP Snooping be enabled on the VLAN prior to enabling DAI. If DHCP Snooping is not enabled before you enable DAI, none of the devices in that VLAN will be able to use ARP to connect to any other device in their VLAN, including the default gateway. The result will be a self-imposed denial of service to any device in that VLAN.

Figure 4-8 Using DHCP Snooping and DAI to Block ARP Attacks


10.1.1.1
MAC A
None matching ARPs in the bit bucket
ARP 10.1.1.1 saying 10.1.1.2 is MAC C
DHCP snooping enabled; Dynamic ARP Inspection enabled
10.1.1.3
MAC C
ARP 10.1.1.2 saying 10.1.1.1 is MAC C
10.1.1.2
MAC B
148496

Because of the importance of the DHCP Snooping binding table to the use of DAI, it is important to back up the binding table. The DHCP Snooping binding table can be backed up to bootflash, File Transfer Protocol (FTP), Remote Copy Protocol (RCP), slot0, and Trivial File Transfer Protocol (TFTP). If the DHCP Snooping binding table is not backed up, the Cisco Unified IP Phones could lose contact with the default gateway during a switch reboot. For example, assume that the DHCP Snooping binding table is not backed up and that you are using Cisco Unified IP Phones with a power adapter instead of line power. When the switch comes back up after a reboot, there will be no DHCP Snooping binding table entry for the phone, and the phone will not be able to communicate with the default gateway unless the DHCP Snooping binding table is backed up and loads the old information before traffic starts to flow from the phone.

Incorrect configurations of this feature can deny network access to approved users. If a device has no entry in the DHCP Snooping binding table, then that device will not be able to use ARP to connect to the default gateway and therefore will not be able to send traffic. If you use static IP addresses, those addresses will have to be entered manually into the DHCP Snooping binding table. If you have devices that do not use DHCP again to obtain their IP addresses when a link goes down (some UNIX or Linux machines behave this way), then you must back up the DHCP Snooping binding table.

802.1X Port-Based Authentication

The 802.1X authentication feature can be used to identify and validate the device credentials of a Cisco Unified IP Phone before granting it access to the network. 802.1X is a MAC-layer protocol that interacts between an end device and a RADIUS server. It encapsulates the Extensible Authentication Protocol (EAP) over LAN, or EAPOL, to transport the authentication messages between the end devices and the switch. In the 802.1X authentication process, the Cisco Unified IP Phone acts as an 802.1X supplicant and initiates the request to access the network. The Cisco Catalyst Switch, acting as the authenticator, passes the request to the authentication server and then either allows or restricts the phone from accessing the network.

802.1X can also be used to authenticate the data devices attached to the Cisco Unified IP Phones. An EAPOL pass-through mechanism is used by the Cisco Unified IP Phones, allowing the locally attached PC to pass EAPOL messages to the 802.1X authenticator. The Cisco Catalyst Switch port needs to be configured in multiple-authentication mode to permit one device on the voice VLAN and multiple authenticated devices on the data VLAN.

Note Cisco recommends authenticating the IP phone before the attached data device is authenticated.

The multiple-authentication mode assigns authenticated devices to either a data or voice VLAN, depending on the attributes received from the authentication server when access is approved. The 802.1X port is divided into a data domain and a voice domain.

In multiple-authentication mode, a guest VLAN can be enabled on the 802.1x port. The switch assigns end clients to a guest VLAN when the authentication server does not receive a response to its EAPOL identity frame or when EAPOL packets are not sent by the client. This allows data devices attached to a Cisco IP Phone, that do not support 802.1X, to be connected to the network.

A voice VLAN must be configured for the IP phone when the switch port is in a multiple-host mode. The RADIUS server must be configured to send a Cisco Attribute-Value (AV) pair attribute with a value of **device-traffic-class=voice**. Without this value, the switch treats the IP phone as a data device.

Dynamic VLAN assignment from a RADIUS server is supported only for data devices.

When a data or a voice device is detected on a port, its MAC address is blocked until authorization succeeds. If the authorization fails, the MAC address remains blocked for 5 minutes.

When the 802.1x authentication is enabled on an access port on which a voice VLAN is configured and to which a Cisco IP Phone is already connected, the phone loses connectivity to the switch for up to 30 seconds.

Most Cisco IP Phones support authentication by means of X.509 certificates using the EAP-Transport Layer Security (EAP-TLS) or EAP-Flexible Authentication with Secure Tunneling (EAP-FAST) methods of authentication. Some of the older models that do not support either method can be authenticated using MAC Authentication Bypass (MAB), which enables a Cisco Catalyst Switch to check the MAC address of the connecting device as the method of authentication.

To determine support for the 802.1X feature configuration, refer to the product guides for the Cisco Unified IP Phones and the Cisco Catalyst Switches, available at http://www.cisco.com.

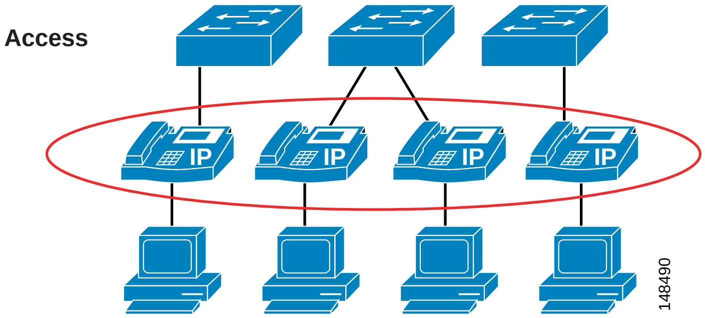
For configuration information, refer to the *IP Telephony for 802.1x Design Guide*, available at

http://www.cisco.com/en/US/docs/solutions/Enterprise/Security/TrustSec_1.99/IP_Tele/IP_Telephony_DIG.html

Endpoint Security

Cisco Unified IP Phones contain built-in features to increase security on an IP Telephony network. These features can be enabled or disabled on a phone-by-phone basis to increase the security of an IP Telephony deployment. Depending on the placement of the phones, a security policy will help determine if these features need to be enabled and where they should be enabled. (See Figure 4-9.)

Figure 4-9 Security at the Phone Level



The following security considerations apply to IP phones:

- PC Port on the Phone, page 4-15
- PC Voice VLAN Access, page 4-16
- Web Access Through the Phone, page 4-17
- Settings Access, page 4-17
- Authentication and Encryption, page 4-17
- VPN Client for IP Phones, page 4-18

Before attempting to configure the security features on a phone, check the documentation at the following link to make sure the features are available on that particular phone model:

http://www.cisco.com/en/US/products/sw/voicesw/index.html

PC Port on the Phone

The phone has the ability to turn on or turn off the port on the back of the phone, to which a PC would normally be connected. This feature can be used as a control point to access the network if that type of control is necessary.

Depending on the security policy and placement of the phones, the PC port on the back of any given phone might have to be disabled. Disabling this port would prevent a device from plugging into the back of the phone and getting network access through the phone itself. A phone in a common area such as a lobby would typically have its port disabled. Most companies would not want someone to get into the network on a non-controlled port because physical security is very weak in a lobby. Phones in a normal work area might also have their ports disabled if the security policy requires that no device should ever get access to the network through a phone PC port. Depending on the model of phone deployed, Cisco

Unified Communications Manager (Unified CM) can disable the PC port on the back of the phone. Before attempting to enable this feature, check the documentation at the following link to verify that this features is supported on your particular model of Cisco Unified IP Phone:

http://www.cisco.com/en/US/products/hw/phones/ps379/tsd_products_support_series_home.html

PC Voice VLAN Access

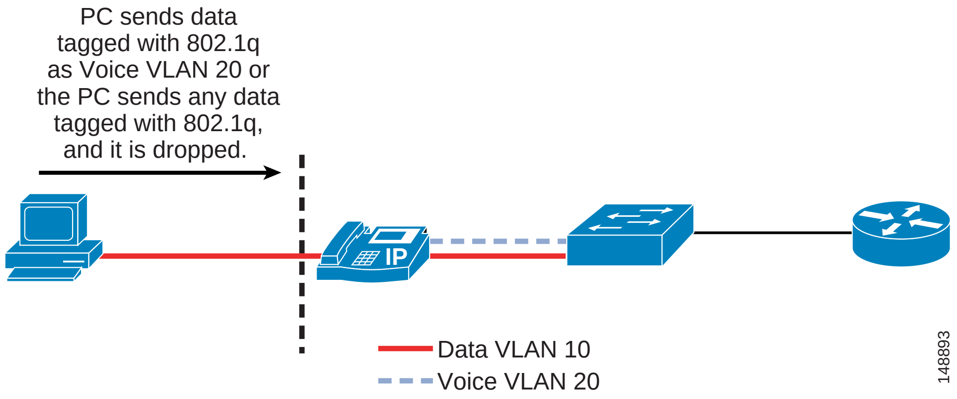
Because there are two VLANs from the switch to the phone, the phone needs to protect the voice VLAN from any unwanted access. The phones can prevent unwanted access into the voice VLAN from the back of the phone. A feature called PC Voice VLAN Access prevents any access to the voice VLAN from the PC port on the back of the phone. When disabled, this feature does not allow the devices plugged into the PC port on the phone to "jump" VLANs and get onto the voice VLAN by sending 802.1q tagged information destined for the voice VLAN to the PC port on the back of the phone. The feature operates one of two ways, depending on the phone that is being configured. On the more advanced phones, the phone will block any traffic destined for the voice VLAN that is sent into the PC port on the back of the phone. In the example shown in Figure 4-10, if the PC tries to send any voice VLAN traffic (with an 802.1q tag of 200 in this case) to the PC port on the phone, that traffic will be blocked. The other way this feature can operate is to block all traffic with an 802.1q tag (not just voice VLAN traffic) that comes into the PC port on the phone.

Currently, 802.1q tagging from an access port is not normally used. If that feature is a requirement for the PC plugged into the port on the phone, you should use a phone that allows 802.1q tagged packets to pass through the phone.

Before attempting to configure the PC Voice VLAN Access feature on a phone, check the documentation at the following link to make sure the feature is available on that particular phone model:

http://www.cisco.com/en/US/products/hw/phones/ps379/tsd_products_support_series_home.html

Figure 4-10 Blocking Traffic to the Voice VLAN from the Phone PC Port


PC sends data tagged with 802.1q as Voice VLAN 20 or the PC sends any data tagged with 802.1q, and it is dropped.
IP
Data VLAN 10
Voice VLAN 20
148893

Web Access Through the Phone

Each Cisco Unified IP Phone has a web server built into it to help with debugging and remote status of the phone for management purposes. The web server also enables the phones to receive applications pushed from Cisco Unified Communications Manager (Unified CM) to the phones. Access to this web server can be enabled or disabled on a phone by means of the Web Access feature in the Unified CM configuration. This setting can be global, or it could be enabled or disabled on a phone-by-phone basis.

If the web server is globally disable but it is needed to help with debugging, then the administrator for Unified CM will have to enable this feature on the phones. The ability to get to this web page can be controlled by an ACL in the network, leaving network operators with the capability to get to the web page when needed.

With the Web Access feature disabled, the phones will be unable to receive applications pushed to them from Unified CM.

Unified CM can be configured to use either HTTPS only or both HTTPS and HTTP for web traffic to and from the IP phones. However, if HTTPS only is configured, this does not by itself close port 80 on the IP phone's web server. It is preferable to use ACLs to restrict HTTP traffic, and configure Unified CM for HTTPS only.

Settings Access

Each Cisco Unified IP Phone has a network settings page that lists many of the network elements and detailed information that is needed for the phone to operate. This information could be used by an attacker to start a reconnaissance on the network with some of the information that is displayed on the phone's web page. For example, an attacker could look at the settings page to determine the default gateway, the TFTP server, and the Unified CM IP address. Each of these pieces of information could be used to gain access to the voice network or to attack a device in the voice network.

This access can be disabled on a phone-by-phone basis to prevent end users or attackers from obtaining the additional information such as Unified CM IP address and TFTP server information. With access to the phone settings page disabled, end users lose the ability to change many of the settings on the phone that they would normally be able to control, such as speaker volume, contrast, and ring type. It might not be practical to use this security feature because of the limitations it places on end users with respect to the phone interface.

For more information on the phone settings page, refer to the latest version of the *Cisco Unified Communications Manager Administration Guide*, available at

http://www.cisco.com/en/US/products/sw/voicesw/ps556/prod_maintenance_guides_list.html

Authentication and Encryption

Unified CM can be configured to provide multiple levels of security to the phones within a voice system, if those phones support those features. This includes device authentication and media and signaling encryption using X.509 certificates. Depending on your security policy, phone placement, and phone support, the security can be configured to fit the needs of your company.

For information on which Cisco Unified IP Phone models support specific security features, refer to the documentation available at

http://www.cisco.com/en/US/products/hw/phones/ps379/tsd_products_support_series_home.html

To enable security on the phones and in the Unified CM cluster, refer to the *Cisco Unified Communications Manager Security Guide*, available at

http://www.cisco.com/en/US/products/sw/voicesw/ps556/prod_maintenance_guides_list.html

When the Public Key Infrastructure (PKI) security features are properly configured in Unified CM, all supported phones will have the following capabilities:

- Integrity — Does not allow TFTP file manipulation but does allow Transport Layer Security (TLS) signaling to the phones when enabled.
- Authentication — The image for the phone is authenticated from Unified CM to the phone, and the device (phone) is authenticated to Unified CM. All signaling messages between the phone and Unified CM are verified as being sent from the authorized device.
- Encryption — For supported devices, signaling and media can be encrypted to prevent eavesdropping.
- Secure Real-time Transport Protocol (SRTP) — Is supported to Cisco IOS gateways and on phone-to-phone communications. Cisco Unity also supports SRTP for voicemail.

Unified CM supports authentication, integrity, and encryption for calls between two Cisco Unified IP Phones but not for all devices or phones. To determine if your device supports these features, refer to the documentation available at

http://www.cisco.com/en/US/products/hw/phones/ps379/tsd_products_support_series_home.html

Unified CM uses certificates for securing identities and enabling encryption. The certificates can be either Manufacturing Installed Certificates (MIC) or Locally Significant Certificates (LSC). MICs are already pre-installed and LSCs are installed by Unified CM's Cisco Certificate Authority Proxy Function (CAPF). Unified CM creates self-signed certificates, but signing of certificates by a third-party certificate authority (CA) using PKCS #10 Certificate Signing Request (CSR) is also supported. When using third-party CAs, the CAPF can be signed by the CA, but the phone LSCs are still generated by the CAPF. When MICs are used, the Cisco CA and the Cisco Manufacturing CA certificates act as the root certificates. When LSCs are generated for natively registered endpoints, the CAPF certificate is the root certificate.

Auto-registration does not work if you configure the cluster for mixed mode, which is required for device authentication. The cluster mixed-mode information is included in the CTL file downloaded by the endpoints. The CTL file configuration requires using a CTL client to sign the file. The CTL client is a separate application that is installed on a Windows PC, and it uses the Cisco Security Administrator Security Token (SAST), USB hardware device, to sign the CTL file.

Application layer protocol inspection and Application Layer Gateways (ALGs) that allow IP Telephony traffic to traverse firewalls and Network Address Translation (NAT) also do not work with signaling encryption. Not all gateways, phones, or conference are supported with encrypted media.

Encrypting media makes recording and monitoring of calls more difficult and expensive. It also makes troubleshooting VoIP problems more challenging.

VPN Client for IP Phones

Cisco Unified IP Phones with an embedded VPN client provide a secure option for connecting phones outside the network to the Unified Communications solution in the enterprise. This functionality does not require an external VPN router at the remote location, and it provides a secure communications tunnel for Layer 3 and higher traffic over an untrusted network between the phone at the deployed location and the corporate network.

The VPN client in Cisco Unified IP Phones uses Cisco SSL VPN technology and can connect to both the Cisco ASA 5500 Series VPN head-end and the Cisco Integrated Services Routers with the Cisco IOS SSL VPN software feature. The voice traffic is carried in UDP and protected by Datagram Transport Layer Security (DTLS) protocol as part of the VPN tunnel. The integrated VPN tunnel applies only to voice and IP phone services. A PC connected to the PC port cannot use this tunnel and needs to establish its own VPN tunnel for any traffic from the PC.

Note Cisco Unified CM 9.*x* allows any VXI clients connected to an IP phone’s PC port to join the phone's VPN tunnel. The VXI client must be configured in the phone’s device profile to allow it to use the tunnel.

For a phone with the embedded VPN client, you must first configure the phone with the VPN configuration parameters, including the VPN concentrator addresses, VPN concentrator credentials, user or phone ID, and credential policy. Because of the sensitivity of this information, the phone must be provisioned within the corporate network before the phone can attempt connecting over an untrusted network. Deploying the phone without first staging the phone in the corporate network is not supported.

The settings menu on the phone's user interface allows the user to enable or disable VPN tunnel establishment. When the VPN tunnel establishment is enabled, the phone starts to establish a VPN tunnel. The phone can be configured with up to three VPN concentrators to provide redundancy. The VPN client supports redirection from a VPN concentrator to other VPN concentrators as a load balancing mechanism.

For instructions on configuring the phones for the VPN client, refer to the latest version of the *Cisco Unified Communications Manager Administration Guide*, available at:

http://www.cisco.com/en/US/products/sw/voicesw/ps556/prod_maintenance_guides_list.html

Quality of Service

Quality of Service (QoS) is a vital part of any security policy for an enterprise network. Even though most people think of QoS as setting the priority of traffic in a network, it also controls the amount of data that is allowed into the network. In the case of Cisco switches, that control point is at the port level when the data comes from the phone to the Ethernet switch. The more control applied at the edge of the network at the access port, the fewer problems will be encountered as the data aggregates in the network.

QoS can be used to control not only the priority of the traffic in the network but also the amount of traffic that can travel through any specific interface. Cisco Smartports templates have been created to assist in deploying voice QoS in a network at the access port level.

A rigorous QoS policy can control and prevent denial-of-service attacks in the network by throttling traffic rates.

As mentioned previously in the lobby phone example, you can provide enough flow control of the traffic at the access port level to prevent any attacker from launching a denial-of-service (DoS) attack from that port in the lobby. The configuration for that example was not as aggressive as it could be because the QoS configuration allowed traffic sent to the port to exceed the maximum rate, but the traffic was remarked to the level of scavenger class. Given a more aggressive QoS policy, any amount of traffic that exceeded that maximum limit of the policy could just be dropped at the port, and that "unknown" traffic would never make it into the network. QoS should be enabled across the entire network to give the IP Telephony data high priority from end to end.

For more information on QoS, refer to the chapter on Network Infrastructure, page 3-1, and the *Enterprise QoS Solution Reference Network Design (SRND) Guide* available at

http://www.cisco.com/go/designzone

Access Control Lists

This section covers access control lists (ACLs) and their uses in protecting voice data.

VLAN Access Control Lists

You can use VLAN access control lists (ACLs) to control data that flows on a network. Cisco switches have the capability of controlling Layers 2 to 4 within a VLAN ACL. Depending on the types of switches in a network, VLAN ACLs can be used to block traffic into and out of a particular VLAN. They can also be used to block intra-VLAN traffic to control what happens inside the VLAN between devices.

If you plan to deploy a VLAN ACL, you should verify which ports are needed to allow the phones to function with each application used in your IP Telephony network. Normally any VLAN ACL would be applied to the VLAN that the phones use. This would allow control at the access port, as close as possible to the devices that are plugged into that access port.

Refer to the following product documentation for information on configuring VLAN ACLs:

- Cisco Catalyst 3750 Switches

  http://www.cisco.com/en/US/products/hw/switches/ps5023/products_installation_and_configuration_guides_list.html

- Cisco Catalyst 4500 Series Switches

  http://www.cisco.com/en/US/products/hw/switches/ps4324/products_installation_and_configuration_guides_list.html

- Cisco Catalyst 6500 Series Switches

  http://www.cisco.com/en/US/products/hw/switches/ps708/products_installation_and_configuration_guides_list.html

For more details on how to apply VLAN ACLs, refer to the following documentation:

- Cisco Catalyst 3750 Switches

  http://www.cisco.com/en/US/products/hw/switches/ps5023/products_installation_and_configuration_guides_list.html

- Cisco Catalyst 4500 Series Switches

  http://www.cisco.com/en/US/products/hw/switches/ps4324/products_installation_and_configuration_guides_list.html

- Cisco Catalyst 6500 Series Switches

  http://www.cisco.com/en/US/products/hw/switches/ps708/products_installation_and_configuration_guides_list.html

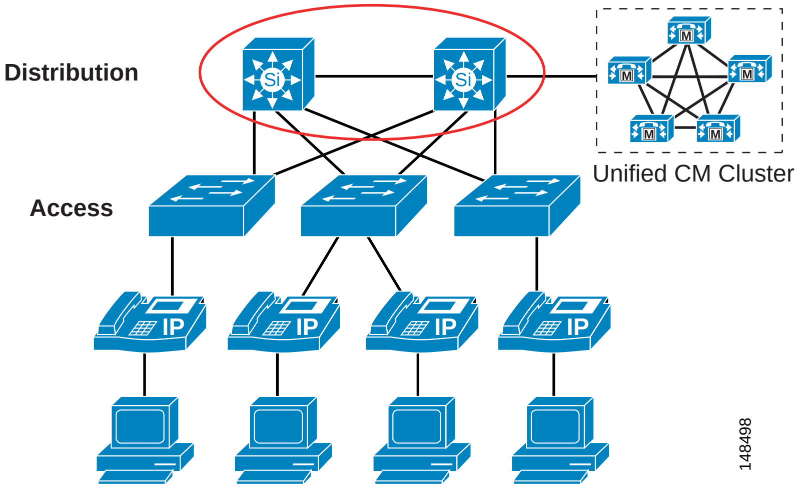
ACLs provide the ability to control the network traffic in and out of a VLAN as well as the ability to control the traffic within the VLAN.

VLAN ACLs are very difficult to deploy and manage at an access-port level that is highly mobile. Because of these management issues, care should be taken when deploying VLAN ACLs at the access port in the network.

Router Access Control Lists

As with VLAN ACLs, routers have the ability to process both inbound and outbound ACLs by port. The first Layer 3 device is the demarcation point between voice data and other types of data when using voice and data VLANs, where the two types of data are allowed to send traffic to each other. Unlike the VLAN ACLs, router ACLs are not deployed in every access device in your network. Rather, they are applied at the edge router, where all data is prepared for routing across the network. This is the perfect location to apply a Layer 3 ACL to control which areas the devices in each of the VLANs have the ability to access within a network. Layer 3 ACLs can be deployed across your entire network to protect devices from each other at points where the traffic converges. (See Figure 4-11.)

Figure 4-11 Router ACLs at Layer 3



There are many types of ACLs that can be deployed at Layer 3. For descriptions and examples of the most common types, refer to *Configuring Commonly Used IP ACLs*, available (with Cisco partner login required) at

http://cisco.com/en/US/partner/tech/tk648/tk361/technologies_configuration_example09186a0080100548.shtml

Depending on your security policy, the Layer 3 ACLs can be as simple as not allowing IP traffic from the non-voice VLANS to access the voice gateway in the network, or the ACLs can be detailed enough to control the individual ports and the time of the day that are used by other devices to communicate to IP Telephony devices. As the ACLs become more granular and detailed, any changes in port usage in a network could break not only voice but also other applications in the network.

If there are software phones in the network, if web access to the phone is allowed, or if you use the Attendant Console or other applications that need access to the voice VLAN subnets, the ACLs are much more difficult to deploy and control.

For IP phones restricted to specific subnets and limited to a voice VLAN, ACLs can be written to block all traffic (by IP address or IP range) to Unified CMs, voice gateways, phones, and any other voice application that is being used for voice-only services. This method simplifies the ACLs at Layer 3 compared to the ACLs at Layer 2 or VLAN ACLs.

Firewalls

Firewalls can be used in conjunction with ACLs to protect the voice servers and the voice gateways from devices that are not allowed to communicate with IP Telephony devices. Because of the dynamic nature of the ports used by IP Telephony, having a firewall does help to control opening up a large range of ports needed for IP Telephony communications. Given the complexities that firewalls introduce into a network design, you must take care in placing and configuring the firewalls and the devices around the firewalls to allow the traffic that is considered correct to pass while blocking the traffic that needs to be blocked.

IP Telephony networks have unique data flows. The phones use a client/server model for signaling for call setup, and Unified CM controls the phones through that signaling. The data flows for the IP Telephony RTP streams are more like a peer-to-peer network, and the phones or gateways talk directly to each other via the RTP streams. If the signaling flows do not go through the firewall so that the firewall can inspect the signaling traffic, the RTP streams could be blocked because the firewall will not know which ports need to be opened to allow the RTP streams for a conversation.

A firewall placed in a correctly designed network can force all the data through that device, so capacities and performance need to be taken into account. Performance includes the amount of latency, which can be increased by a firewall if the firewall is under high load or even under attack. The general rule in an IP Telephony deployment is to keep the CPU usage of the firewalls to less than 60% for normal usage. If the CPU runs over 60%, it increases the chance of impacting IP phones, call setup, and registration. If the CPU usage stays at a sustained level above 60%, the registered IP phones will be affected, quality of calls in progress will degrade, and call setup for new calls will suffer. In the worst case, if the sustained CPU usage stays above 60%, phones will start to unregister. When this happens, they will attempt to re-register with Unified CM, thus increasing the load on the firewalls even more. If this were to happen, the effect would be a rolling blackout of phones unregistering and attempting to re-register with Unified CM. Until the CPU usage of the firewall decreases to under 60% sustained load, this rolling blackout would continue and most (if not all) of the phones would be affected. If you are currently using a Cisco firewall in your network, you should monitor the CPU usage carefully when adding IP Telephony traffic to your network so that you do not adversely affect that traffic.

There are many ways to deploy firewalls. This section concentrates on the Cisco Adaptive Security Appliance (ASA) in the active/standby mode in both routed and transparent scenarios. Each of the configurations in this section is in single-context mode within the voice sections of the firewall configurations.

All of the Cisco firewalls can run in either multiple-context or single-context mode. In single-context mode, the firewall is a single firewall that controls all traffic flowing through it. In multiple-context mode, the firewalls can be turned into many virtual firewalls. Each of these contexts or virtual firewalls have their own configurations and can be controlled by different groups or administrators. Each time a new context is added to a firewall, it will increase the load and memory requirements on that firewall. When you deploy a new context, make sure that the CPU requirements are met so that voice RTP streams are not adversely affected.

Adaptive Security Appliances have limited support for application inspection of IPv6 traffic for Unified Communications application servers and endpoints. Cisco recommends not using IPv6 for Unified Communications if ASAs are deployed in your network.

Note An ASA with No Payload Encryption model disables Unified Communications features.

A firewall provides a security control point in the network for applications that run over the network. A firewall also provides dynamic opening of ports for IP Telephony conversations if that traffic is running through the firewall.

Using its application inspection capability, the firewall can inspect the traffic that runs though it to determine if that traffic is really the type of traffic that the firewall is expecting. For example, does the HTTP traffic really look like HTTP traffic, or is it an attack? If it is an attack, then the firewall drops that packet and does not allow it to get to the HTTP server behind the firewall.

Not all IP Telephony application servers or applications are supported with firewall application layer protocol inspection. Some of these applications include Cisco Unity voicemail servers, Cisco Unified Attendant Console, Cisco Unified Contact Center Enterprise, and Cisco Unified Contact Center Express. ACLs can be written for these applications to allow traffic to flow through a firewall.

Note The timers for failover on the firewalls are set quite high by default. To keep from affecting voice RTP streams as they go through the firewall if there is a failover, Cisco recommends reducing those timer settings to less than one second. If this is done, and if there is a failover, the amount of time that the RTP streams could be affected will be less because the firewalls will fail-over quicker and there will be less impact on the RTP streams during the failover time.

When firewalls are placed between different Unified Communications components, the application inspection must be enabled for all protocols used for communications between the components. Application inspection can fail in call flow scenarios used by features such as Silent Monitoring by Unified Communications Manager, when the firewall is between the remote agent phones and the supervisor phones.

Unified Communications devices using TCP, such as Cisco Unified Communications Manager, support the TCP SACK option to speed up data transfer in case of packet loss. But not all firewalls support the TCP SACK option. In that case, TCP sessions established between Unified Communications devices through such a firewall will encounter problems if they attempt to use the TCP SACK option, and the TCP session might fail. Therefore, the firewalls should provide full support for the TCP SACK option. If support is not available, then the firewalls should be able to modify the TCP packets during the three-way handshake and to disable TCP SACK option support so that the endpoints will not attempt to use this option.

To determine if the applications running on your network are supported with the version of firewall in the network or if ACLs have to be written, refer to the appropriate application documentation available at

http://www.cisco.com

Routed ASA

The ASA firewall in routed mode acts as a router between connected networks, and each interface requires an IP address on a different subnet. In single-context mode, the routed firewall supports Open Shortest Path First (OSPF) and Routing Information Protocol (RIP) in passive mode. Multiple-context mode supports static routes only. ASA version 8.*x* also supports Enhanced Interior Gateway Routing Protocol (EIGRP). Cisco recommends using the advanced routing capabilities of the upstream and downstream routers instead of relying on the security appliance for extensive routing needs. For more information on the routed mode, refer to the *Cisco Security Appliance Command Line Configuration Guide*, available at

http://www.cisco.com/en/US/products/ps6120/products_installation_and_configuration_guides_list.html

The routed ASA firewall supports QoS, NAT, and VPN termination to the box, which are not supported in the transparent mode (see Transparent ASA, page 4-24). With the routed configuration, each interface on the ASA would have an IP address. In the transparent mode, there would be no IP address on the interfaces other then the IP address to manage the ASA remotely.

The limitations of this mode, when compared to the transparent mode, are that the device can be seen in the network and, because of that, it can be a point of attack. In addition, placing a routed ASA firewall in a network changes the network routing because some of the routing can be done by the firewall. IP addresses must also be available for all the interfaces on the firewall that are going to be use, so changing the IP addresses of the routers in the network might also be required. If a routing protocol or RSVP is to be allowed through the ASA firewall, then an ACL will have to be put on the inside (or most trusted) interface to allow that traffic to pass to the outside (or lesser trusted) interfaces. That ACL must also define all other traffic that will be allowed out of the most trusted interface.

Transparent ASA

The ASA firewall can be configured to be a Layer 2 firewall (also known as "bump in the wire" or "stealth firewall"). In this configuration, the firewall does not have an IP address (other than for management purposes), and all of the transactions are done at Layer 2 of the network. Even though the firewall acts as a bridge, Layer 3 traffic cannot pass through the security appliance unless you explicitly permit it with an extended access list. The only traffic allowed without an access list is Address Resolution Protocol (ARP) traffic.

This configuration has the advantage that an attacker cannot see the firewall because it is not doing any dynamic routing. Static routing is required to make the firewall work even in transparent mode.

This configuration also makes it easier to place the firewall into an existing network because routing does not have to change for the firewall. It also makes the firewall easier to manage and debug because it is not doing any routing within the firewall. Because the firewall is not processing routing requests, the performance of the firewall is usually somewhat higher with **inspect** commands and overall traffic than the same firewall model and software that is doing routing.

With transparent mode, if you are going to pass data for routing, you will also have to define the ACLs both inside and outside the firewall to allow traffic, unlike with the same firewall in routed mode. Cisco Discovery Protocol (CDP) traffic will not pass through the device even if it is defined. Each directly connected network must be on the same subnet. You cannot share interfaces between contexts; if you plan on running multiple-context mode, you will have to use additional interfaces. You must define all non-IP traffic, such as routing protocols, with an ACL to allow that traffic through the firewall. QoS is not supported in transparent mode. Multicast traffic can be allowed to go through the firewall with an extended ACL, but it is not a multicast device. In transparent mode, the firewall does not support VPN termination other than for the management interface.

If a routing protocol or RSVP is to be allowed through the ASA firewall, then an ACL will have to be put on the inside (or most trusted) interface to allow that traffic to pass to the outside (or lesser trusted) interfaces. That ACL must also define all other traffic that will be allowed out of the most trusted interface.

For more information on the transparent mode, refer to the *Cisco Security Appliance Command Line Configuration Guide*, available at

http://www.cisco.com/en/US/products/ps6120/products_installation_and_configuration_guides_list.html

Note Using NAT in transparent mode requires ASA version 8.0(2) or later. For more information, refer to the *Cisco ASA 5500 Series Release Notes* at http://www.cisco.com/en/US/docs/security/asa/asa80/release/notes/asarn80.html.

ASA Intercompany Media Engine Proxy

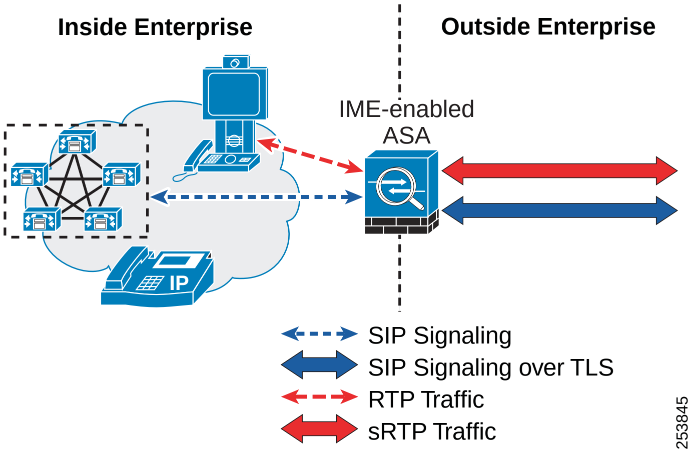
The ASA Cisco Intercompany Media Engine (IME) proxy is a required component of the Cisco IME solution for IME call processing. The IME enables secure business-to-business communication systems that support enhanced Unified Communications features and that do not have to go through the PSTN network. For more information about the IME call processing phase of the solution, see Cisco Intercompany Media Engine, page 5-75. The IME-enabled ASA provides perimeter security functions such as anti-spam blocking of non-IME calls and audio quality monitoring for the Fallback feature, inspects SIP messages, and acts as a proxy for SIP to SIP/TLS and RTP/SRTP conversions. The IME-enabled ASA terminates and re-initiates connections, which allows it to inspect the SIP messaging and apply SIP ALG processing. The ASA will convert the SIP/TLS traffic to TCP going toward Unified CM if Unified CM is not secure, or it will connect through TLS if Unified CM is secure. The following deployment models apply to the IME-enabled ASA:

- Basic (Inline)
- Offpath

Basic Deployment

In a basic (inline) deployment, the Internet ASA is configured with the IME feature, and all Internet-bound traffic from the Unified CM cluster will naturally traverse this IME-enabled ASA. As shown in Figure 4-12, the IME-enabled ASA resides on the edge of the enterprise and proxies all IME-related SIP trunk signaling and audio/video RTP media to remote enterprises.

Figure 4-12 Intercompany Media Engine ASA Basic (Inline) Deployment Model

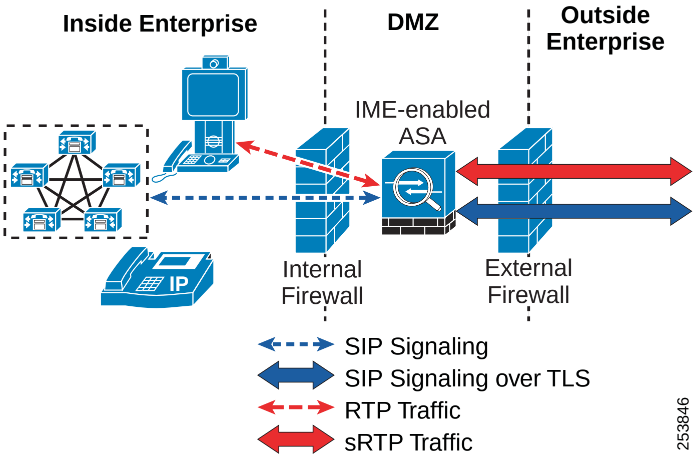


Offpath Deployment

In deployments where there are existing firewalls in the enterprise network, it might not be possible to replace or upgrade the existing firewall to support the IME feature or to change the existing security architecture by adding an IME-enabled ASA inline with the Internet firewall. In this scenario, the ASA can be implemented in an offpath model for IME. Offpath is the recommended deployment method.

In an offpath deployment, inbound and outbound IME calls pass through an IME-enabled ASA that is located in the DMZ, as illustrated in figure 4-20. Unified CM is configured to direct all SIP signaling to the IME-enabled ASA. All other Internet-bound traffic does not flow through the IME-enabled ASA.

Figure 4-13 Intercompany Media Engine ASA Offpath Deployment Model



Inbound IME calls from remote enterprises are addressed to the outside interface of the IME-enabled ASA, which utilizes static NAT or PAT to create a mapping to each Unified CM node on the inside. This behavior is the same for both deployment options. For outbound IME calls, offpath deployment requires that Unified CM send calls directly to the offpath IME-enabled ASA. This is accomplished via a mapping service protocol. Unified CM sends a mapping service request for the IME-enabled ASA to provide an internal IP address and port number to be used as the destination IP address and port number of the remote destination in the IME learned route. Unified CM then addresses the SIP Invite for this IME call to this internal IP address, which will guarantee the packet is forwarded to the IME-enabled ASA. Once the packet is received by the IME-enabled ASA, it then forwards the calls to the external IP address of the called party.

Mid-Call PSTN Fallback

The IME solution also provides a mechanism to allow calls to fall back to the PSTN if the quality of service (QoS) degrades below an acceptable level. The IME-enabled ASAs on the originating and terminating sides monitor all audio streams (not video) incoming from the internet and analyze the media against an algorithm with configurable sensitivity settings. Based on the observed loss and jitter measurements of an RTP stream, if the IME-enabled ASA determines call quality has deteriorated past its sensitivity threshold, it sends a SIP Refer message to its Unified CM to trigger the fallback. While the IME call remains active, the Unified CM on the originating side sets up a PSTN call in the background to the specific IME fallback DID (obtained during SIP call setup) of the remote enterprise. Once the terminating side Unified CM identifies the PSTN call as the fallback call for the IME call and a connection is established, both Unified CMs instruct the endpoints to switch media to the respective PSTN gateways. This change is seamless to the user. Any advanced features such as video are lost, but the audio portion of the call remains intact.

Cisco recommends starting with the default fallback sensitivity level and making revisions after determining how many calls are in fact falling back to PSTN connectivity. For more details regarding the IME solution and ASA configuration, refer to the Cisco Intercompany Media Engine Proxy information in the *Cisco ASA 5500 Series Configuration Guide using the CLI*, available at

http://www.cisco.com/en/US/docs/security/asa/asa83/configuration/guide/config.html

Design Considerations

The IME-enabled ASA requires at least two external (global) IP addresses, one for SIP signaling and one for media termination if PAT is used for incoming calls from remote enterprises. If NAT is implemented, more may be required. The external IP address on the IME-enabled ASA for SIP signaling is what is advertised in IME learned routes.

The IME-enabled ASA also requires at least two internal IP addresses, one for SIP signaling and one for media termination. PAT is used for incoming IME calls from Unified CM.

Note Although the IME-enabled ASA interfaces are referred to as external and internal, if the ASA is deployed in a DMZ, both interfaces may be on subnets that exist within the DMZ. At a minimum, the external interface subnet needs to be accessible from the Internet, and the internal interface subnet must be accessible from the intranet.

For any non-IME firewalls in the network that separate two components of the solution, it is imperative to open the proper pinholes to allow IME communications between the following components:

- IME server and Unified CM
- IME server and GoDaddy Enrollment Server
- IME server and the peer-to-peer IME server network (Distributed Cache Ring)
- IME-enabled ASA (internal) and Unified CM
- IME-enabled ASA (internal) and IME internal endpoints (media)
- IME-enabled ASA (external) and remote enterprise IME-enabled ASA

For a complete list of ports for the IME solution components, refer to the *Cisco Intercompany Media Engine Installation and Configuration Guide*, available at

http://cisco.com/en/US/products/sw/voicesw/ps556/prod_maintenance_guides_list.html

If there is an intranet firewall between the IME-enabled ASA and the Unified CM that is performing NAT, the following conditions must be met:

- This intranet firewall must be a Cisco ASA capable of SIP ALG functionality to allow the proper fixup of incoming and outgoing SIP messaging.
- There must be a static NAT entry to translate Unified CM's real IP address to an address reachable by the IME-enabled ASA.

The Cisco IME-enabled ASA typically has a default route for reaching Internet subnets. It also requires IP routes to all potential subnets containing internal endpoints. This includes data subnets that may include Cisco Unified Video Advantage cameras.

The IME solution requires its own Certificate Authority for validating ASA certificates used for establishing SIP/TLS connections. The IME-enabled ASA must verify SIP SSL certificates against this Certificate Authority (CA).

Note GoDaddy.com is the only authorized certificate provider for establishing secure SIP TLS connections with remote enterprises.

High Availability

IME-enabled ASAs can be deployed in an active/standby failover mode to provide stateless failover of IME communications. If an outage occurs, all calls being established, as well as existing calls, will be lost. Stateful failover is not supported.

With the offpath deployment method, Unified CM is capable of configuring multiple IME Services (sets of enrolled and excluded DIDs), each with its own IME firewall. This can add further resiliency to the solution.

Capacity Planning

Each model of ASA is rated for handling a certain number of audio and video calls. For current IME call capacities for the ASA, see Capacity Planning, page 5-83.

With the offpath model, each IME Service (set of enrolled and excluded DIDs) configured in Unified CM is associated with an IME-enabled ASA. Multiple IME Services can exist in Unified CM, allowing an administrator to spread the load across multiple IME-enabled ASAs, thus increasing overall capacity.

Note Cisco Unified CM 9.*x* currently does not support the ASA Phone Proxy and TLS Proxy features.

Data Center

Within the data center, the security policy should define what security is needed for the IP Telephony applications servers. Because the Cisco Unified Communications servers are based on IP, the security that you would put on any other time-sensitive data within a data center could be applied to those servers as well.

If clustering over the WAN is being used between data centers, any additional security that is applied both within and between those data centers has to fit within the maximum round-trip time that is allowed between nodes in a cluster. In a multisite or redundant data center implementation that uses clustering over the WAN, if your current security policy for application servers requires securing the traffic between servers across data center firewalls, then Cisco recommends using IPSec tunnels for this traffic between the infrastructure security systems already deployed.

To design appropriate data center security for your data applications, Cisco recommends following the guidelines presented in the *Data Center Networking: Server Farm Security SRND* (*Server Farm Security in the Business Ready Data Center Architecture*), available at

http://www.cisco.com/go/designzone

Gateways, Trunks, and Media Resources

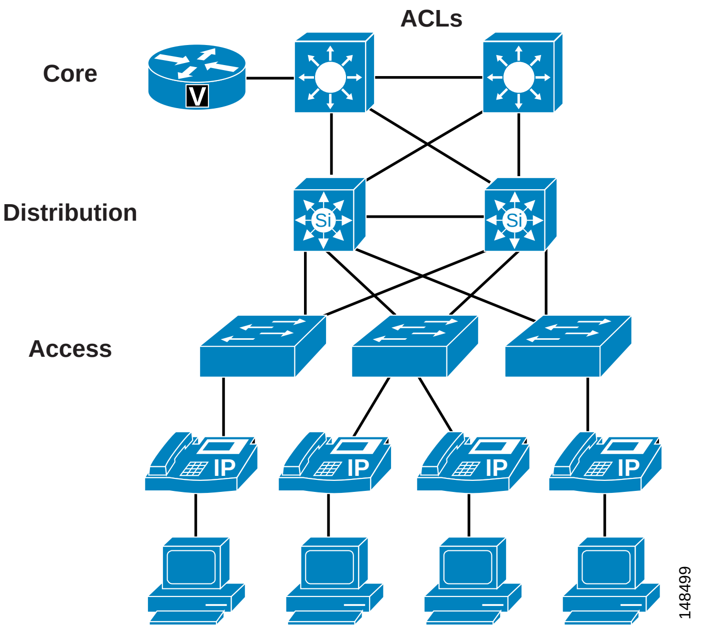
Gateways and media resources are devices that convert an IP Telephony call into a PSTN call. When an outside call is placed, the gateway or media resource is one of the few places within an IP Telephony network to which all the voice RTP streams flow.

Because IP Telephony gateways and media resources can be placed almost anywhere in a network, securing an IP Telephony gateway or media resource might be considered more difficult than securing other devices, depending on your security policy. However, depending on which point trust is established in the network, the gateways and media resources can be quite easy to secure. Because of the way the gateways and media resources are controlled by Unified CM, if the path that the signaling takes to the gateway or media resource is in what is considered a secure section of the network, a simple ACL can be used to control signaling to and from the gateway or media resource. If the network is not considered secure between the gateways (or media resources) and where the Unified CMs are located (such as when a gateway is located at a remote branch), the infrastructure can be used to build IPSec tunnels to the gateways and media resources to protect the signaling. Most networks would most likely use a combination of the two approaches (ACL and IPSec) to secure those devices.

For H.323 videoconferencing devices, an ACL can be written to block port 1720 for H.225 trunks from any H.323 client in the network. This method would block users from initiating an H.225 session with each other directly. Cisco devices might use different ports for H.225, so refer to the product documentation for your equipment to see which port is used. If possible, change the port to 1720 so that only one ACL is needed to control signaling.

Because we use QoS at the edge of the network, if an attacker can get into the voice VLAN and determine where the gateways and media resources are, QoS at the port would limit how much data the attacker would be able to send to the gateway or media resource. (See Figure 4-14.)

Figure 4-14 Securing Gateways and Media Resources with IPSec, ACLs, and QoS



Some gateways and media resources support Secure RTP (SRTP) to the gateways and media resources from the phones, if the phone is enabled for SRTP. To determine if a gateway or media resource supports SRTP, refer to the appropriate product documentation at:

http://www.cisco.com

For more information on IPSec tunnels, refer to the *Site-to-Site IPSec VPN Solution Reference Network Design (SRND)*, available at:

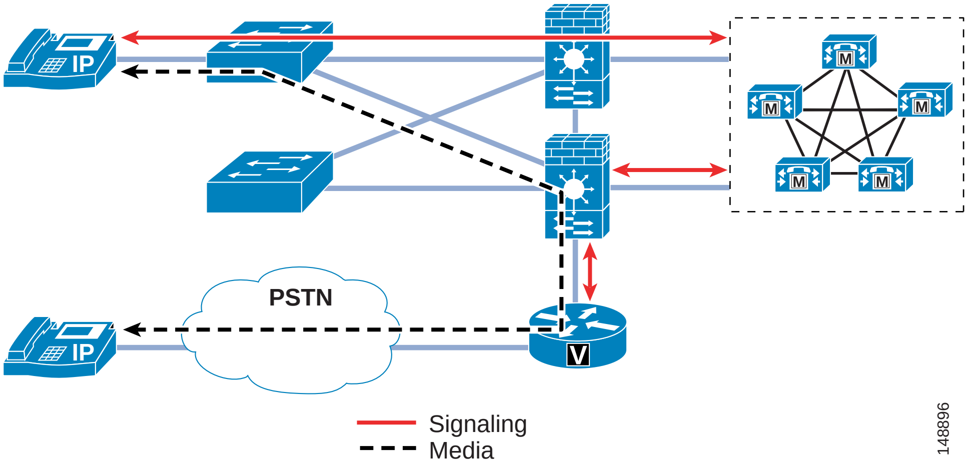
http://www.cisco.com/go/designzone

Putting Firewalls Around Gateways

Some very interesting issues arise from placing firewalls between a phone making a call and the gateway to the PSTN network. Stateful firewalls look into the signaling messages between Unified CM, the gateway, and the phone, and they open a pinhole for the RTP streams to allow the call to take place. To do the same thing with a normal ACL, the entire port ranges used by the RTP streams would have to be open to the gateway.

There are two ways to deploy gateways within a network: behind a firewall and in front of a firewall. If you place the gateway behind a firewall, all the media from the phones that are using that gateway have to flow through the firewall, and additional CPU resources are required to run those streams through the firewall. In turn, the firewall adds control of those streams and protects the gateway from denial-of-service attacks. (See Figure 4-15.)

Figure 4-15 Gateway Placed Behind a Firewall



The second way to deploy the gateway is outside the firewall. Because the only type of data that is ever sent to the gateway from the phones is RTP streams, the access switch's QoS features control the amount of RTP traffic that can be sent to that gateway. The only thing that Unified CM sends to the gateway is the signaling to set up the call. If the gateway is put in an area of the network that is trusted, the only communication that has to be allowed between Unified CM and the gateway is that signaling. (See Figure 4-15.) This method of deployment decreases the load on the firewall because the RTP streams are not going through the firewall.

Unlike an ACL, most firewall configurations will open only the RTP stream port that Unified CM has told the phone and the gateway to use between those two devices as long as the signaling goes through the firewall. The firewall also has additional features for DoS attacks and Cisco Intrusion Prevention System (IPS) signatures to look at interesting traffic and determine if any attackers are doing something they should not be doing.

As stated in the section on Firewalls, page 4-22, when a firewall is looking at all the signaling and RTP streams from phones to a gateway, capacity could be an issue. Also, if data other than voice data is running through the firewall, CPU usage must be monitored to make sure that the firewall does not affect the calls that are running through the firewall.

Firewalls and H.323

H.323 utilizes H.245 for setting up the media streams between endpoints, and for the duration of that call the H.245 session remains active between Unified CM and the H.323 gateway. Subsequent changes to the call flow are done through H.245.

By default, a Cisco firewall tracks the H.245 session and the associated RTP streams of calls, and it will time-out the H.245 session if no RTP traffic crosses the firewall for longer than 5 minutes. For topologies where at least one H.323 gateway and the other endpoints are all on one side of the firewall, the firewall will not see the RTP traffic. After 5 minutes, the H.245 session will be blocked by the firewall, which stops control of that stream but does not affect the stream itself. For example, no supplementary services will be available. This default behavior can be changed in firewall configuration so that the maximum anticipated call duration is specified.

The advantage of the configuration change from default is that it prevents H.323 from losing any call functionality when all endpoints are on the same side of the firewall.

SAF Service

Unified CM employs the Cisco Service Advertisement Framework (SAF) network service for its Call Control Discovery (CCD) feature (see Service Advertisement Framework (SAF), page 3-69). This capability uses a SIP trunk or a non-gatekeeper controlled H.323 trunk associated with the Call Control Discovery advertising service. The service advertises the call negotiation information for these trunks, including the dynamic port number for the H.323 trunk, the standard port 5060 for the SIP trunk, and the SIP route header information.

The Adaptive Security Appliances do not have application inspection for the SAF network service. When Unified CM uses a SAF-enabled H.323 trunk to place a call, the ASA cannot inspect the SAF packet to learn the ephemeral port number used in the H.225 signalling. Therefore, in scenarios where call traffic from SAF-enabled H.323 trunks traverses the ASAs, ACLs must be configured on the ASAs to allow this signaling traffic. The ACL configuration must account for all the ports used by the H.225 and H.245 signaling. ACL configuration is not required when SAF-enabled SIP trunks with the standard 5060 port are used.

Unified CM Trunk Integration with Cisco Unified Border Element

Unified CM trunks provide an additional point of IP connectivity between the enterprise network and external networks. Additional security measures must be applied to these interconnects to mitigate threats inherent in data and IP telephony applications. Implementing a Cisco Unified Border Element between the Unified CM trunks and the external network provides for more flexible and secure interoperability options.

The Cisco Unified Border Element is a Cisco IOS software feature that provides voice application demarcation and security threat mitigation techniques applicable to both voice and data traffic. Cisco Unified Border Element can be configured in conjunction with Cisco IOS Firewall, Authentication, and VPN features on the same device to increase security for the Unified CM trunks integrated with service provider networks or other external networks. These Cisco IOS security features can serve as a defense against outside attacks and as a checkpoint for the internal traffic exiting to the service provider's network through the router. Infrastructure access control lists (ACLs) can also be used to prevent unauthorized access, DoS attacks, or distributed DoS (DDoS) attacks that originate from the service provider or a network connected to the service provider's network, as well as to prevent intrusions and data theft.

Cisco Unified Border Element is a back-to-back user agent (B2BUA) that provides the capability to hide network topology on signaling and media. It enables security and operational independence of the network and provides NAT service by substituting the Cisco Unified Border Element IP address on all traffic.

Cisco Unified Border Element can be used to re-mark DSCP QoS parameters on media and signaling packets between networks. This ensures that traffic adheres to QoS policies within the network.

Cisco IOS Firewall features, used in combination with Cisco Unified Border Element, provide Application Inspection and Control (AIC) to match signaling messages and manage traffic. This helps prevent SIP trunk DoS attacks and allows message filtering based on content and rate limiting.

Cisco Unified Border Element allows for SIP trunk registration. This capability is not available in Unified CM SIP trunks.

Cisco Unified Border Element can register the enterprise network's E.164 DID numbers to the service provider's SIP trunk on behalf of the endpoints behind it. If Cisco Unified Border Element is used to proxy the network's E.164 DID numbers, the status of the actual endpoint is not monitored. Therefore unregistered endpoints might still be seen as available.

Cisco Unified Border Element can connect RTP enterprise networks with SRTP over an external network. This allows secure communications without the need to deploy SRTP within the enterprise. It also supports RTP-SRTP interworking, but this is limited to a small number of codecs, including G.711 mulaw, G.711 alaw, G.729abr8, G.729ar8, G.729br8, and G.729r8.

Certain SIP service providers require SIP trunks to be registered before they allow call service. This ensures that calls originate only from well-known endpoints, thus making the service negotiation between the enterprise and the service provider more secure. Unified CM does not support registration on SIP trunks natively, but this support can be accomplished by using a Cisco Unified Border Element. The Cisco Unified Border Element registers to the service provider with the phone numbers of the enterprise on behalf of Cisco Unified Communications Manager.

For configuration and product details about Cisco Unified Border Element, refer to the documentation at:

- http://www.cisco.com/en/US/products/sw/voicesw/ps5640/index.html
- http://www.cisco.com/en/US/products/sw/voicesw/ps5640/products_installation_and_configuration_guides_list.html

Applications Servers

For a list of the Unified CM security features and how to enable them, refer to the *Cisco Unified Communications Manager Security Guide*, available at

http://www.cisco.com/en/US/products/sw/voicesw/ps556/prod_maintenance_guides_list.html

Before enabling any of the Unified CM security features, verify that they will satisfy the security requirements specified in your enterprise security policy for these types of devices in a network. For more information, refer to the *Cisco ASA 5500 Series Release Notes* at

http://www.cisco.com/en/US/docs/security/asa/asa80/release/notes/asarn80.html

Single Sign-On

The Single Sign-On (SSO) feature was introduced in Cisco Unified CM 8.5(1), and it allows end users to log into a Windows domain and have secure access to the Unified Communication Manager's User Options page and the Cisco Unified Communications Integration for Microsoft Office Communicator (CUCIMOC) application.

Configuring Single Sign-On requires integration of Cisco Unified CM with third-party applications, including Microsoft Windows Servers, Microsoft Active Directory, and the ForgeRock Open Access Manager (OpenAM). For configuration details, refer to the latest version of the *Cisco Unified Communications Manager Features and Services Guide*, available at

http://www.cisco.com/en/US/products/sw/voicesw/ps556/prod_maintenance_guides_list.html

SELinux on the Unified CM and Application Servers

Security Enhanced Linux (SELinux) has replaced the Cisco Security Agent on Cisco Unified Communications System application servers as the Host Intrusion Prevention software. SELinux enforces policies, similar to the Cisco Security Agent, that look at the behavior of the traffic to and from the server, and the way the applications are running on that server, to determine if everything is working correctly. If something is considered abnormal, then SELinux's access rules prevent that activity from happening.

Connection rate limiting for DoS protection, and network shield protection for blocking specific ports, are configured using IPTables. The settings for the host-based firewall can be accessed using the Operating System Administration page of the Cisco Unified Communications server.

SELinux cannot be disabled by an administrator, but it can be set to a permissive mode. It should be made permissive strictly for troubleshooting purposes. Disabling SELinux requires root access and can be done only by remote support from Cisco Technical Assistance Center (TAC).

General Server Guidelines

Your Unified CM and other IP Telephony application servers should not be treated as normal servers. Anything you do while configuring the system could affect calls that are trying to be places or that are in progress. As with any other business-class application, major configuration changes should be done within maintenance windows to keep from disrupting phone conversations.

Standard security policies for application servers might not be adequate for IP Telephony servers. Unlike email servers or web servers, voice servers will not allow you to refresh a screen or re-send a message. The voice communications are real-time events. Any security policy for IP Telephony servers should ensure that work that is not related to configuring or managing the voice systems is not done on the IP Telephony servers at any time. Activities that might be considered normal on application servers within a network (for example, surfing the internet) should not take place on the IP Telephony servers.

In addition, Cisco provides a well defined patch system for the IP Telephony servers, and it should be applied based on the patch policy within your IT organization. You should not patch the system normally using the OS vendor's patch system unless it is approved by Cisco Systems. All patches should be downloaded from Cisco or from the OS vendor as directed by Cisco Systems, and applied according to the patch installation process.

You should use the OS hardening techniques if your security policy requires you to lock down the OS even more than what is provided in the default installation.

To receive security alerts, you can subscribe to the Cisco Notification service at:

http://www.cisco.com/cisco/support/notifications.html

Deployment Examples

This section presents examples of what could be done from a security perspective for a lobby phone and a firewall deployment. A good security policy should be in place to cover deployments similar to these types.

Lobby Phone Example

The example in this section illustrates one possible way to configure a phone and a network for use in an area with low physical security, such as a lobby area. None of the features in this example are required for a lobby phone, but if your security policy states more security is needed, then you could use the features listed in this example.

Because you would not want anyone to gain access to the network from the PC port on the phone, you should disable the PC port on the back of the phone to limit network access (see PC Port on the Phone, page 4-15). You should also disable the settings page on the phone so that potential attackers cannot see the IP addresses of the network to which the lobby phone is connected (see Settings Access, page 4-17). The disadvantage of not being able to change the settings on the phone usually will not matter for a lobby phone.

Because there is very little chance that a lobby phone will be moved, you could use a static IP address for that phone. A static IP address would prevent an attacker from unplugging the phone and then plugging into that phone port to get a new IP address (see IP Addressing, page 4-4). Also, if the phone is unplugged, the port state will change and the phone will no longer be registered with Unified CM. You can track this event in just the lobby phone ports to see if someone is trying to attach to the network.

Using static port security for the phone and not allowing the MAC address to be learned would mean that an attacker would have to change his MAC address to that of the phone, if he were able to discover that address. Dynamic port security could be used with an unlimited timer to learn the MAC address (but never unlearn it), so that it would not have to be added. Then the switch port would not have to be changed to clear that MAC address unless the phone is changed. The MAC address is listed in a label on the bottom of the phone. If listing the MAC address is considered a security issue, the label can be removed and replaced with a "Lobby Phone" label to identify the device. (See Switch Port, page 4-6.)

A single VLAN could be used and Cisco Discovery Protocol (CDP) could be disabled on the port so that attackers would not be able to see any information from the Ethernet port about that port or switch to which it is attached. In this case, the phone would not have a CDP entry in the switch for E911 emergency calls, and each lobby phone would need either a label or an information message to local security when an emergency number is dialed.

A static entry in the DHCP Snooping binding table could be made because there would be no DHCP on the port (see DHCP Snooping: Prevent Rogue DHCP Server Attacks, page 4-10). Once the static entry is in the DHCP Snooping binding table, Dynamic ARP Inspection could be enabled on the VLAN to keep the attacker from getting other information about one of the Layer 2 neighbors on the network (see Requirement for Dynamic ARP Inspection, page 4-12).

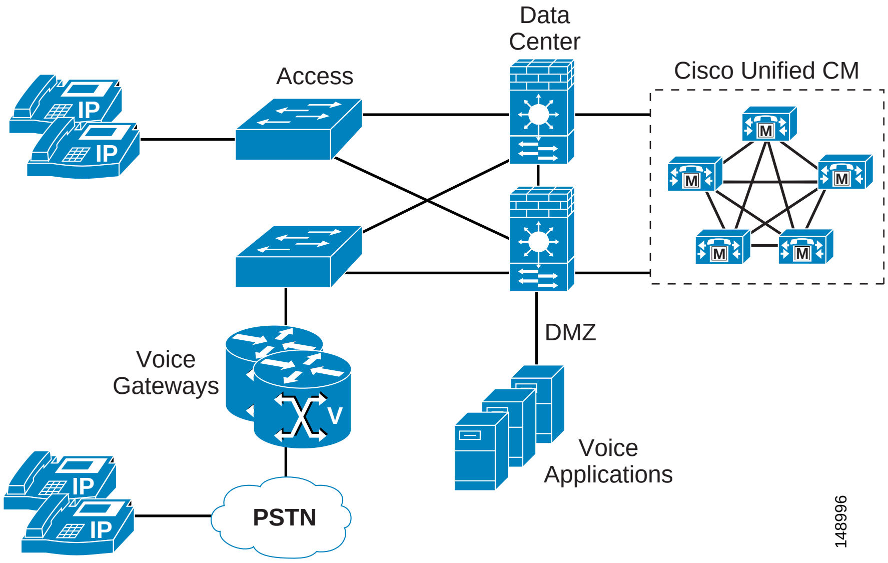
With a static entry in the DHCP Snooping binding table, IP Source Guard could be used. If an attacker got the MAC address and the IP address and then started sending packets, only packets with the correct IP address could be sent.

A VLAN ACL could be written to allow only the ports and IP addresses that are needed for the phones to operate (see VLAN Access Control Lists, page 4-20). The following example contains a very small ACL that can be applied to a port at Layer 2 or at the first Layer 3 device to help control access into the network (see Router Access Control Lists, page 4-21). This example is based on a Cisco 7960 IP Phone being used in a lobby area, without music on hold to the phone or HTTP access from the phone.

Firewall Deployment Example (Centralized Deployment)

The example in this section is one way that firewalls could be deployed within the data center, with Unified CMs behind them (see Figure 4-16). In this example, the Unified CMs are in a centralized deployment, single cluster with all the phones outside the firewalls. Because the network in this deployment already contained firewalls that are configured in routed mode within the corporate data center, the load was reviewed before the placement of gateways was determined. After reviewing the average load of the firewall, it was decided that all the RTP streams would not transverse the firewall in order to keep the firewalls under the 60% CPU load (see Putting Firewalls Around Gateways, page 4-31). The gateways are placed outside the firewalls, and ACLs within the network are used to control the TCP data flow to and from the gateways from the Unified CMs. An ACL is also written in the network to control the RTP streams from the phones because the IP addresses of the phones are well defined (see IP Addressing, page 4-4). The voice applications servers are placed within the demilitarized zone (DMZ), and ACLs are used at the firewalls to control access to and from the Unified CMs and to the users in the network. This configuration will limit the amount of RTP streams through the firewall using inspects, which will minimize the impact to the firewalls when the new voice applications are added to the existing network.

Figure 4-16 Firewall Deployment Example



Securing Network Virtualization

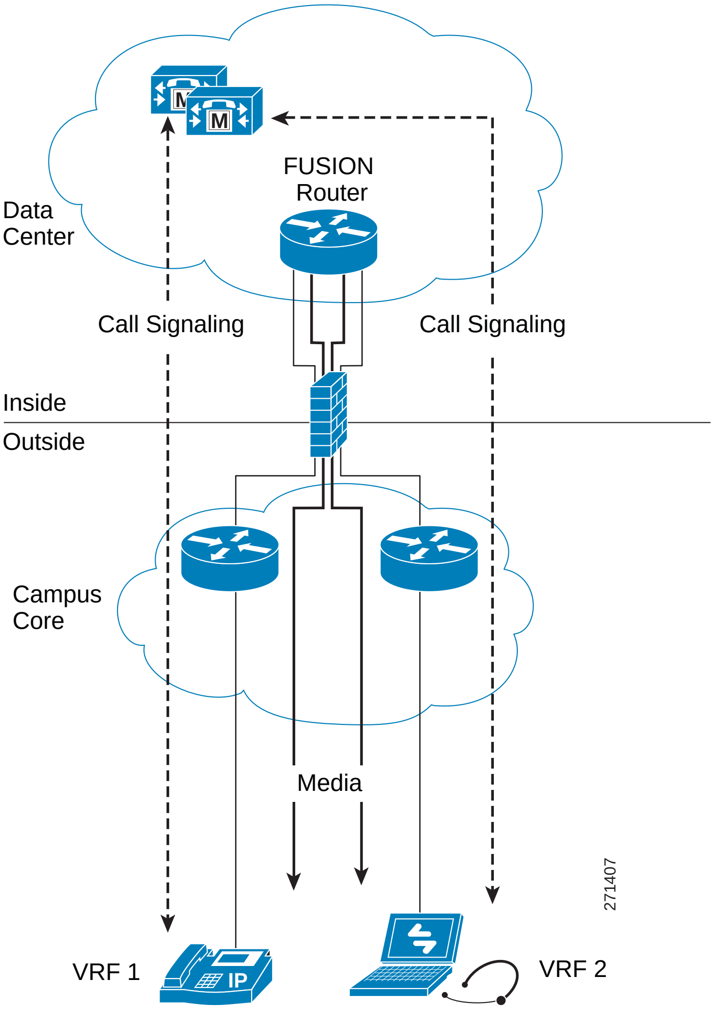
This section describes the challenges with providing homogenous connectivity for communications between virtual networks and a technique for overcoming these challenges. It assumes familiarity with Virtual Route Forwarding and Network Virtualization. Network design principles for these technologies are described in the Network Virtualization documentation available at http://www.cisco.com/go/designzone.

This discussion is not meant as an endorsement to use virtualization as a method to increase the security of a Unified Communications solution. Its purpose is to explain how such deployments can layer Unified Communications onto the existing infrastructure. Refer to the Network Virtualization documentation for evaluating the advantages and disadvantages of virtualization technology.

When a network is based on virtualization technology, there is a logical separation of traffic at Layer 3, and separate routing tables exist for each virtual network. Due to the lack of routing information, devices in different virtual networks cannot communicate with one another. This environment works well for client-server deployments where all user endpoints communicate with devices in the data center only, but it has issues for providing peer-to-peer communication. Regardless of how the virtual networks are arranged – whether by department, location, type of traffic (data or voice), or some other basis – the core issue is the same: endpoints in different Virtual Private Network Routing and Forwarding tables (VRFs) do not have the capability to communicate to one another. Figure 4-17 shows a solution that uses a shared VRF located in the data center to provide connectivity between a software-based phone located in one VRF and a hardware phone located in another VRF. This solution may also apply to other variants of this situation. Network Virtualization requires that fire-walling of the data center be implemented for the demarcation between the data center and the campus networks, and the following discussion shows how this can be implemented.

Scenario 1: Single Data Center

Figure 4-17 Single Data Center



This scenario is the simplest to implement and is an incremental configuration change beyond the usual network virtualization implementation. This design incorporates a data center router with the capability to route packets to any VRF, and it is called the fusion router. (Refer to the Network Virtualization documentation for details on the configuration of the fusion router.) The deployment scenario for enabling peer-to-peer communications traffic utilizes the fusion router for routing between VRFs and the firewall capabilities for securing access to the data center.

The following base requirements apply to this scenario:

- Campus routers send packets for other campus VRFs toward the fusion router via default routing, so all router hops must route by default to the fusion router. The data center shared VRF has route information about each campus VRF. All VRFs other than the shared VRF have no direct connectivity.
- A Unified CM cluster is located in a shared VRF in the data center, and communication within that shared VRF is totally unhindered.

- The shared VRF is located in the data center. If multiple data centers exist, the shared VRF spans all the data centers.

The application layer gateway at the data center edge specifies access lists to open ports for TFTP and SCCP or SIP sessions originated on the outside toward the Unified CM cluster in the data center. TFTP is required to allow phones to download their configuration and software images from their TFTP server, and SCCP or SIP is required to allow them to register with the Unified CM cluster. Refer to Unified CM product documentation for a list of appropriate port numbers for the particular version of software used.

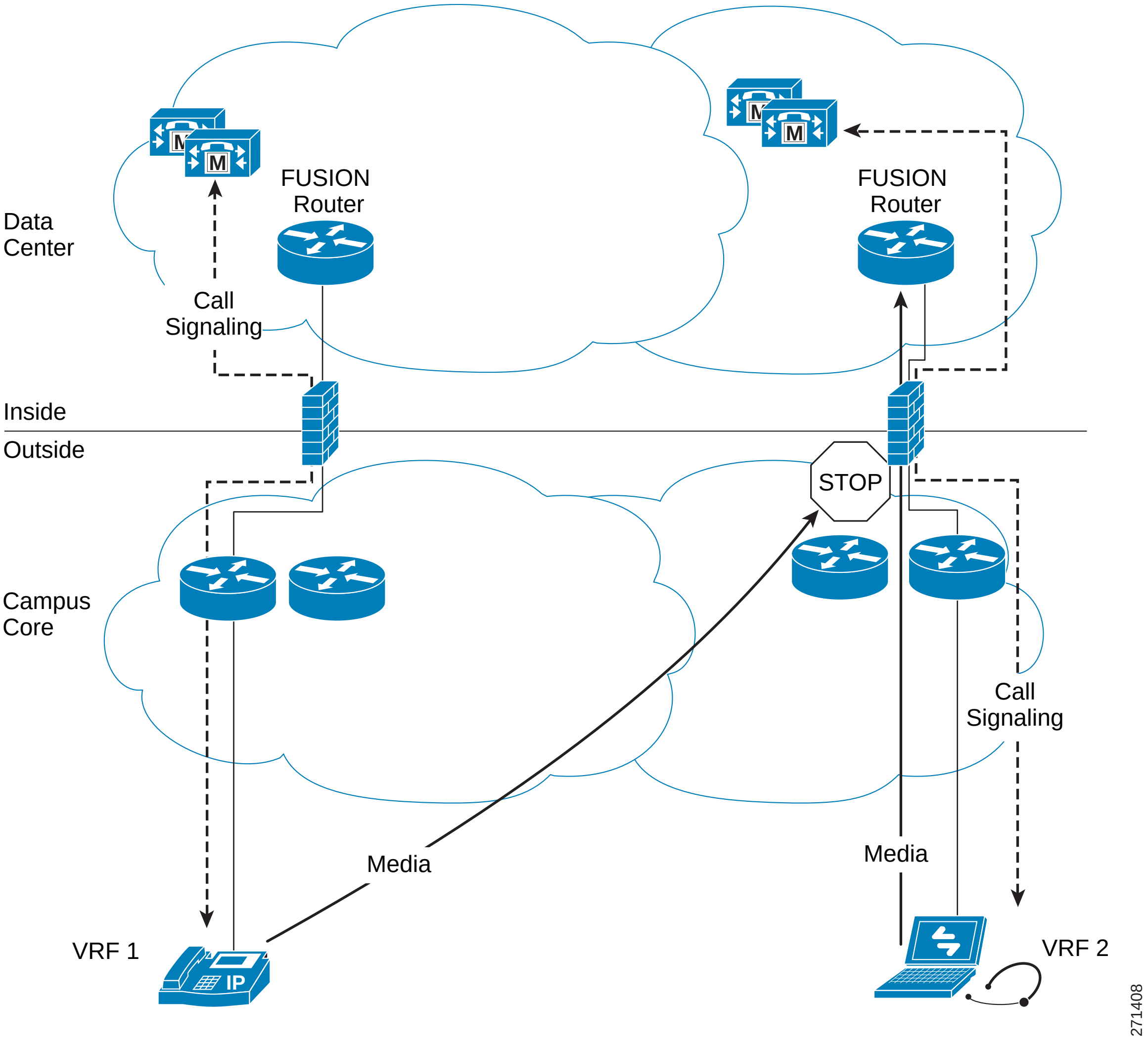
In this scenario, all call signaling from communication devices in each VRF passes through the application layer gateway, and inspection of that signaling allows the application layer gateway to dynamically open the necessary UDP pinholes for each VRF for the RTP traffic to pass from the outside of the firewall toward the fusion router. Without the inspection occurring on the firewalls, each RTP stream that originates from an endpoint on the outside is not allowed to pass through the firewall. It is the inspection of the call control signaling that allows the UDP traffic to be forwarded through the firewall.

This deployment model provides a method to allow communication devices on a VRF-enabled network to have peer-to-peer connectivity. The application layer gateway provides secure access to the shared VLAN and the fusion router. All media streams between different VRFs do not take the most direct path between endpoints. The media is backhauled to the data center to be routed via the fusion router.

Scenario 2: Redundant Data Centers

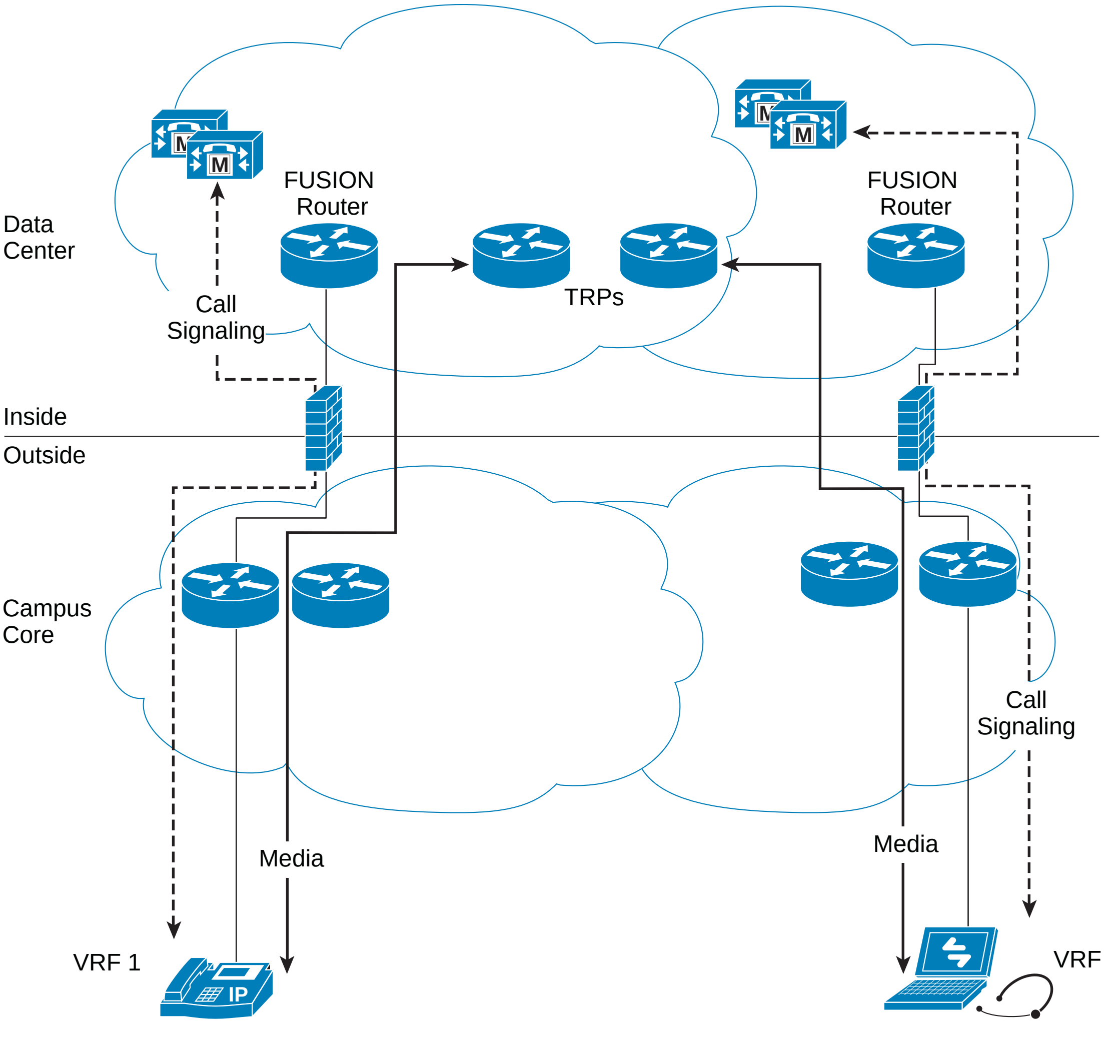
When redundant data centers are involved, the scenario becomes more complicated. It is necessary to ensure that the call setup signaling passes though the same application layer gateway that the corresponding RTP stream is going to use. If the signaling and media take different paths, a UDP pinhole is not opened. Figure 4-18 illustrates an example of a problematic scenario. The hardware phone on the left is controlled by the subscribers in the data center on the left, and the corresponding call control signaling passes through the left firewall. Pinholes are opened in that firewall for the RTP stream. However, the routing might not guarantee that the RTP media stream follows the same path, and the firewall on the right blocks that stream.

Figure 4-18 Call Signaling and Media Take Different Paths



The solution is to utilize Trusted Relay Point (TRP) functionality. (See Figure 4-19.) Subscribers in each data center can invoke TRPs that provide anchoring of the media and ensure that the media streams flow through the appropriate firewall. A phone controlled by a subscriber in the left data center must invoke a TRP in that data center, and a phone controlled by a subscriber in the right data center must invoke a TRP located in the right data center. The TRP provides an IP address that enables a specific host route for media that can ensure the exact same routing path as the call signaling. This is used to ensure that signaling and media pass via the same firewall, thus solving the issue.

Figure 4-19 Redundant Data Centers with TRPs



TRPs are media termination point resources that are invoked at the device level for any call involving that device. Each device has a configuration checkbox that specifies whether a TRP should be invoked.

Conclusion

This chapter did not cover all of the security that could be enabled to protect the voice data within your network. The techniques presented here are just a subset of all the tools that are available to network administrators to protect all the data within a network. On the other hand, even these tools do not have to be enabled within a network, depending on what level of security is required for the data within the network overall. Choose your security methods wisely. As the security within a network increases, so do the complexity and troubleshooting problems. It is up to each enterprise to define both the risks and the requirements of its organization and then to apply the appropriate security within the network and on the devices attached to that network.

CHAPTER 5

Unified Communications Deployment Models

Revised: April 30, 2013; OL-27282-05

This chapter describes the deployment models for Cisco Unified Communications Systems.

Earlier versions of this chapter based the deployment models discussion on the call processing deployment models for Cisco Unified Communications Manager (Unified CM) exclusively. The current version of this chapter, by contrast, introduces a site-based approach to the design guidance for the constituent technologies of the Cisco Unified Communications System. The intent is to offer design guidance for the entire Cisco Unified Communications System, which includes much more than just the call processing service.

For design guidance with earlier releases of Cisco Unified Communications, refer to the Cisco Unified Communications Solution Reference Network Design (SRND) documentation available at

http://www.cisco.com/go/ucsrnd

What's New in This Chapter

Table 5-1 lists the topics that are new in this chapter or that have changed significantly from previous releases of this document.

Table 5-1 New or Changed Information Since the Previous Release of This Document

New or Revised Topic	Described in	Revision Date
Minor correction to Cisco IOS extended ping example	Delay Testing, page 5-36	April 30, 2013
Cisco Unified Survivable Remote Site Telephony (SRST) Manager	Cisco Unified Survivable Remote Site Telephony Manager, page 5-18	August 31, 2012
Minor updates for Cisco Unified Communications System Release 9.0	Various sections throughout this chapter	June 28, 2012

Deployment Model Architecture

In general terms, the deployment model architecture follows that of the enterprise it is deployed to serve. Deployment models describe the reference architecture required to satisfy the Unified Communications needs of well-defined, typical topologies of enterprises. For example, a centralized call processing deployment model caters to enterprises whose operational footprint is based on multiple sites linked to one or few centralized headquarters offices.

In some cases, the deployment model of a technology will depart from that of the enterprise, due to technological constraints. For example, if an enterprise has a single campus whose scale exceeds that of a single service instance (such as a call processing service provided by Cisco Unified Communications Manager), then a single campus might require more than a single instance of a call processing cluster or a single messaging product.

Another option for customers who exceed the sizing limits of a standard cluster is to consider deploying a megacluster, which can provide increased scalability. For more information about megaclusters, see Megacluster, page 8-26.

Note Unless otherwise specified, all information contained within this SRND that relates to call processing deployments (including capacity, high availability, and general design considerations) applies only to a standard cluster.

High Availability for Deployment Models

Unified Communications services offer many capabilities aimed at achieving high availability. They may be implemented in various ways, such as:

- Failover redundancy

  For services that are considered essential, redundant elements should be deployed so that no single point of failure is present in the design. The redundancy between the two (or more) elements is automated. For example, the clustering technology used in Cisco Unified Communications Manager (Unified CM) allows for up to three servers to provide backup for each other. This type of redundancy may cross technological boundaries. For example, a phone may have as its first three preferred call control agents, three separate Unified CM servers belonging to the same call processing cluster. As a fourth choice, the phone can also be configured to rely on a Cisco IOS router for call processing services.

- Redundant links

  In some instances, it is advantageous to deploy redundant IP links, such as IP WAN links, to guard against the failure of a single WAN link.

- Geographical diversity

  Some products support the distribution of redundant service nodes across WAN links so that, if an entire site is off-line (such as would be the case during an extended power outage exceeding the capabilities of provisioned UPS and generator backup systems), another site in a different location can ensure business continuance.

Capacity Planning for Deployment Models

The capacities of various deployment models are typically integrally linked to the capacities of the products upon which they are based. Where appropriate in this chapter, capacities are called out. For some of the products supporting services covered in more detail in other sections of this document, the capacities of those products are discussed in their respective sections.

Site-Based Design

Across all technologies that make up the Cisco Unified Communications System, the following common set of criteria emerges as the main drivers of design:

Size

In this context, size generally refers to the number of users, which translates into a quantity of IP telephones, voice mail boxes, presence watchers, and so forth. Size also can be considered in terms of processing capacity for sites where few (or no) users are present, such as data centers.

Network Connectivity

The site's connectivity into the rest of the system has three main components driving the design:

- Bandwidth enabled for Quality of Service (QoS)
- Latency
- Reliability

These components are often considered adequate in the Local Area Network (LAN): QoS is achievable with all LAN equipment, bandwidth is typically in the Gigabit range, latency is minimal (in the order of a few milliseconds), and excellent reliability is the norm.

The Metropolitan Area Network (MAN) often approaches the LAN in all three dimensions: bandwidth is still typically in the multiple Megabit range, latency is typically in the low tens of milliseconds, and excellent reliability is common. Packet treatment policies are generally available from MAN providers, so that end-to-end QoS is achievable.

The Wide Area Network (WAN) generally requires extra attention to these components: the bandwidth is at a cost premium, the latencies may depend not only on effective serialization speeds but also on actual transmission delays related to physical distance, and the reliability can be impacted by a multitude of factors. The QoS performance can also require extra operational costs and configuration effort.

Bandwidth has great influence on the types of Unified Communications services available at a site, and on the way these services are provided. For example, if a site serving 20 users is connected with 1.5 Mbps of bandwidth to the rest of the system, the site's voice, presence, instant messaging, email, and video services can readily be hosted at a remote datacenter site. If that same site is hosting 1000 users, some of the services would best be hosted locally to avoid saturating the comparatively limited bandwidth with signaling and media flows. Another alternative is to consider increasing the bandwidth to allow services to be delivered across the WAN from a remote datacenter site.

The influence of latency on design varies, based on the type of Unified Communications service considered for remote deployment. If a voice service is hosted across a WAN where the one-way latency is 200 ms, for example, users might experience issues such as delay-to-dialtone or increased media cut-through delays. For other services such as presence, there might be no problem with a 200 ms latency.

Reliability of the site's connectivity into the rest of the network is a fundamental consideration in determining the appropriate deployment model for any technology. When reliability is high, most Unified Communications components allow for the deployment of services hosted from a remote site; when reliability is inconsistent, some Unified Communications components might not perform reliably when hosted remotely; if the reliability is poor, co-location of the Unified Communications services at the site might be required.

High Availability Requirements

The high availability of services is always a design goal. Pragmatic design decisions are required when balancing the need for reliability and the cost of achieving it. The following elements all affect a design's ability to deliver high availability:

- Bandwidth reliability, directly affecting the deployment model for any Unified Communications service
- Power availability

  Power loss is a very disruptive event in any system, not only because it prevents the consumption of services while the power is out, but also because of the ripple effects caused by power restoration. A site with highly available power (for example, a site whose power grid connection is stable, backed-up by uninterruptible power supplies (UPSs) and by generator power) can typically be chosen to host any Unified Communications service. If a site has inconsistent power availability, it would not be judicious to use it as a hosting site.
- Environmental factors such as heat, humidity, vibration, and so forth
- Availability of qualified personnel

  Some Unified Communications services are delivered though the use of equipment such as servers that require periodical maintenance. Some Unified Communications functions such as the hosting of Unified Communications call agent servers are best deployed at sites staffed with qualified personnel.

Site-Based Design Guidance

Throughout this document, design guidance is organized along the lines of the various Unified Communications services and technologies. For instance, the call processing chapter contains not only the actual description of the call processing services, but also design guidance pertaining to deploying IP phones and Cisco Unified Communications servers based on a site's size, network connectivity, and high availability requirements. Likewise, the call admission control chapter focuses on the technical explanation of that technology while also incorporating site-based design considerations.

Generally speaking, most aspects of any given Unified Communications service or technology are applicable to all deployments, no matter the site's size or network connectivity. When applicable, site-based design considerations are called out. Services can be centralized, distributed, inter-networked, and geographically diversified.

Centralized Services

For applications where enterprise branch sites are geographically dispersed and interconnected over a Wide Area Network, the Cisco Unified Communications services can be deployed at a central location while serving endpoints over the WAN connections. For example, the call processing service can be deployed in a centralized manner, requiring only IP connectivity with the remote sites to deliver telephony services. Likewise, voice messaging services, such as those provided by the Cisco Unity Connection platform, can also be provisioned centrally to deliver services to endpoints remotely connected across an IP WAN.

Centrally provisioned Unified Communications services can be impacted by WAN connectivity interruptions; for each service, the available local survivability options should be planned. As an example, the call processing service as offered by Cisco Unified CM can be configured with local survivability functionality such as SRST or Cisco Unified Communications Manager Express (Unified CME). Likewise, a centralized voice messaging service such as that of Cisco Unity Connection can be provisioned to allow remote sites operating under SRST or Unified CME to access voice messaging services at the central site, through the PSTN.

The centralization of services need not be uniform across all Unified Communications services. For example, a system can be deployed where multiple sites rely on a centralized call processing service, but can also be provisioned with a de-centralized (distributed) voice messaging service such as Cisco Unity Express. Likewise, a Unified Communications system could be deployed where call processing is provisioned locally at each site through Cisco Unified Communications Manager Express, with a centralized voice messaging service such as Cisco Unity Connection.

In many cases, the main criteria driving the design for each service are the availability and quality of the IP network between sites. The centralization of Unified Communications services offers advantages of economy of scale in both capital and operational expenses associated with the hosting and operation of equipment in situations where the IP connectivity between sites offers the following characteristics:

- Enough bandwidth for the anticipated traffic load, including peak hour access loads such as those generated by access to voicemail, access to centralized PSTN connectivity, and inter-site on-net communications including voice and video
- High availability, where the WAN service provider adheres to a Service Level Agreement to maintain and restore connectivity promptly
- Low latency, where local events at the remote site will not suffer if the round-trip time to the main central site imparts some delays to the system's response times

Also, when a given service is deployed centrally to serve endpoints at multiple sites, there are often advantages of feature transparency afforded by the use of the same processing resources for users at multiple sites. For example, when two sites are served by the same centralized Cisco Unified Communications Manager cluster, the users can share line appearances between the two sites. This benefit would not be available if each site were served by different (distributed) call processing systems.

These advantages of feature transparency and economies of scale should be evaluated against the relative cost of establishing and operating a WAN network configured to accommodate the demands of Unified Communications traffic.

Distributed Services

Unified Communications services can also be deployed independently over multiple sites, in a distributed fashion. For example, two sites (or more) can be provisioned with independent call processing Cisco Unified CME nodes, with no reliance on the WAN for availability of service to their co-located endpoints. Likewise, sites can be provisioned with independent voice messaging systems such as Cisco Unity Express.

The main advantage of distributing Unified Communications services lies in the independence of the deployment approach from the relative availability and cost of WAN connectivity. For example, if a company operates a site in a remote location where WAN connectivity is not available, is very expensive, or is not reliable, then provisioning an independent call processing node such as Cisco Unified Communications Manager Express within the remote site will avoid any call processing interruptions if the WAN goes down.

Inter-Networking of Services

If two sites are provisioned with independent services, they can still be interconnected to achieve some degree of inter-site feature transparency. For example, a distributed call processing service provisioned through Cisco Unified Communications Manager Express can be inter-networked through H.323 or SIP trunks to permit IP calls between the sites. Likewise, separate instances of Cisco Unity Connection or Cisco Unity Express can partake in the same messaging network to achieve the routing of messages and the exchange of subscriber and directory information within a unified messaging network.

Geographical Diversity of Unified Communications Services

Some services can be provisioned in multiple redundant nodes across the IP WAN, allowing for continued service through site disruptions such as loss of power, network outages, or even compromises in the physical integrity of a site by events such as fire or earthquake.

To achieve such geographical diversity, the individual service must support redundant nodes as well as the deployment of these nodes across the latency and bandwidth constraints of the IP WAN. For example, the call processing service of Unified CM does support the deployment of a single cluster's call processing nodes across an IP WAN as long as the total end-to-end round-trip time between the nodes does not exceed 80 ms and an appropriate quantity of QoS-enabled bandwidth is provisioned. By contrast, Unified CME does not offer redundancy, and thus cannot be deployed in a geographically diverse configuration.

Table 5-2 summarizes the ability of each Cisco Unified Communications service to be deployed in the manners outlined above.

Table 5-2 Available Deployment Options for Cisco Unified Communications Services

Service	Centralized	Distributed	Inter-Networked	Geographical Diversity
Cisco Unified CM	Yes	Yes	Yes	Yes
Cisco Unified CME	No	Yes	Yes	No
Cisco Business Edition 6000	Yes	Yes	Yes	Yes
Cisco Business Edition 5000	Yes	Yes	Yes	No
Cisco Business Edition 3000	Yes	No	No	No
Cisco Unity Express	No	Yes	Yes, with Cisco Unified Messaging Gateway	No
Cisco Unity Connection	Yes	Yes (One Cisco Unity Connection per site)	Yes, with Cisco Unified Messaging Gateway	Yes
Cisco Emergency Responder	Yes	Yes (One Emergency Responder group per site)	Yes, through Emergency Responder clustering	Yes

Table 5-2 Available Deployment Options for Cisco Unified Communications Services (continued)

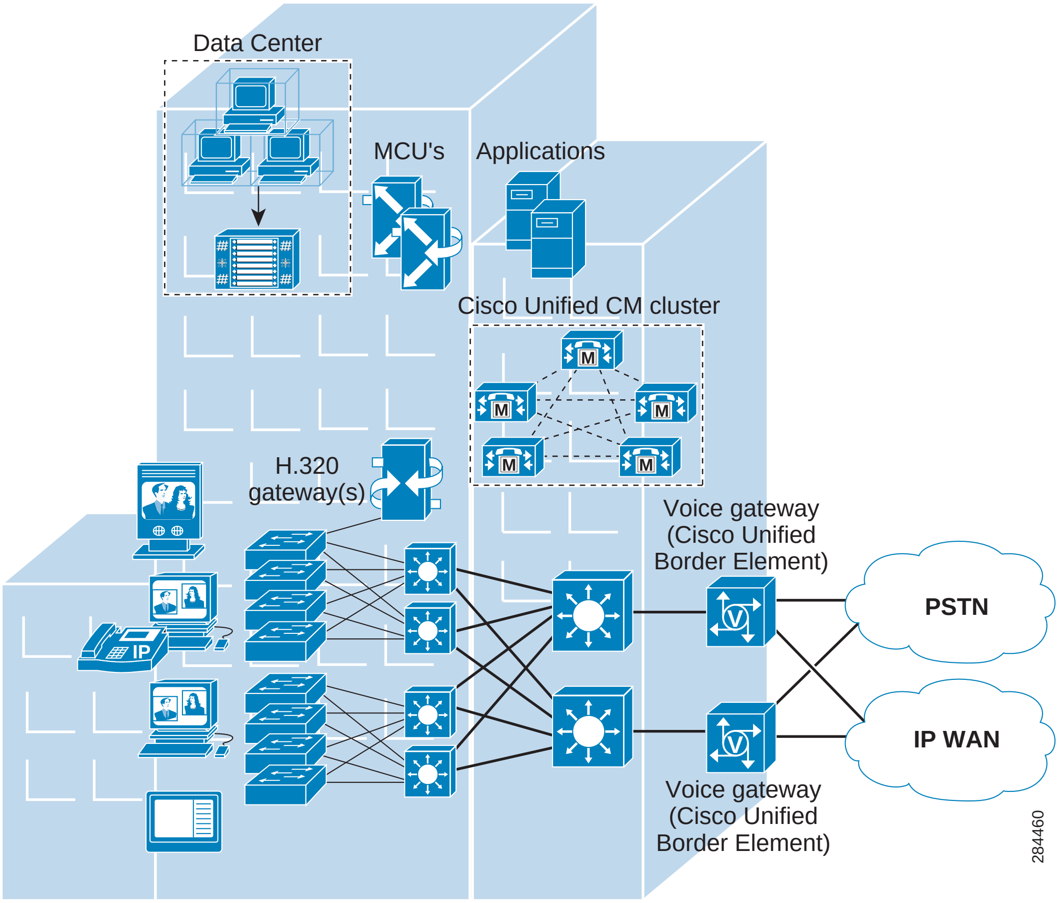
Service	Centralized	Distributed	Inter-Networked	Geographical Diversity
Cisco IM and Presence	Yes	Yes (one Cisco IM and Presence Service per site)	Yes, through inter-domain federation	Yes
Cisco Unified Mobility	Yes	Yes, as Unified CM Single Number Reach	No	Yes

Because call processing is a fundamental service, the basic call processing deployment models are introduced in this chapter. For a detailed technical discussion on Cisco Unified Communications Manager call processing, refer to the chapter on Call Processing, page 8-1.

Campus

In this call processing deployment model, the Unified Communications services and the endpoints are co-located in the campus, and the QoS-enabled network between the service nodes, the endpoints, and applications is considered highly available, offering virtually unlimited bandwidth with less than 15 ms of latency end-to-end. Likewise, the quality and availability of power are very high, and services are hosted in an appropriate data center environment. Communications between the endpoints traverses a LAN or a MAN, and communications outside the enterprise goes over an external network such as the PSTN. An enterprise would typically deploy the campus model over a single building or over a group of buildings connected by a LAN or MAN.

Figure 5-1 Example of a Campus Deployment



The campus model typically has the following design characteristics:

- Single Cisco Unified CM cluster. Some campus call processing deployments may require more than one Unified CM cluster, for instance, if scale calls for more endpoints than can be serviced by a single cluster or if a cluster needs to be dedicated to an application such as a call center.
- Alternatively for smaller deployments, Cisco Business Edition 3000, 5000, or 6000 may be deployed in the campus.
- Maximum of 40,000 configured and registered Skinny Client Control Protocol (SCCP) or Session Initiation Protocol (SIP) IP phones, Cisco Cius, video endpoints, mobile clients, and Cisco Virtualization Experience Clients (VXC) per Unified CM cluster.
- Maximum of 2,100 gateways and trunks (that is, the total number of H.323 gateways, H.323 trunks, digital MGCP devices, and SIP trunks) per Unified CM cluster.
- Trunks and/or gateways (IP or PSTN) for all calls to destinations outside the campus.
- Co-located digital signal processor (DSP) resources for conferencing, transcoding, and media termination point (MTP).
- Other Unified Communications services, such as messaging (voicemail), presence, and mobility are typically co-located.

- Interfaces to legacy voice services such as PBXs and voicemail systems are connected within the campus, with no operational costs associated with bandwidth or connectivity.
- Multipoint Control Unit (MCU) resources are required for multipoint video conferencing. Depending on conferencing requirements, these resources may be either SCCP or H.323, or both.
- H.323 and H.320 video gateways are needed to communicate with H.320 videoconferencing devices on the public ISDN network.
- High-bandwidth audio is available (for example, G.711 or G.722) between devices within the site.
- High-bandwidth video (for example, 384 kbps to 1.5 Mbps) is available between devices within the site.

Best Practices for the Campus Model

Follow these guidelines and best practices when implementing the single-site model:

- Ensure that the infrastructure is highly available, enabled for QoS, and configured to offer resiliency, fast convergence, and inline power.
- Know the calling patterns for your enterprise. Use the campus model if most of the calls from your enterprise are within the same site or to PSTN users outside your enterprise.
- Use G.711 codecs for all endpoints. This practice eliminates the consumption of digital signal processor (DSP) resources for transcoding, and those resources can be allocated to other functions such as conferencing and media termination points (MTPs).
- Implement the recommended network infrastructure for high availability, connectivity options for phones (in-line power), Quality of Service (QoS) mechanisms, and security. (See Network Infrastructure, page 3-1.)
- Follow the provisioning recommendations listed in the chapter on Call Processing, page 8-1.

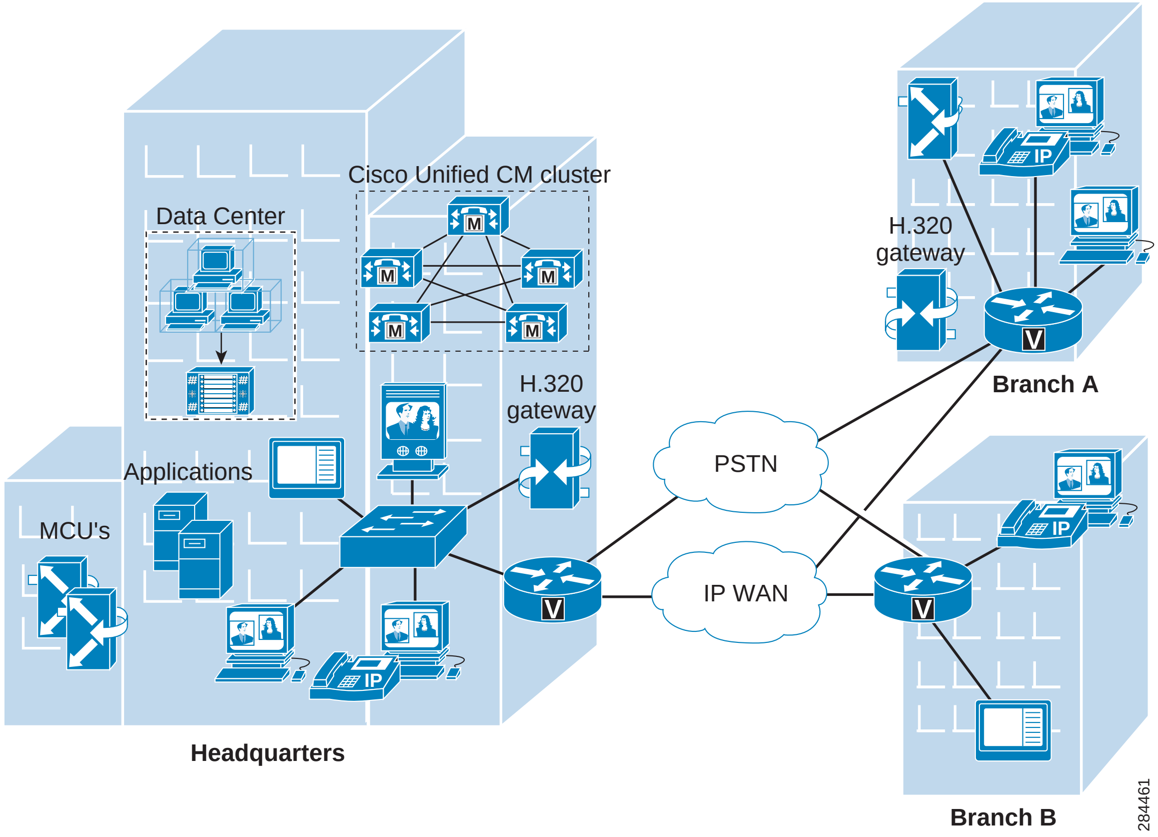
Multisite with Centralized Call Processing

In this call processing deployment model, endpoints are remotely located from the call processing service, across a QoS-enabled Wide Area Network. Due to the limited quantity of bandwidth available across the WAN, a call admission control mechanism is required to manage the number of calls admitted on any given WAN link, to keep the load within the limits of the available bandwidth. On-net communication between the endpoints traverses either a LAN/MAN (when endpoints are located in the same site) or a WAN (when endpoints are located in different sites). Communication outside the enterprise goes over an external network such as the PSTN, through a gateway or Cisco Unified Border Element (CUBE) session border controller (SBC) that can be co-located with the endpoint or at a different location (for example, when using a centralized gateway at the main site or when doing Tail End Hop Off (TEHO) across the enterprise network).

The IP WAN also carries call control signaling between the central site and the remote sites. Figure 5-2 illustrates a typical centralized call processing deployment, with a Unified CM cluster as the call processing agent at the central site and a QoS-enabled IP WAN to connect all the sites. In this deployment model, other Unified Communications services such as voice messaging, presence and mobility are often hosted at the central site as well to reduce the overall costs of administration and maintenance. In situations where the availability of the WAN is unreliable or when WAN bandwidth costs are high, it is possible to consider decentralizing some Unified Communications services such as voice messaging (voicemail) so that the service's availability is not impacted by WAN outages.

Note In each solution for the centralized call processing model presented in this document, the various sites connect to an IP WAN with QoS enabled.

Figure 5-2 Multisite Deployment with Centralized Call Processing



The multisite model with centralized call processing has the following design characteristics:

- Single Unified CM cluster. Some centralized call processing deployments may require more than one Unified CM cluster, for instance, if scale calls for more endpoints than can be serviced by a single cluster or if a cluster needs to be dedicated to an application such as a call center.
- For smaller deployments, Cisco Business Edition 3000 may be deployed in centralized call processing configurations for up to 9 remote sites.
- Cisco Business Edition 5000 may be deployed in centralized call processing configurations for up to 19 remote sites.
- Cisco Business Edition 6000 may be deployed in centralized call processing configurations for up to 49 remote sites.
- Maximum of 40,000 configured and registered Skinny Client Control Protocol (SCCP) or Session Initiation Protocol (SIP) IP phones, Cisco Cius, video endpoints, mobile clients, and Cisco Virtualization Experience Clients (VXC) per Unified CM cluster.
- Maximum of 2,000 locations or branch sites per Unified CM cluster.

- Maximum of 2,100 gateways and trunks (that is, the total number of H.323 gateways, H.323 trunks, digital MGCP devices, and SIP trunks) per Unified CM cluster.
- PSTN connectivity for all off-net calls.
- Digital signal processor (DSP) resources for conferencing, transcoding, and media termination point (MTP) are distributed locally to each site to reduce WAN bandwidth consumption on calls requiring DSPs.
- Capability to integrate with legacy private branch exchange (PBX) and voicemail systems. Interfaces to legacy voice services such as PBXs and voicemail systems can connected within the central site, with no operational costs associated with bandwidth or connectivity. Connectivity to legacy systems located at remote sites may require the operational expenses associated with the provisioning of extra WAN bandwidth.
- MCU resources are required for multipoint video conferencing. Depending on conferencing requirements, these resources may be either SCCP or H.323, or both, and may all be located at the central site or may be distributed to the remote sites if local conferencing resources are required.
- H.323/H.320 video gateways are needed to communicate with H.320 videoconferencing devices on the public ISDN network. These gateways may all be located at the central site or may be distributed to the remote sites if local ISDN access is required.
- The system allows for the automated selection of high-bandwidth audio (for example, G.711 or G.722) between devices within the site, while selecting low-bandwidth audio (for example, G.729) between devices in different sites.
- The system allows for the automated selection of high-bandwidth video (for example, 384 kbps to 1.5 Mbps) between devices in the same site, and low-bandwidth video (for example, 128 kbps) between devices at different sites.
- A minimum of 768 kbps or greater WAN link speed should be used when video is to be placed on the WAN.
- Call admission control is achieved through Enhanced Locations CAC or RSVP.
- For voice and video calls, automated alternate routing (AAR) provides the automated rerouting of calls through the PSTN when call admission control denies a call due to lack of bandwidth. AAR relies on a gateway being available to route the call from the calling phone toward the PSTN, and another gateway to accept the call from the PSTN at the remote site, to be connected to the called phone.
- Call Forward Unregistered (CFUR) functionality provides the automated rerouting of calls through the PSTN when an endpoint is considered unregistered due to a remote WAN link failure. CFUR relies on a gateway being available to route the call from the calling phone toward the PSTN, and another gateway to accept the call from the PSTN at the remote site, to be connected to the called phone.
- Survivable Remote Site Telephony (SRST) for video. SCCP video endpoints located at remote sites become audio-only devices if the WAN connection fails.
- Cisco Unified Communications Manager Express (Unified CME) may be used for remote site survivability instead of an SRST router.
- Cisco Unified Communications Manager Express (Unified CME) can be integrated with the Cisco Unity Connection server in the branch office or remote site. The Cisco Unity Connection server is registered to the Unified CM at the central site in normal mode and can fall back to Unified CME in SRST mode when Unified CM is not reachable, or during a WAN outage, to provide the users at the branch offices with access to their voicemail with MWI.

- As with other call processing types that support multisite centralized call processing, Cisco Business Edition 3000 allows PSTN routing through both central and remote site gateways. Providing a local gateway at remote sites for local PSTN breakout is a necessary requirement for countries providing emergency services for users located at remote sites. The local gateway at the remote site provides call routing to the local PSAP of the remote site location. Local PSTN breakout at remote sites might also be needed or required for countries having strict regulations requiring separation of IP telephony networks from the PSTN. Where regulations allow, local PSTN breakout through the remote site gateway can be used to enable toll bypass or tail-end hop off (TEHO). Business Edition 3000 provides country-based dial plan configuration to enable routing to configured PSTN gateways as well as policy mechanisms to control PSTN access restrictions (as applicable based on local country regulations). Business Edition 3000 supports local PSTN breakout only through the MGCP-controlled Cisco 2901 Integrated Services Router (ISR). Local breakout at a remote site can also be provided through analog trunks using a Cisco SPA8800 IP Telephony Gateway or through SIP trunks using Cisco Unified Border Element on a Cisco SPA8800 or SPA8900 IP Telephony Gateway (sometimes referred to as "CUBE Lite").
- Business Edition 3000 does not support SRST or remote site survivability.

Connectivity options for the IP WAN include:

- Leased lines
- Frame Relay
- Asynchronous Transfer Mode (ATM)
- ATM and Frame Relay Service Inter-Working (SIW)
- Multiprotocol Label Switching (MPLS) Virtual Private Network (VPN)
- Voice and Video Enabled IP Security Protocol (IPSec) VPN (V3PN)

Routers that reside at the WAN edges require quality of service (QoS) mechanisms, such as priority queuing and traffic shaping, to protect the voice traffic from the data traffic across the WAN, where bandwidth is typically scarce. In addition, a call admission control scheme is needed to avoid oversubscribing the WAN links with voice traffic and deteriorating the quality of established calls. For centralized call processing deployments, *locations* (static or RSVP-enabled) configured within Unified CM provide call admission control. (Refer to the chapter on Call Admission Control, page 11-1, for more information on locations.)

A variety of Cisco gateways can provide the remote sites with TDM and/or IP-based PSTN access. When the IP WAN is down, or if all the available bandwidth on the IP WAN has been consumed, calls from users at remote sites can be rerouted through the PSTN. The Cisco Unified Survivable Remote Site Telephony (SRST) feature, available for both SCCP and SIP phones, provides call processing at the branch offices for Cisco Unified IP Phones if they lose their connection to the remote primary, secondary, or tertiary Unified CM or if the WAN connection is down. Cisco Unified SRST functionality is available on Cisco IOS gateways running the SRST feature or on Cisco Unified CME running in SRST mode. Unified CME running in SRST mode provides more features for the phones than SRST on a Cisco IOS gateway.

Best Practices for the Centralized Call Processing Model

Follow these guidelines and best practices when implementing multisite centralized call processing deployments:

- Minimize delay between Unified CM and remote locations to reduce voice cut-through delays (also known as clipping).
- Configure Enhanced Locations CAC or RSVP in Unified CM to provide call admission control into and out of remote branches. See the chapter on Call Admission Control, page 11-1, for details on how to apply this mechanism to the various WAN topologies.
- The number of IP phones and line appearances supported in Survivable Remote Site Telephony (SRST) mode at each remote site depends on the branch router platform, the amount of memory installed, and the Cisco IOS release. SRST on a Cisco IOS gateway supports up to 1,500 phones, while Unified CME running in SRST mode supports 450 phones. (For the latest SRST or Unified CME platform and code specifications, refer to the SRST and Unified CME documentation available at http://www.cisco.com.) Generally speaking, however, the choice of whether to adopt a centralized call processing or distributed call processing approach for a given site depends on a number of factors such as:
  - IP WAN bandwidth or delay limitations
  - Criticality of the voice network
  - Feature set needs
  - Scalability
  - Ease of management
  - Cost

  If a distributed call processing model is deemed more suitable for the customer's business needs, the choices include installing a Unified CM cluster at each site or running Unified CME at the remote sites.
- At the remote sites, use the following features to ensure call processing survivability in the event of a WAN failure:
  - For SCCP phones, use SRST on a Cisco IOS gateway or Unified CME running in SRST mode.
  - For SIP phones, use SIP SRST.
  - For MGCP phones, use MGCP Gateway Fallback.

  SRST or Unified CME in SRST mode, SIP SRST, and MGCP Gateway Fallback can reside with each other on the same Cisco IOS gateway.

Remote Site Survivability

When deploying Cisco Unified Communications across a WAN with the centralized call processing model, you should take additional steps to ensure that data and voice services at the remote sites are highly available. Table 5-3 summarizes the different strategies for providing high availability at the remote sites. The choice of one of these strategies may depend on several factors, such as specific business or application requirements, the priorities associated with highly available data and voice services, and cost considerations.

Table 5-3 Strategies for High Availability at the Remote Sites

Strategy	High Availability for Data Services?	High Availability for Voice Services?
Redundant IP WAN links in branch router	Yes	Yes
Redundant branch router platforms + Redundant IP WAN links	Yes	Yes
Data-only ISDN backup + SRST or Unified CME	Yes	Yes
Data and voice ISDN backup	Yes	Yes (see rules below)
Cisco Unified Survivable Remote Site Telephony (SRST) or Unified CME in SRST mode	No	Yes

The first two solutions listed in Table 5-3 provide high availability at the network infrastructure layer by adding redundancy to the IP WAN access points, thus maintaining IP connectivity between the remote IP phones and the centralized Unified CM at all times. These solutions apply to both data and voice services, and are entirely transparent to the call processing layer. The options range from adding a redundant IP WAN link at the branch router to adding a second branch router platform with a redundant IP WAN link.

The third and forth solutions in Table 5-3 use an ISDN backup link to provide survivability during WAN failures. The two deployment options for ISDN backup are:

- Data-only ISDN backup

  With this option, ISDN is used for data survivability only, while SRST or Unified CME in SRST mode is used for voice survivability. Note that you should configure an access control list on the branch router to prevent traffic from telephony signaling protocols such as Skinny Client Control Protocol (SCCP), H.323, Media Gateway Control Protocol (MGCP), or Session Initiation Protocol (SIP) from entering the ISDN interface, so that signaling from the IP phones does not reach the Unified CM at the central site. This is to ensure that the telephony endpoints located at the branch detect the WAN's failure and rely on local SRST resources.

- Data and voice ISDN backup

  With this option, ISDN is used for both data and voice survivability. In this case, SRST or Unified CME in SRST mode is not used because the IP phones maintain IP connectivity to the Unified CM cluster at all times. However, Cisco recommends that you use ISDN to transport data and voice traffic only if all of the following conditions are true:

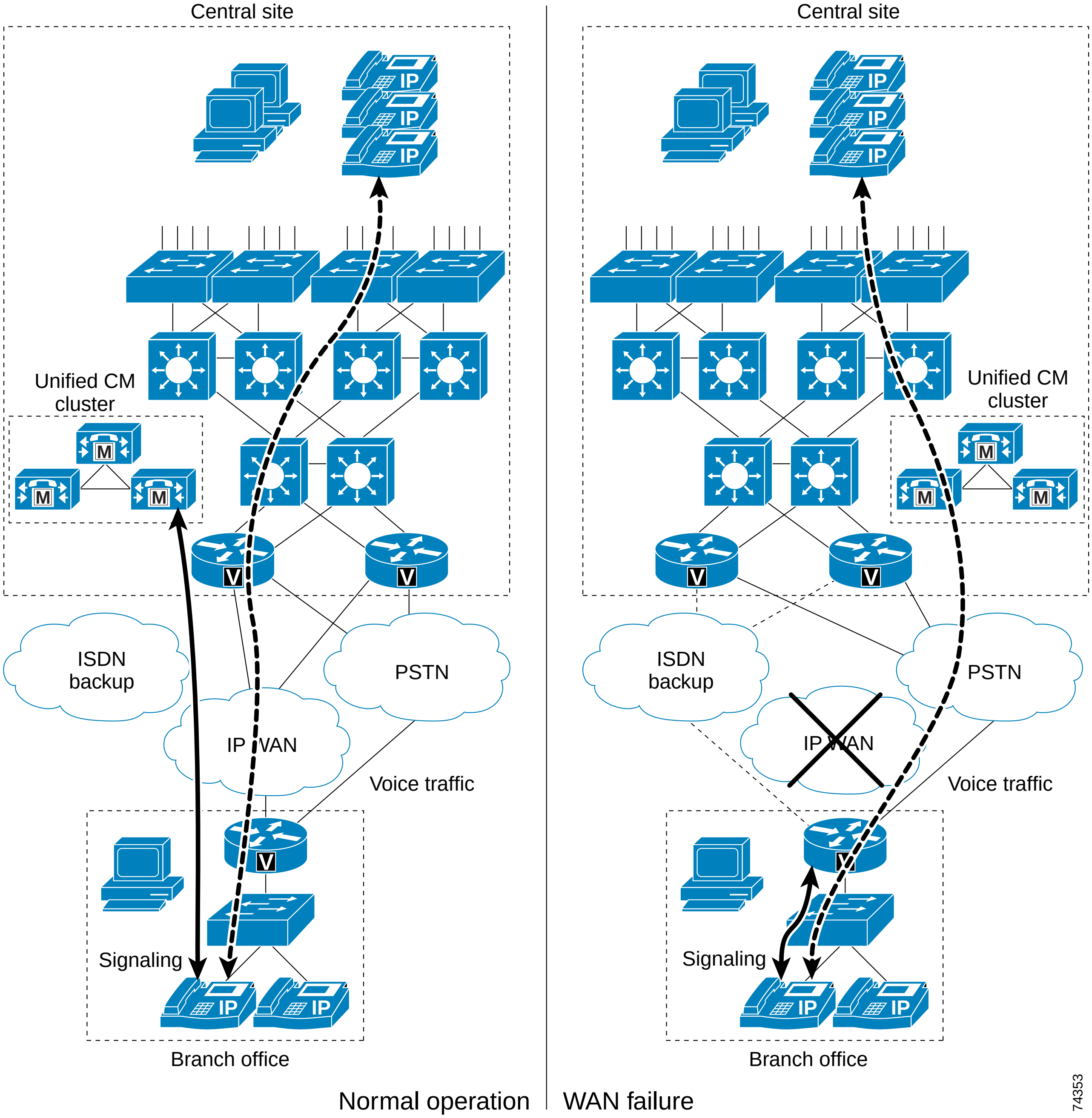
  – The bandwidth allocated to voice traffic on the ISDN link is the same as the bandwidth allocated to voice traffic on the IP WAN link.

  – The ISDN link bandwidth is fixed.

  – All the required QoS features have been deployed on the router's ISDN interfaces. Refer to the chapter on Network Infrastructure, page 3-1, for more details on QoS.

The fifth solution listed in Table 5-3, Survivable Remote Site Telephony (SRST) or Unified CME in SRST mode, provides high availability for voice services only, by providing a subset of the call processing capabilities within the remote office router and enhancing the IP phones with the ability to “re-home” to the call processing functions in the local router if a WAN failure is detected. Figure 5-3 illustrates a typical call scenario with SRST or Unified CME in SRST mode.

Figure 5-3 Survivable Remote Site Telephony (SRST) or Unified CME in SRST Mode



Under normal operations shown in the left part of Figure 5-3, the branch office connects to the central site via an IP WAN, which carries data traffic, voice traffic, and call signaling. The IP phones at the branch office exchange call signaling information with the Unified CM cluster at the central site and place their calls across the IP WAN. The branch router or gateway forwards both types of traffic (call signaling and voice) transparently and has no knowledge of the IP phones.

If the WAN link to the branch office fails, or if some other event causes loss of connectivity to the Unified CM cluster, the branch IP phones re-register with the branch router in SRST mode. The branch router, SRST, or Unified CME running in SRST mode, queries the IP phones for their configuration and uses this information to build its own configuration automatically. The branch IP phones can then make and receive calls either within the branch's network or through the PSTN. The phone displays the message "Unified CM fallback mode," and some advanced Unified CM features are unavailable and are grayed out on the phone display.

When WAN connectivity to the central site is reestablished, the branch IP phones automatically re-register with the Unified CM cluster and resume normal operation. The branch SRST router deletes its information about the IP phones and reverts to its standard routing or gateway configuration. Unified CME running in SRST mode at the branch can choose to save the learned phone and line configuration to the running configuration on the Unified CME router by using the auto-provision option. If **auto-provision none** is configured, none of the auto-provisioned phone or line configuration information is written to the running configuration of the Unified CME router. Hence, no configuration change is required on Unified CME if the IP phone is replaced and the MAC address changes.

Note When WAN connectivity to the central site is reestablished, or when Unified CM is reachable again, phones in SRST mode with active calls will not immediately re-register to Unified CM until those active calls are terminated.

Note The remote site survivability features explained above are not supported with Business Edition 3000.

Unified CME in SRST Mode

When Unified CME is used in SRST mode, it provides more call processing features for the IP phones than are available with the SRST feature on a router. In addition to the SRST features such as call preservation, auto-provisioning, and failover, Unified CME in SRST mode also provides most of the Unified CME telephony features for the SCCP phones, including:

- Paging
- Conferencing
- Hunt groups
- Basic automatic call distribution (B-ACD)
- Call park, call pickup, call pickup groups
- Overlay-DN, softkey templates
- Cisco IP Communicator
- Cisco Unified Video Advantage
- Integration with Cisco Unity with MWI support at remote sites, with distributed Microsoft Exchange or IBM Lotus Domino server

Unified CME in SRST mode provides call processing support for SCCP phones in case of a WAN failure. However, Unified CME in SRST mode does not provide fallback support for MGCP phones or endpoints. To enable SIP and MGCP phones to fall back if they lose their connection to the SIP proxy server or Unified CM, or if the WAN connection fails, you can additionally configure both the SIP SRST feature and the MGCP Gateway Fallback feature on the same Unified CME server running as the SRST fallback server.

Best Practices for Unified CME in SRST Mode

- Use the Unified CME IP address as the IP address for SRST reference in the Unified CM configuration.
- The Connection Monitor Duration is a timer that specifies how long phones monitor the WAN link before initiating a fallback from SRST to Unified CM. The default setting of 120 seconds should be used in most cases. However, to prevent phones in SRST mode from falling back and re-homing to Unified CM with flapping links, you can set the Connection Monitor Duration parameter on Unified CM to a longer period so that phones do not keep registering back and forth between the SRST router and Unified CM. Do not set the value to an extensively longer period because this will prevent the phones from falling back from SRST to Unified CM for a long amount of time.
- Phones in SRST fallback mode will not re-home to Unified CM when they are in active state.
- Phones in SRST fallback mode revert to non-secure mode from secure conferencing.
- Configure **auto-provision none** to prevent writing any learned ephone-dn or ephone configuration to the running configuration of the Unified CME router. This eliminates the need to change the configuration if the IP phone is replaced or the MAC address changes.

For more information on using Unified CME in SRST mode, refer to the *Cisco Unified Communications Manager Express System Administrator Guide*, available at

http://www.cisco.com/en/US/products/sw/voicesw/ps4625/products_installation_and_configuration_guides_list.html

For more information on SIP SRST, refer to the *Cisco Unified SIP SRST System Administrator Guide*, available at

http://www.cisco.com/en/US/products/sw/voicesw/ps2169/products_installation_and_configuration_guides_list.html

For more information on MGCP Gateway fallback, refer to the information on MGCP gateway fallback in the *Cisco CallManager and Cisco IOS Interoperability Guide*, available at

http://www.cisco.com/en/US/docs/ios/12_3/vvf_c/interop/ccm_c.html

Best Practices for SRST Router

Use a Cisco Unified SRST router, rather than Unified CME in SRST mode, for the following deployment scenarios:

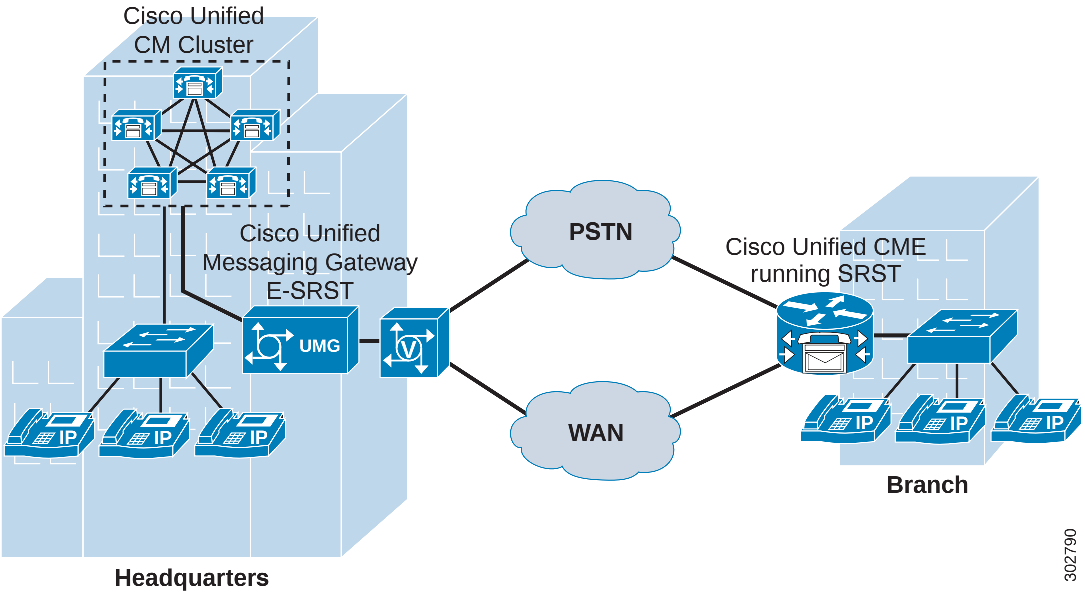
- For supporting a maximum of 1,500 phones on a single SRST router. (Unified CME in SRST mode supports a maximum of 450 phones.)
- For up to 3,000 phones, use two SRST routers. Dial plans must be properly configured to route the calls back and forth between the SRST routers.
- For simple, one-time configuration of basic SRST functions.
- For SRTP media encryption, which is available only in Cisco Unified SRST (Secure SRST).
- For support of the Cisco VG248 Voice Gateway.

For routing calls to and from phones that are unreachable or not registered to the SRST router, use the **alias** command.

Cisco Unified Survivable Remote Site Telephony Manager

Cisco Unified Survivable Remote Site Telephony (SRST) Manager simplifies the deployment of Cisco Unified CME running SRST as well as traditional SRST in the branch. (See Figure 5-4.) It is Linux-based software running inside a virtual machine on Cisco supported virtualized platforms (for example, Cisco UCS). Cisco Unified SRST Manager supports only the centralized call processing deployment model, where the Cisco Unified CM cluster runs in the central location. Cisco Unified SRST Manager can be deployed in the central location along with the Cisco Unified CM cluster or in the remote branch location. Figure 5-4 illustrates the deployment of Cisco Unified SRST Manager in the central location. During normal operation, Cisco Unified SRST Manager regularly retrieves configurations (for example, calling search space, partition, hunt group, call park, call pickup, and so forth, if configured) from Cisco Unified CM and uploads them to provision the branch router with similar functionality for use in SRST mode. Thus, Cisco Unified SRST Manager reduces manual configuration required in the branch SRST router and enables users to have a similar calling experience in both SRST and normal modes.

Figure 5-4 Cisco Unified Survivable Remote Site Telephony Manager Deployed in the Central Location



Cisco Unified SRST Manager consumes bandwidth from the WAN link when uploading the Unified CM configurations to provision the branch router. The Cisco Unified SRST Manager software does not perform packet marking, therefore the Cisco Unified SRST Manager traffic will travel as best-effort on the network. Cisco recommends maintaining this best-effort marking, which is IP Precedence 0 (DSCP 0 or PHB BE), to ensure that it does not interfere with real-time high priority voice traffic. To ensure that Cisco Unified SRST Manager traffic does not cause congestion and to reduce the chances of packet drop, Cisco recommends scheduling the configuration upload to take place during non-peak hours (for example, in the evening hours or during the weekend). The configuration upload schedule can be set from the Cisco Unified SRST Manager web interface.

Consider the following guidelines when you deploy Cisco Unified SRST Manager:

- Cisco Unified SRST Manager is not supported with the Cisco Unified Communications 500 Series platform or the Cisco Business Edition 3000 and 5000 platforms.
- The branch voice gateway must be co-resident with (reside on) the SRST router.
- There is no high availability support with Cisco Unified SRST Manager. If Cisco Unified SRST Manager is unavailable, configuration upload is not possible.
- Cisco Unified SRST Manager is not supported in deployments where NAT is used between the headquarters and branch locations.

Voice Over the PSTN as a Variant of Centralized Call Processing

Centralized call processing deployments can be adapted so that inter-site voice media is sent over the PSTN instead of the WAN. With this configuration, the signaling (call control) of all telephony endpoints is still controlled by the central Unified CM cluster, therefore this Voice over the PSTN (VoPSTN) model variation still requires a QoS-enabled WAN with appropriate bandwidth configured for the signaling traffic.

You can implement VoPSTN in one of the following ways:

- Using the automated alternate routing (AAR) feature. (For more information on AAR, see the section on Automated Alternate Routing, page 9-117.)
- Using a combination of dial plan constructs in both Unified CM and the PSTN gateways.

VoPSTN can be an attractive option in deployments where IP WAN bandwidth is either scarce or expensive with respect to PSTN charges, or where IP WAN bandwidth upgrades are planned for a later date but the Cisco Unified Communications system is already being deployed.

Note VoPSTN deployments offer basic voice functionality that is a reduced subset of the Unified CM feature set.

In particular, regardless of the implementation choice, the system designer should address the following issues, among others:

- Centralized voicemail requires:
  - A telephony network provider that supports redirected dialed number identification service (RDNIS) end-to-end for all locations that are part of the deployment. RDNIS is required so that calls redirected to voicemail carry the redirecting DN, to ensure proper voicemail box selection.
  - If the voicemail system is accessed through an MGCP gateway, the voicemail pilot number must be a fully qualified E.164 number.
- The Extension Mobility feature is limited to IP phones contained within a single branch site.
- All on-net (intra-cluster) calls will be delivered to the destination phone with the same call treatment as an off-net (PSTN) call. This includes the quantity of digits delivered in the call directories such as Missed Calls and Received Calls.
- Each inter-branch call generates two independent call detail records (CDRs): one for the call leg from the calling phone to the PSTN, and the other for the call leg from the PSTN to the called phone.
- There is no way to distinguish the ring type for on-net and off-net calls.
- All destination phones require a fully qualified Direct Inward Dial (DID) PSTN number that can be called directly. Non-DID DNs cannot be reached directly from a different branch site.

- With VoPSTN, music on hold (MoH) is limited to cases where the holding party is co-located with the MoH resource. If MoH servers are deployed at the central site, then only calls placed on hold by devices at the central site will receive the hold music.
- Transfers to a destination outside the branch site will result in the hairpinning of the call through the branch's gateway. Traffic engineering of the branch's gateway resources must be adjusted accordingly.
- Call forwarding of any call coming into the branch's gateway to a destination outside the branch site will result in hairpinning of the call through the gateway, thus using two trunk ports. This behavior applies to:
  - Calls forwarded to a voicemail system located outside the branch
  - Calls forwarded to an on-net abbreviated dialing destination located in a different branch

  The gateway port utilization resulting from these call forwarding flows should be taken into account when sizing the trunks connecting the branch to the PSTN.
- Conferencing resources must be co-located with the phone initiating the conference.
- VoPSTN does not support applications that require streaming of IP audio from the central site (that is, not traversing a gateway). These applications include, but are not limited to:
  - Centralized music on hold (MoH) servers
  - Interactive Voice Response (IVR)
  - CTI-based applications
- Use of the Attendant Console outside of the central site can require a considerable amount of bandwidth if the remote sites must access large user account directories without caching them.
- Because all inter-branch media (including transfers) are sent through the PSTN, the gateway trunk group must be sized to accommodate all inter-branch traffic, transfers, and centralized voicemail access.
- Cisco recommends that you do not deploy shared lines across branches, such that the devices sharing the line are in different branches.

In addition to these general considerations, the following sections present recommendations and issues specific to each of the following implementation methods:

- VoPSTN Using AAR, page 5-20
- VoPSTN Using Dial Plan, page 5-22

VoPSTN Using AAR

This method consists of configuring the Unified CM dial plan as in a traditional centralized call processing deployment, with the automated alternate routing (AAR) feature also properly configured. AAR provides transparent re-routing over the PSTN of inter-site calls when the locations mechanism for call admission control determines that there is not enough available WAN bandwidth to accept an additional call.

To use the PSTN as the primary (and only) voice path, you can configure the call admission control bandwidth of each location (branch site) to be 1 kbps, thus preventing *all* calls from traversing the WAN. With this configuration, all inter-site calls trigger the AAR functionality, which automatically re-routes the calls over the PSTN.

The AAR implementation method for VoPSTN offers the following benefits:

- An easy migration path to a complete Cisco Unified Communications deployment. When bandwidth becomes available to support voice media over the WAN, the dial plan can be maintained intact, and the only change needed is to update the location bandwidth value for each site.
- Support for some supplementary features, such as callback on busy.

In addition to the general considerations listed for VoPSTN, the following design guidelines apply to the AAR implementation method:

- AAR functionality must be configured properly.
- As a general rule, supported call initiation devices include IP phones, gateways, and line-side gateway-driven analog phones.
- Inter-branch calls can use AAR only if the destination devices are IP phones or Cisco Unity ports.
- Inter-branch calls to other endpoints must use a fully qualified E.164 number.
- All on-net, inter-branch calls will display the message, "Network congestion, rerouting."
- If destination phones become unregistered (for example, due to WAN connectivity interruption), AAR functionality will not be invoked and abbreviated dialing will be possible only if Call Forward Unregistered (CFUR) is configured. If the destination phone has registered with an SRST router, then it can also be reached by directly dialing its PSTN DID number.
- If originating phones become unregistered (for example, due to WAN connectivity interruption), they will go into SRST (or Unified CME as SRST) mode. To preserve abbreviated dialing functionality under these conditions, configure the SRST (or Unified CME as SRST) router with an appropriate set of translation rules to match the abbreviated dialing form of the destination and translate it into the form required by the PSTN to route calls to the destination.
- Shared lines within the same branch should be configured in a partition included only in that branch's calling search spaces. Inter-site access to the shared line requires one of the following:
  - The originating site dials the DID number of the shared line.
  - If inter-site abbreviated dialing to the shared line is desired, use a translation pattern that expands the user-dialed abbreviated string to the DID number of the shared line.

Note In this case, direct dialing of the shared line's DN from another branch would trigger multiple AAR-based PSTN calls.

VoPSTN Using Dial Plan

This method relies on a specific dial plan configuration within Unified CM and the PSTN gateways to route all inter-site calls over the PSTN. The dial plan must place IP phone DN's at each site into a different partition, and their calling search space must provide access only to the site's internal partition and a set of route patterns that point to the local PSTN gateway.

Abbreviated inter-site dialing can still be provided via a set of translations at each branch site, one for each of the other branch sites. These translations are best accomplished with H.323 gateways and translation rules within Cisco IOS.

The dial plan method for implementing VoPSTN offers the following benefits:

- Easier configuration because AAR is not needed.
- Abbreviated dialing automatically works even under WAN failure conditions on either the originating or destination side, because the Cisco IOS translation rules within the H.323 gateway are effective in SRST mode.

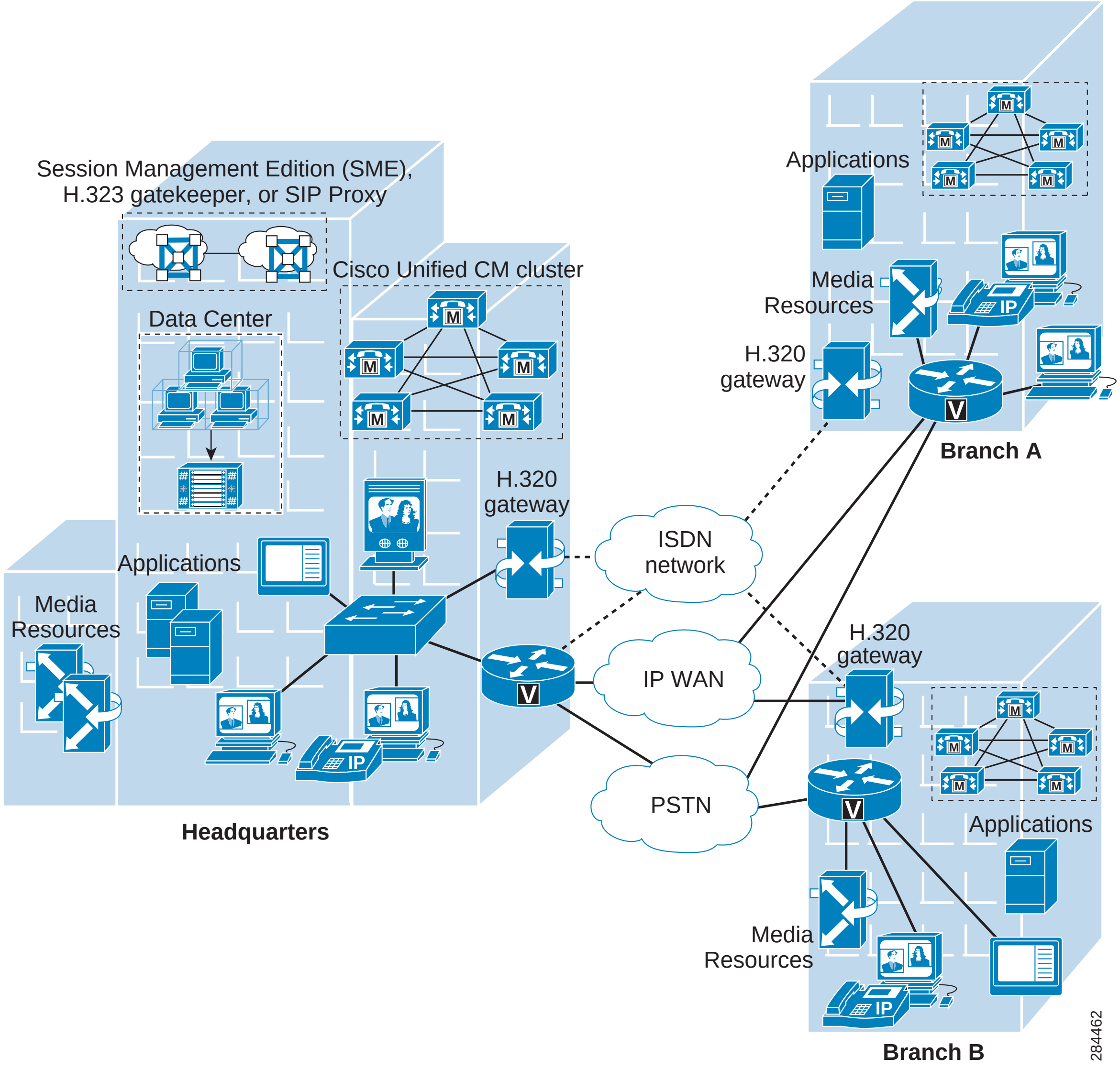
In addition to the general considerations listed for VoPSTN, the following design guidelines apply to the dial plan implementation method:

- There is no support for supplementary features such as callback on busy.
- Some CTI-based applications do not support overlapping extensions (that is, two or more phones configured with the same DN, although in different partitions).
- There is no easy migration to a complete Cisco Unified Communications deployment because the dial plan needs to be redesigned.

Multisite with Distributed Call Processing

The model for a multisite deployment with distributed call processing consists of multiple independent sites, each with its own call processing agent cluster connected to an IP WAN that carries voice traffic between the distributed sites. Figure 5-5 illustrates a typical distributed call processing deployment.

Figure 5-5 Multisite Deployment with Distributed Call Processing



Each site in the distributed call processing model can be one of the following:

- A single site with its own call processing agent, which can be either:
  - Cisco Unified Communications Manager (Unified CM)
  - Cisco Business Edition 5000 and Business Edition 6000
  - Cisco Unified Communications Manager Express (Unified CME)

  - Other IP PBX
- A centralized call processing site and all of its associated remote sites
- A legacy PBX with Voice over IP (VoIP) gateway

The multisite model with distributed call processing has the following design characteristics:

- Maximum of 40,000 configured and registered Skinny Client Control Protocol (SCCP) or Session Initiation Protocol (SIP) IP phones, Cisco Cius, video endpoints, mobile clients, and Cisco Virtualization Experience Clients (VXC) per Unified CM cluster.
- Maximum of 2,100 gateways and trunks (that is, the total number of H.323 gateways, H.323 trunks, digital MGCP devices, and SIP trunks) per Unified CM cluster.
- TDM or IP-based PSTN for all external calls.
- Digital signal processor (DSP) resources for conferencing, transcoding, and media termination point (MTP) are distributed locally to each site to reduce WAN bandwidth consumption on calls requiring DSPs.
- Voicemail, unified messaging, and Cisco IM and Presence components.
- Capability to integrate with legacy private branch exchange (PBX) and voicemail systems.
- Cisco Unified Communications Manager Session Management Edition (SME) clusters, H.323 gatekeepers, or Session Initiation Protocol (SIP) proxy servers can be used to provide intercluster call routing and dial plan aggregation in multisite distributed call processing deployments.
- MCU resources are required in each cluster for multipoint video conferencing. Depending on conferencing requirements, these resources may be either SCCP or H.323, or both, and may all be located at the regional sites or may be distributed to the remote sites of each cluster if local conferencing resources are required.
- H.323/H.320 video gateways are needed to communicate with H.320 videoconferencing devices on the public ISDN network. These gateways may all be located at the regional sites or may be distributed to the remote sites of each cluster if local ISDN access is required.
- High-bandwidth audio (for example, G.711 or G.722) between devices in the same site, but low-bandwidth audio (for example, G.729) between devices in different sites.
- High-bandwidth video (for example, 384 kbps to 1.5 Mbps) between devices in the same site, but low-bandwidth video (for example, 128 kbps) between devices at different sites.
- Minimum of 768 kbps or greater WAN link speeds. Video is *not* recommended on WAN connections that operate at speeds lower than 768 kbps.
- Call admission control is achieved through Enhanced Locations CAC or RSVP.

An IP WAN interconnects all the distributed call processing sites. Typically, the PSTN serves as a backup connection between the sites in case the IP WAN connection fails or does not have any more available bandwidth. A site connected only through the PSTN is a standalone site and is not covered by the distributed call processing model. (See Campus, page 5-7.)

Connectivity options for the IP WAN include:

- Leased lines
- Frame Relay
- Asynchronous Transfer Mode (ATM)
- ATM and Frame Relay Service Inter-Working (SIW)
- Multiprotocol Label Switching (MPLS) Virtual Private Network (VPN)
- Voice and Video Enabled IP Security Protocol (IPSec) VPN (V3PN)

Best Practices for the Distributed Call Processing Model

A multisite deployment with distributed call processing has many of the same requirements as a single site or a multisite deployment with centralized call processing. Follow the best practices from these other models in addition to the ones listed here for the distributed call processing model. (See Campus, page 5-7, and Multisite with Centralized Call Processing, page 5-9.)

Cisco Unified Communications Manager Session Management Edition clusters, H.323 gatekeepers, or Session Initiation Protocol (SIP) proxy servers can be used to provide intercluster call routing and dial plan aggregation in multisite distributed call processing deployments. The following best practices apply to the use of these dial plan aggregation devices:

Unified CM Session Management Edition Clusters

Cisco Unified Communications Manager Session Management Edition is commonly used for intercluster call routing and dial plan aggregation in distributed call processing deployments. Intercluster call routing can be number based using standard numeric route patterns and/or URI based using the Intercluster Look-up Service (ILS). Unified CM Session Management Edition supports multiple protocols (SIP, H.323, MGCP, and SCCP), has sophisticated trunk and digit manipulation features, supports Enhanced Locations CAC and RSVP, and uses the same code and user interface as Unified CM. Unified CM Session Management Edition cluster deployments typically consist of many trunks and no (or very few) Unified Communications endpoints. Unified CM Session Management Edition clusters can use all of the high availability features (such as clustering over the WAN, CallManager Groups, and Run on all Unified CM Nodes) that are available to Unified CM clusters.

For detailed information on Unified CM Session Management Edition cluster deployments, refer to the *Cisco Unified Communizations Manager Session Management Edition Deployment Guide*, available at

http://www.cisco.com/en/US/products/ps10661/products_implementation_design_guides_list.html

Gatekeeper Deployments

- Cisco IOS gatekeepers can be used to provide call admission control into and out of each site.
- To provide high availability of the gatekeeper, use Hot Standby Router Protocol (HSRP) gatekeeper pairs, gatekeeper clustering, and alternate gatekeeper support. In addition, use multiple gatekeepers to provide redundancy within the network. (See Gatekeeper Design Considerations, page 8-37.)
- Size the platforms appropriately to ensure that performance and capacity requirements can be met.
- Use only one type of codec on the WAN because the H.323 specification does not allow for Layer 2, IP, User Data Protocol (UDP), or Real-time Transport Protocol (RTP) header overhead in the bandwidth request. (Header overhead is allowed only in the payload or encoded voice part of the packet.) Using one type of codec on the WAN simplifies capacity planning by eliminating the need to over-provision the IP WAN to allow for the worst-case scenario.

For more information on the various functions performed by gatekeepers, refer to the following sections:

- For gatekeeper call admission control, see Call Admission Control, page 11-1.
- For gatekeeper scalability and redundancy, see Call Processing, page 8-1.
- For gatekeeper dial plan resolution, see Dial Plan, page 9-1.

SIP Proxy Depoloyments

SIP proxies such as the Cisco Unified SIP Proxy provide call routing and SIP signaling normalization.

The following best practices apply to the use of SIP proxies:

- Provide adequate redundancy for the SIP proxies.
- Ensure that the SIP proxies have the capacity for the call rate and number of calls required in the network.
- Planning for call admission control is outside the scope of this document.

Call Processing Agents for the Distributed Call Processing Model

Your choice of call processing agent will vary, based on many factors. The main factors, for the purpose of design, are the size of the site and the functionality required.

For a distributed call processing deployment, each site has its own call processing agent. The design of each site varies with the call processing agent, the functionality required, and the fault tolerance required. For example, in a site with 500 phones, a Unified CM cluster containing two servers can provide one-to-one redundancy, with the backup server being used as a publisher and Trivial File Transfer Protocol (TFTP) server.

The requirement for IP-based applications also greatly affects the choice of call processing agent because only Unified CM provides the required support for many Cisco IP applications.

Table 5-4 lists recommended call processing agents.

Table 5-4 Recommended Call Processing Agents

Call Processing Agent	Recommended Size	Comments
Cisco Unified Communications Manager Express (Unified CME)	Up to 450 phones	• For small remote sites • Capacity depends on Cisco IOS platform
Cisco Business Edition 5000	Up to 575 phones	• For small sites • Supports centralized or distributed call processing
Cisco Business Edition 6000	Up to 1,200 phones	• For small to medium sites • Supports centralized or distributed call processing
Cisco Unified Communications Manager (Unified CM)	50 to 40,000 phones	• Small to large sites, depending on the size of the Unified CM cluster • Supports centralized or distributed call processing
Legacy PBX with VoIP gateway	Depends on PBX	• Number of IP WAN calls and functionality depend on the PBX-to-VoIP gateway protocol and the gateway platform

Unified CM Session Management Edition

Unified Communications deployments using Cisco Unified Communications Manager Session Management Edition are a variation of the multisite distributed call processing deployment model and are typically employed to interconnect large numbers of unified communications systems through a single front-end system, in this case the Unified CM Session Management Edition. This section discusses the relevant design considerations for deploying Unified CM Session Management Edition.

Cisco Unified CM Session Management Edition is essentially a Unified CM cluster with trunk interfaces only and no IP endpoints. It enables aggregation of multiple unified communications systems, referred to as leaf systems.

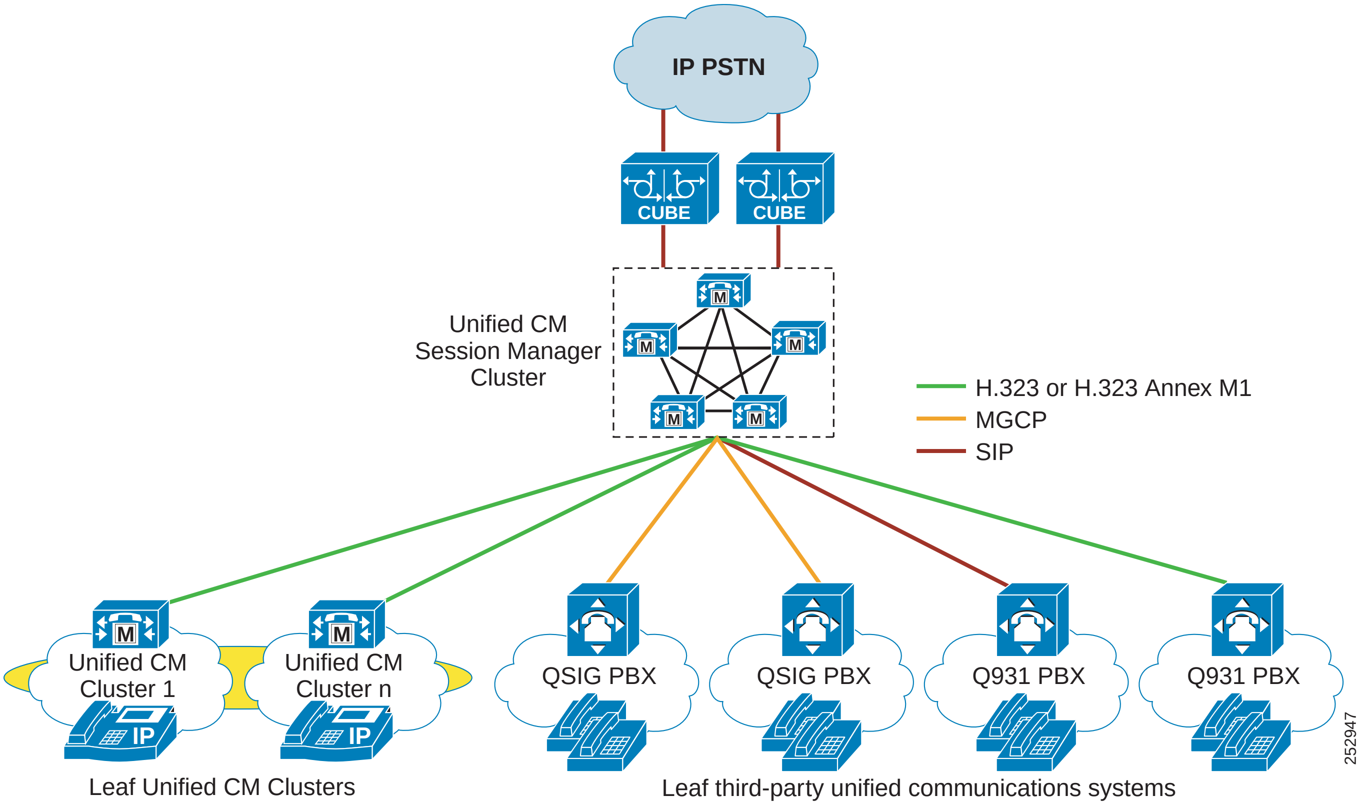
Session Management Edition deployments can be used to migrate a deployment of multiple PBXs and associated phones to a Unified CM cluster with IP phones and relatively few trunks. The Session Management Edition cluster may start with a large number of trunks interconnecting third-party PBXs; and migrate over time to a Unified CM cluster deployment with thousands of IP phones.

With Cisco Unified CM 8.0 and later releases, Unified CM Session Management Edition supports the following features:

- H.323 Annex M1 intercluster trunks
- SIP intercluster trunks
- SIP trunks
- H.323 trunks
- MGCP trunks
- Voice calls
- Video calls
- Encrypted calls
- Fax calls

Unified CM Session Management Edition may also be used to connect to third-party unified communications systems such as PSTN connections, PBXs, and centralized unified communications applications. (See Figure 5-6.) However, as with any standard Unified CM cluster, third-party connections to Unified CM Session Management Edition should be system tested for interoperability prior to use in a production environment.

Figure 5-6 Multisite Deployment with Unified CM Session Management Edition



When to Deploy Unified CM Session Management Edition

Cisco recommends deploying Unified CM Session Management Edition if you want to do any of the following:

- Create and manage a centralized dial plan

  Rather than configuring each unified communications system with a separate dial plan and trunks to connect to all the other unified communications systems, Unified CM Session Management Edition allows you to configure the leaf unified communications systems with a simplified dial plan and trunk(s) pointing to the Session Management cluster. Unified CM Session Management Edition holds the centralized dial plan and corresponding reachability information about all the other unified communications systems.

- Provide centralized PSTN access

  Unified CM Session Management Edition can be used to aggregate PSTN access to one (or more) centralized PSTN trunks. Centralized PSTN access is commonly combined with the reduction, or elimination, of branch-based PSTN circuits.

- Centralize applications

  The deployment of a Unified CM Session Management Edition enables commonly used applications such as conferencing or videoconferencing to connect directly to the Session Management cluster, thus reducing the overhead of managing multiple trunks to leaf systems.

- Aggregate PBXs for migration to a Unified Communications system

  Unified CM Session Management Edition can provide an aggregation point for multiple PBXs as part of the migration from legacy PBXs to a Cisco Unified Communications System.

Differences Between Unified CM Session Management Edition and Standard Unified CM Clusters

The Unified CM Session Management Edition software is exactly the same as Unified CM. However, the software has been enhanced significantly to satisfy the requirements and the constraints of this new deployment model. Unified CM Session Management Edition is designed to support a large number of trunk-to-trunk connections, and as such it is subject to the following design considerations:

- Capacity

  It is important to correctly size the Unified CM Session Management cluster based on the expected BHCA traffic load between leaf Unified Communications systems (for example, Unified CM clusters and PBXs), to and from any centralized PSTN connections, and to any centralized applications. Determine the average BHCA and Call Holding Time for users of your Unified Communications system and share this information with your Cisco account Systems Engineer (SE) or Cisco Partner to size your Unified CM Session Management Edition cluster correctly.

- Trunks

  Where possible, avoid the use of static MTPs on Unified CM trunks (do not enable **MTP required** on the SIP or H.323 trunks of leaf Unified CM or Unified CM Session Management Edition clusters). Trunks that do not use “MTP required” offer more codec choices; support voice, video, and encryption; and do not anchor trunk calls to MTP resources. Dynamically inserted MTPs can be used on trunks (for example, for DTMF translation from in-band to out-of-band). If SIP Early Offer is required by a third-party unified communications system, use either the “Early Offer support for voice and video calls (insert MTP if needed)” on Unified CM SIP trunks or the Delayed Offer to Early Offer feature with Cisco Unified Border Element.

- Unified CM versions

  Both the Unified CM Session Management Edition and Unified CM leaf clusters should be deployed with Cisco Unified CM 7.1(2) or later release. Cisco Unified CM 8.5 or later release is recommended because those versions include features that improve and simplify call routing through Unified CM and Session Management Edition clusters. Earlier versions of Unified CM can be deployed but might experience problems that can be resolved only by upgrading your cluster to Unified CM 7.1(2) or later release.

- Interoperability

  Even though most vendors do conform to standards, differences can and do exist between protocol implementations from various vendors. As with any standard Unified CM cluster, Cisco strongly recommends that you conduct end-to-end system interoperability testing with any unverified third-party unified communications system before deploying the system in a production environment. The interoperability testing should verify call flows and features from Cisco and third-party leaf systems through the Unified CM Session Management cluster. To learn which third-party unified communications systems have been tested by the Cisco Interoperability team, refer to the information available on the Cisco Interoperability Portal at

  http://www.cisco.com/en/US/solutions/ns340/ns414/ns728/interOp_ucSessionMgr.html

- Load balancing for inbound and outbound calls

  Configure trunks on the Unified CM Session Management Edition and leaf unified communications systems so that inbound and outbound calls are evenly distributed across the Unified CM servers within the Session Management cluster. For more information on load balancing for trunk calls, refer to the chapter on Cisco Unified CM Trunks, page 14-1.

- Design guidance and assistance

  For detailed information on Unified CM Session Management Edition designs and deployments, refer to the *Cisco Unified Communications Manager Session Management Edition Deployment Guide*, available at

  http://www.cisco.com/en/US/products/ps10661/products_implementation_design_guides_list.html

  Unified CM Session Management Edition designs should be reviewed by your Cisco SE in conjunction with the Cisco Unified CM Session Management Team.

Hybrid Session Management Edition and SAF CCD Deployments

Session Management Edition deployments provide internal dial plan aggregation. Cisco Service Advertisement Framework (SAF) Call Control Discovery (CCD) deployments distribute both the internal dial plan and the corresponding external "To PSTN" dial plan to participating SAF CCD Unified Communications systems. Combining Session Management Edition and SAF CCD enables Session Management Edition to act as the central Session Manager for all leaf Unified Communications systems, while also using SAF CCD to distribute both the internal and external "To PSTN" dial plans to all SAF CCD participating Unified CM leaf clusters.

A Session Management Edition and SAF hybrid deployment uses a specific configuration of SAF CCD to allow all calls between leaf clusters to be routed only through the Session Management Edition cluster. The SAF configuration consists of two parts:

- Advertising SAF CCD routes to leaf clusters from/through Session Management Edition
- Advertising SAF CCD routes from leaf clusters to Session Management Edition

Note This discussion assumes that you have already configured your Cisco IOS SAF Forwarders and basic SAF CCD configuration on Unified CM (that is, Advertising Service, Requesting Service, SAF enabled Trunks, and so forth). This design uses a single SAF Autonomous System (AS).

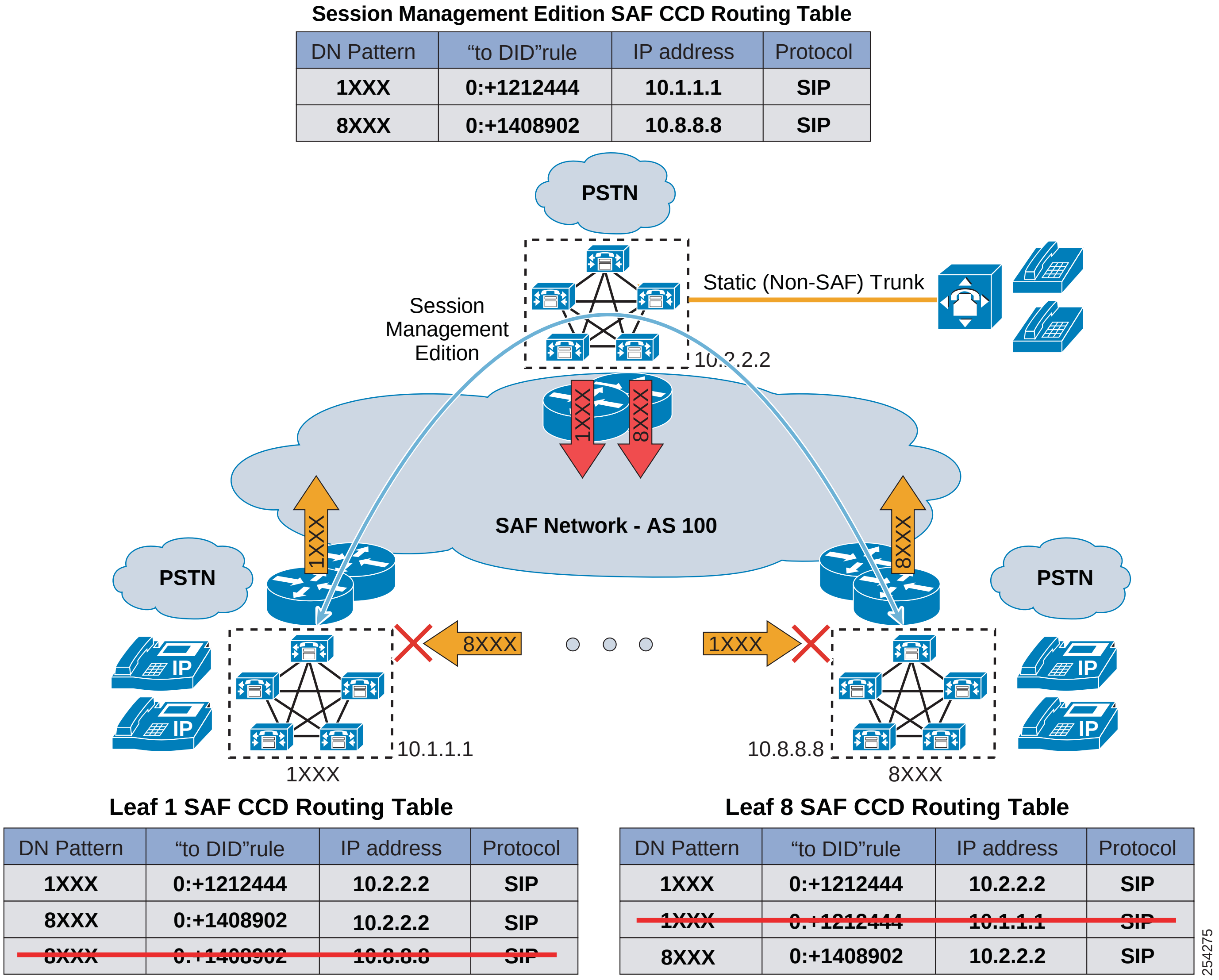
Advertising SAF CCD Routes to Leaf Clusters from/through Session Management Edition

On the Session Management Edition cluster, create the DN patterns, DN Groups, and corresponding "to DID" rules for the internal number ranges and external "To PSTN" numbers hosted by each SAF-enabled leaf cluster. Publish these DN patterns to the SAF AS by associating them with one or more SAF-enabled trunks and advertising services. These DN patterns and corresponding routes to Session Management Edition are learned by all SAF-enabled leaf clusters. While Session Management Edition is reachable through the IP WAN, all intercluster calls are routed through Session Management Edition. When Session Management Edition is unreachable, intercluster calls are routed through the leaf cluster's local PSTN gateway after the called number has been modified using the learned DN pattern's "to DID" rule.

Advertising SAF CCD Routes from Leaf Clusters to Session Management Edition

The purpose of advertising each leaf cluster's hosted DN ranges to the SAF AS is to allow the Session Management Edition cluster to learn about these DN ranges and leaf cluster reachability. These number ranges are also learned by all other leaf clusters. (See Figure 5-7.) To prevent direct leaf-to-leaf routes from being used, in each leaf cluster, block learned routes from all other leaf clusters. Routes can be blocked based on whether they match either the IP address the SAF nodes in each of the leaf clusters or (preferably) the Remote Call Control Entity Name for each leaf cluster. (This is the Unified CM Cluster ID in the Unified CM Enterprise Parameters menu.)

Figure 5-7 Advertising SAF CCD Routes in a Session Management Edition Deployment



Session Management Edition SAF CCD Routing Table

DN Pattern	"to DID"rule	IP address	Protocol
1XXX	0:+1212444	10.1.1.1	SIP
8XXX	0:+1408902	10.8.8.8	SIP

Leaf 1 SAF CCD Routing Table

DN Pattern	"to DID"rule	IP address	Protocol
1XXX	0:+1212444	10.2.2.2	SIP
8XXX	0:+1408902	10.2.2.2	SIP
~~8XXX~~	~~0:+1408902~~	~~10.8.8.8~~	~~SIP~~

Leaf 8 SAF CCD Routing Table

DN Pattern	"to DID"rule	IP address	Protocol
1XXX	0:+1212444	10.2.2.2	SIP
~~1XXX~~	~~0:+1212444~~	~~10.1.1.1~~	~~SIP~~
8XXX	0:+1408902	10.2.2.2	SIP

Operational Considerations for Session Management Edition and SAF CCD Deployments

The following operational considerations apply to deployments of Cisco Unified CM Session Management Edition with Service Advertisement Framework (SAF) Call Control Discovery (CCD).

Leaf Clusters Learning Their Own DN Ranges from Session Management Edition

As can be seen in the SAF CCD routing tables in Figure 5-7, leaf clusters learn about the reachability of their own DN ranges from Session Management Edition. These DN ranges can be blocked in the same way that intercluster DN ranges and routes are blocked. If these Session Management Edition SAF CCD routes are not blocked, they are selected only for intra-cluster calls if the calling search space of the calling device has the SAF CCD learned routes partition ordered above the internal DN's partition. In most cases, the internal DN partition will be ordered above the SAF CCD partition, so that intra-cluster calls are not routed through Session Management Edition.

Routing Calls to the PSTN When IP Routes from Session Management Edition to Leaf Clusters Are Not Available

Two configuration options are available when re-routing calls to the PSTN:

- Re-route calls to the PSTN through a PSTN gateway associated with Session Management Edition

  If the Session Management Edition cluster has PSTN access and you wish to re-route calls that are unreachable through an IP path from Session Management Edition to the destination leaf cluster, make sure each leaf cluster advertises a "to DID" rule for each advertised DN range or group to Session Management Edition. This "to DID" rule is used by Session Management Edition to modify the called number and to route the call through the inbound trunk's Automated Alternate Routing (AAR) calling search space (CSS).

- Re-route calls to the PSTN from the originating leaf cluster

  If the Session Management Edition cluster does not have PSTN access and you wish to re-route calls that are unreachable from Session Management Edition to the destination leaf cluster through the PSTN at the originating leaf cluster, make sure each leaf cluster does not advertises a "to DID" rule for each advertised DN range or group to Session Management Edition. In this case, if a signaling path cannot be established from Session Management Edition to the destination leaf cluster, Session Management Edition signals the call failure to the originating leaf cluster, which in turn uses its "to DID" rule (learned from Session Management Edition) to modify the called number and route the call through the calling device's Automated Alternate Routing (AAR) calling search space (CSS).

Calls to Non-SAF Unified Communications Systems over Static Session Management Edition Trunks

Session Management Edition can use SAF CCD to advertise the DN ranges of non-SAF Unified Communications systems to all SAF-enabled leaf clusters. Calls from leaf clusters to non-SAF Unified Communications systems through the Session Management Edition cluster use SAF trunks to reach Session Management Edition. Session Management Edition then uses a configured route pattern and corresponding static (standard) trunk to reach the non-SAF Unified Communications system.

PSTN Fallback for Calls to Non-SAF Unified Communications Systems

There are two options for PSTN fallback if the non-SAF Unified Communications system is not reachable through a static trunk from Session Management Edition:

- Re-route calls to the PSTN from the originating leaf cluster.

  With this option, a single trunk is configured from Session Management Edition to the destination Unified Communications system. If a signaling path cannot be established from Session Management Edition to the destination Unified Communications system, Session Management

Edition signals the call failure to the originating leaf cluster, which in turn uses its "to DID" rule (learned from Session Management Edition) to modify the called number and route the call through the calling device's Automated Alternate Routing (AAR) calling search space (CSS).

- Re-route calls to the PSTN from Session Management Edition.

  With this option, create two trunks as part of a route list and route group. The first-choice trunk is configured from Session Management Edition to the destination Unified Communications system, while the second-choice trunk is configured from Session Management Edition to its local PSTN gateway. If a signaling path cannot be established from Session Management Edition to the destination Unified Communications system, Session Management Edition chooses the second trunk to the PSTN. The route group that contains the PSTN trunk can be used to modify the internal called number to its PSTN equivalent.

Clustering Over the IP WAN

You may deploy a single Unified CM cluster across multiple sites that are connected by an IP WAN with QoS features enabled. This section provides a brief overview of clustering over the WAN. For further information, refer to the chapter on Call Processing, page 8-1.

Clustering over the WAN can support two types of deployments:

- Local Failover Deployment Model, page 5-37

  Local failover requires that you place the Unified CM subscriber and backup servers at the same site, with no WAN between them. This type of deployment is ideal for two to four sites with Unified CM.

- Remote Failover Deployment Model, page 5-43

  Remote failover allows you to deploy primary and backup call processing servers split across the WAN. Using this type of deployment, you may have multiple sites with Unified CM subscribers being backed up by Unified CM subscribers at another site.

Note Remote failover deployments might require higher bandwidth because a large amount of intra-cluster traffic flows between the subscriber servers.

You can also use a combination of the two deployment models to satisfy specific site requirements. For example, two main sites may each have primary and backup subscribers, with another two sites containing only a primary server each and utilizing either shared backups or dedicated backups at the two main sites.

Some of the key advantages of clustering over the WAN are:

- Single point of administration for users for all sites within the cluster
- Feature transparency
- Shared line appearances
- Extension mobility within the cluster
- Unified dial plan

These features make this solution ideal as a disaster recovery plan for business continuance sites or as a single solution for multiple small or medium sites.

WAN Considerations

For clustering over the WAN to be successful, you must carefully plan, design, and implement various characteristics of the WAN itself. The Intra-Cluster Communication Signaling (ICCS) between Unified CM servers consists of many traffic types. The ICCS traffic types are classified as either priority or best-effort. Priority ICCS traffic is marked with IP Precedence 3 (DSCP 24 or PHB CS3). Best-effort ICCS traffic is marked with IP Precedence 0 (DSCP 0 or PHB BE). The various types of ICCS traffic are described in Intra-Cluster Communications, page 5-34, which also provides further guidelines for provisioning. The following design guidelines apply to the indicated WAN characteristics:

- Delay

  The maximum one-way delay between any two Unified CM servers should not exceed 40 ms, or 80 ms round-trip time. Measuring the delay is covered in Delay Testing, page 5-36. Propagation delay between two sites introduces 6 microseconds per kilometer without any other network delays being considered. This equates to a theoretical maximum distance of approximately 6,000 km for 40 ms delay or approximately 3,720 miles. These distances are provided only as relative guidelines and in reality will be shorter due to other delay incurred within the network.

- Jitter

  Jitter is the varying delay that packets incur through the network due to processing, queue, buffer, congestion, or path variation delay. Jitter for the IP Precedence 3 ICCS traffic must be minimized using Quality of Service (QoS) features.

- Packet loss and errors

  The network should be engineered to provide sufficient prioritized bandwidth for all ICCS traffic, especially the priority ICCS traffic. Standard QoS mechanisms must be implemented to avoid congestion and packet loss. If packets are lost due to line errors or other “real world” conditions, the ICCS packet will be retransmitted because it uses the TCP protocol for reliable transmission. The retransmission might result in a call being delayed during setup, disconnect (teardown), or other supplementary services during the call. Some packet loss conditions could result in a lost call, but this scenario should be no more likely than errors occurring on a T1 or E1, which affect calls via a trunk to the PSTN/ISDN.

- Bandwidth

  Provision the correct amount of bandwidth between each server for the expected call volume, type of devices, and number of devices. This bandwidth is in addition to any other bandwidth for other applications sharing the network, including voice and video traffic between the sites. The bandwidth provisioned must have QoS enabled to provide the prioritization and scheduling for the different classes of traffic. The general rule of thumb for bandwidth is to over-provision and under-subscribe.

- Quality of Service

  The network infrastructure relies on QoS engineering to provide consistent and predictable end-to-end levels of service for traffic. Neither QoS nor bandwidth alone is the solution; rather, QoS-enabled bandwidth must be engineered into the network infrastructure.

Intra-Cluster Communications

In general, intra-cluster communications means all traffic between servers. There is also a real-time protocol called Intra-Cluster Communication Signaling (ICCS), which provides the communications with the Cisco CallManager Service process that is at the heart of the call processing in each server or node within the cluster.

The intra-cluster traffic between the servers consists of the following:

- Database traffic from the IBM Informix Dynamic Server (IDS) database that provides the main configuration information. The IDS traffic may be re-prioritized in line with Cisco QoS recommendations to a higher priority data service (for example, IP Precedence 1 if required by the particular business needs). An example of this is extensive use of Extension Mobility, which relies on IDS database configuration.
- Firewall management traffic, which is used to authenticate the subscribers to the publisher to access the publisher's database. The management traffic flows between all servers in a cluster. The management traffic may be prioritized in line with Cisco QoS recommendations to a higher priority data service (for example, IP Precedence 1 if required by the particular business needs).
- ICCS real-time traffic, which consists of signaling, call admission control, and other information regarding calls as they are initiated and completed. ICCS uses a Transmission Control Protocol (TCP) connection between all servers that have the Cisco CallManager Service enabled. The connections are a full mesh between these servers. This traffic is priority ICCS traffic and is marked dependant on release and service parameter configuration.
- CTI Manager real-time traffic is used for CTI devices involved in calls or for controlling or monitoring other third-party devices on the Unified CM servers. This traffic is marked as priority ICCS traffic and exists between the Unified CM server with the CTI Manager and the Unified CM server with the CTI device.

Note For detailed information on various types of traffic between Unified CM servers, refer to the TCP and UDP port usage documents at http://www.cisco.com/en/US/products/sw/voicesw/ps556/prod_maintenance_guides_list.html.

Unified CM Publisher

The publisher server replicates a partial read-only copy of the master database to all other servers in the cluster. Most of the database modifications are done on the publisher. If changes such as administration updates are made in the publisher’s master database during a period when another server in the cluster is unreachable, the publisher will replicate the updated database when communications are re-established. Database modifications for user-facing call processing features are made on the subscriber servers to which the IP phones are registered. These features include:

- Call Forward All (CFA)
- Message Waiting Indication (MWI)
- Privacy Enable/Disable
- Do Not Disturb (DND) Enable/Disable
- Extension Mobility Login (EM)
- Monitor (for future use; currently no updates at the user level)
- Hunt Group Logout
- Device Mobility
- CTI Certificate Authority Proxy Function (CAPF) status for end users and application users
- Credential hacking and authentication

Each subscriber replicates these changes to every other server in the cluster. Any other configuration changes cannot be made on the database during the period when the publisher is unreachable or offline. Most normal operations of the cluster, including the following, will *not* be affected during the period of publisher failure:

- Call processing
- Failover
- Registration of previously configured devices

Other services or applications might also be affected, and their ability to function without the publisher should be verified when deployed.

Call Detail Records (CDR) and Call Management Records (CMR)

Call detail records and call management records, when enabled, are collected by each subscriber and uploaded to the publisher periodically. During a period that the publisher is unreachable, the CDRs and CMRs are stored on the subscriber's local hard disk. When connectivity is re-established to the publisher, all outstanding CDRs are uploaded to the publisher, which stores the records in the CDR Analysis and Reporting (CAR) database.

Delay Testing

The maximum round-trip time (RTT) between any two servers must not exceed 80 ms. This time limit must include all delays in the transmission path between the two servers. Verifying the round trip delay using the **ping** utility on the Unified CM server will not provide an accurate result. The ping is sent as a best-effort tagged packet and is not transported using the same QoS-enabled path as the ICCS traffic. Therefore, Cisco recommends that you verify the delay by using the closest network device to the Unified CM servers, ideally the access switch to which the server is attached. Cisco IOS provides a extended ping capable to set the Layer 3 type of service (ToS) bits to make sure the ping packet is sent on the same QoS-enabled path that the ICCS traffic will traverse. The time recorded by the extended ping is the round-trip time (RTT), or the time it takes to traverse the communications path and return.

The following example shows a Cisco IOS extended ping with the IP Precedence bits set to 3 (ToS byte value set to 96):

```
Access_SW#ping
Protocol [ip]:
Target IP address: 10.10.10.10
Repeat count [5]:
Datagram size [100]:
Timeout in seconds [2]:
Extended commands [n]: y
Source address or interface:
Type of service [0]: 96
Set DF bit in IP header? [no]:
Validate reply data? [no]:
Data pattern [0xABCD]:
Loose, Strict, Record, Timestamp, Verbose[none]:
Sweep range of sizes [n]:
Type escape sequence to abort.
Sending 5, 100-byte ICMP Echos to 10.10.10.10, timeout is 2 seconds:
!!!!!
Success rate is 100 percent (5/5), round-trip min/avg/max = 1/2/4 ms
```

Error Rate

The expected error rate should be zero. Any errors, dropped packets, or other impairments to the IP network can have an impact to the call processing performance of the cluster. This may be noticeable by delay in dial tone, slow key or display response on the IP phone, or delay from off-hook to connection of the voice path. Although Unified CM will tolerate random errors, they should be avoided to avoid impairing the performance of the cluster.

Troubleshooting

If the Unified CM subscribers in a cluster are experiencing impairment of the ICCS communication due to higher than expected delay, errors, or dropped packets, some of the following symptoms might occur:

- IP phones, gateways, or other devices on a remote Unified CM server within the cluster might temporarily be unreachable.
- Calls might be disconnected or might fail during call setup.
- Users might experience longer than expected delays before hearing dial tone.
- Busy hour call completions (BHCC) might be low.
- The ICCS (SDL session) might be reset or disconnected.
- The time taken to upgrade a subscriber and synchronize its database with the publisher will increase.

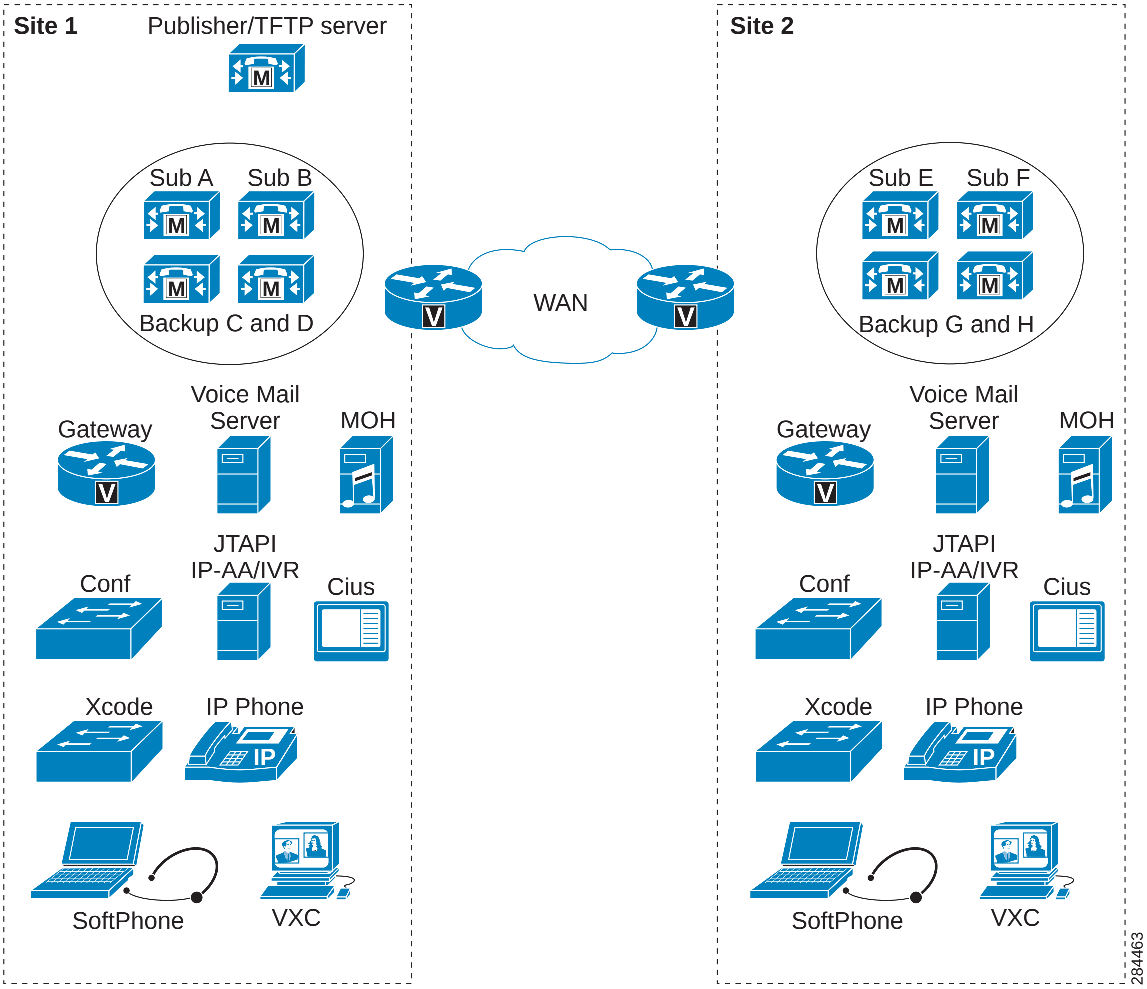
In summary, perform the following tasks to troubleshoot ICCS communication problems:

- Verify the delay between the servers.
- Check all links for errors or dropped packets.
- Verify that QoS is correctly configured.
- Verify that sufficient bandwidth is provisioned for the queues and across the WAN to support all the traffic.

Local Failover Deployment Model

The local failover deployment model provides the most resilience for clustering over the WAN. Each of the sites in this model contains at least one primary Unified CM subscriber and one backup subscriber. This configuration can support up to four sites. The maximum number of phones and other devices will be dependant on the quantity and type of servers deployed. The maximum total number of IP phones for all sites is 40,000. (See Figure 5-8.)

Figure 5-8 Example of Local Failover Model



Observe the following guidelines when implementing the local failover model:

- Configure each site to contain at least one primary Unified CM subscriber and one backup subscriber.
- Configure Unified CM *groups* and *device pools* to allow devices within the site to register with only the servers at that site under all conditions.
- Cisco highly recommends that you replicate key services (TFTP, DNS, DHCP, LDAP, and IP Phone Services), all media resources (conference bridges and music on hold), and gateways at each site to provide the highest level of resiliency. You could also extend this practice to include a voicemail system at each site.
- Under a WAN failure condition, sites without access to the publisher database will lose some functionality. For example, system administration at the remote site will not be able to add, modify, or delete any part of the configuration. However, users can continue to access the user-facing features listed in the section on Unified CM Publisher, page 5-35.
- Under WAN failure conditions, calls made to phone numbers that are not currently communicating with the subscriber placing the call, will result in either a fast-busy tone or a call forward (possibly to voicemail or to a destination configured under Call Forward Unregistered).

- The maximum allowed round-trip time (RTT) between any two servers in the Unified CM cluster is 80 ms.

  Note At a higher round-trip delay time and higher busy hour call attempts (BHCA), voice cut-through delay might be higher, causing initial voice clipping when a voice call is established.

- A minimum of 1.544 Mbps (T1) bandwidth is required for Intra-Cluster Communication Signaling (ICCS) for 10,000 busy hour call attempts (BHCA) between sites that are clustered over the WAN. This is a minimum bandwidth requirement for call control traffic, and it applies to deployments where directory numbers are not shared between sites that are clustered over the WAN. The following equation may be used as a guideline to calculate the bandwidth for more than 10,000 BHCA between non-shared directory numbers at a specific delay:

  Total Bandwidth (Mbps) = (Total BHCA/10,000) ∗ (1 + 0.006 ∗ Delay), where
  Delay = RTT delay in ms

  This call control traffic is classified as priority traffic. Priority ICCS traffic is marked with IP Precedence 3 (DSCP 24 or PHB CS3).

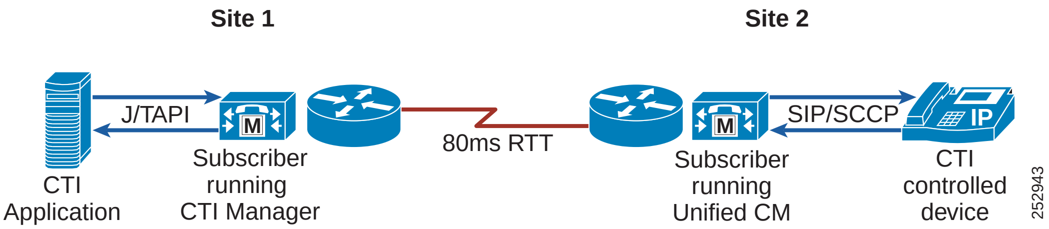
- In addition to the bandwidth required for Intra-Cluster Communication Signaling (ICCS) traffic, a minimum of 1.544 Mbps (T1) bandwidth is required for database and other inter-server traffic for every subscriber server remote to the publisher.

- For customers who also want to deploy CTI Manager over the WAN (see Figure 5-9), the following formula can be used to calculate the bandwidth (Mbps) for the CTI Intra-Cluster Communication Signaling (ICCS) traffic between the Unified CM subscriber running the CTI Manager service and the Unified CM subscriber to which the CTI controlled endpoint is registered:

  With Unified CM 8.6(1) and earlier releases, CTI ICCS bandwidth (Mbps)
  = (Total BHCA/10,000) ∗ 1.25

  With Unified CM 8.6(2) and later releases, CTI ICCS bandwidth (Mbps)
  = (Total BHCA/10,000) ∗ 0.53

Figure 5-9 CTI Over the WAN

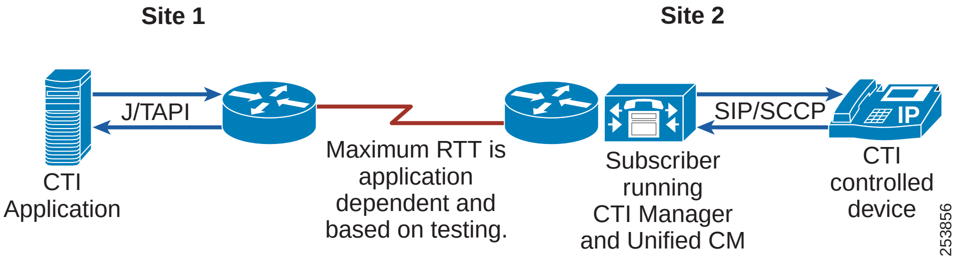


- For deployments where the J/TAPI application is remote from the Unified CM subscriber (see Figure 5-10), the following formula can be used to calculate the Quick Buffer Encoding (QBE) J/TAPI bandwidth for a typical J/TAPI application with Unified CM 8.6(2) and later releases:

  J/TAPI bandwidth (Mbps) = (Total BHCA/10,000) ∗ 0.28

  The bandwidth may vary depending on the J/TAPI application. Check with the application developer or provider to validate the bandwidth requirement.

Figure 5-10 J/TAPI Over the WAN


Site 1
Site 2
J/TAPI
SIP/SCCP
IP
CTI Application
Maximum RTT is application dependent and based on testing.
Subscriber running CTI Manager and Unified CM
CTI controlled device
253856

Example 5-1 Bandwidth Calculation for Two Sites

Consider two sites, Site 1 and Site 2, with Unified CM clustered over the WAN across these two sites that are 80 ms round-trip time apart. Site 1 has one publisher, one combined TFTP and music on hold (MoH) server, and two Unified CM subscriber servers. Site 2 has one TFTP/MoH server and two Unified CM subscriber servers. Site 1 has 5000 phones, each having one DN; and Site 2 has 5000 phones, each having one DN. During the busy hour, 2500 phones in Site 1 call 2500 phones in Site 2, each at 3 BHCA. During that same busy hour, 2500 phones in Site 2 also call 2500 phones in Site 1, each at 3 BHCA. In this case:

Total BHCA during the busy hour = 2500∗3 + 2500∗3 = 15,000

Total bandwidth required between the sites = Total ICCS bandwidth + Total database bandwidth

Because total BHCA is 15,000 (greater than 10,000), we can use the formula to calculate:
Total ICCS bandwidth = (15,000/10,000) ∗ (1 + 0.006∗80) = 2.22 Mbps

Total database bandwidth = (Number of servers remote to the publisher) ∗ 1.544 = 3 ∗ 1.544 = 4.632 Mbps

Total bandwidth required between the sites = 2.22 Mbps + 4.632 Mbps = 6.852 Mbps (Approximately 7 Mbps)

- When directory numbers are shared between sites that are clustered over the WAN, additional bandwidth must be reserved. This overhead or additional bandwidth (in addition to the minimum 1.544 Mbps bandwidth) for 10,000 BHCA between shared DNs can be calculated using the following equation:

  Overhead = (0.012 ∗ Delay ∗ Shared-line) + (0.65 ∗ Shared-line), where:

  Delay = RTT delay over the IP WAN, in ms

  Shared-line = Average number of additional phones on which a directory number is shared across the WAN.

  The following equation may be used as a guideline to calculate the bandwidth for more than 10,000 BHCA between shared directory numbers at a specific delay:

  Total bandwidth (Mbps) = (Total BHCA/10,000) ∗ (1 + 0.006 ∗ Delay + 0.012 ∗ Delay ∗ Shared-line + 0.65 ∗ Shared-line), where:

  Delay = RTT delay in ms

  Shared-line = Average number of additional phones on which a directory number is shared across the WAN.

Example 5-2 Bandwidth Calculation for Two Sites with Shared Directory Numbers

Consider two sites, Site 1 and Site 2, with Unified CM clustered over the WAN across these two sites that are 80 ms round-trip time apart. Site 1 has one publisher, one combined TFTP and music on hold (MoH) server, and two Unified CM subscriber servers. Site 2 has one TFTP/MoH server and two Unified CM subscriber servers. Site 1 has 5000 phones, each having one DN; and Site 2 has 5000 phones, each sharing a DN with the 5000 phones in Site 1. Thus, each DN is shared across the WAN with an average of one additional phone. During the busy hour, 2500 phones in Site 1 call 2500 phones in Site 2, each at 3 BHCA. This also causes the phones in Site 1 to ring. During that same busy hour, 2500 phones in Site 2 call 2500 phones in Site 1, each at 3 BHCA. This also causes the phones in Site 2 to ring. In this case:

Total BHCA during the busy hour = 2500 ∗ 3 + 2500 ∗ 3 = 15,000

Total bandwidth required between the sites = Total ICCS bandwidth + Total database bandwidth

Because total BHCA is 15,000 (greater than 10,000), we can use the formula to calculate:
Total ICCS bandwidth = (15,000/10,000) ∗ (1 + 0.006∗80 + 0.012∗80∗1 + 0.65∗1) = 4.635 Mbps

Total database bandwidth = (Number of servers remote to the publisher) ∗ 1.544 = 3 ∗ 1.544
= 4.632 Mbps

Total bandwidth required between the sites = 4.635 Mbps + 4.632 Mbps = 9.267 Mbps
(Approximately 10 Mbps)

Note The bandwidth requirements stated above are strictly for ICCS, database, and other inter-server traffic. If calls are going over the IP WAN, additional bandwidth must be provisioned for voice or media traffic, depending on the voice codec used for the calls.

- Subscriber servers in the cluster read their local database. Database modifications can occur in both the local database as well as the publisher database, depending on the type of changes. Informix Dynamic Server (IDS) database replication is used to synchronize the databases on the various servers in the cluster. Therefore, when recovering from failure conditions such as the loss of WAN connectivity for an extended period of time, the Unified CM databases must be synchronized with any changes that might have been made during the outage. This process happens automatically when database connectivity is restored to the publisher and other servers in the cluster. This process can take longer over low bandwidth and/or higher delay links. In rare scenarios, manual reset or repair of the database replication between servers in the cluster might be required. This is performed by using the commands such as **utils dbreplication repair all** and/or **utils dbreplication reset all** at the command line interface (CLI). Repair or reset of database replication using the CLI on remote subscribers over the WAN causes all Unified CM databases in the cluster to be re-synchronized, in which case additional bandwidth above 1.544 Mbps might be required. With longer delays and lower bandwidth between the publisher and subscriber nodes, it can take longer for database replication repair or reset to complete.

  Note Repairing or resetting of database replication on multiple subscribers at the same remote location can result in increased time for database replication to complete. Cisco recommends repairing or resetting of database replication on these remote subscribers one at a time. Repairing or resetting of database replication on subscribers at different remote locations may be performed simultaneously.

- If remote branches using centralized call processing are connected to the main sites via clustering over the WAN, pay careful attention to the configuration of call admission control to avoid oversubscribing the links used for clustering over the WAN.
  – If the bandwidth is not limited on the links used for clustering over the WAN (that is, if the interfaces to the links are OC-3s or STM-1s and there is no requirement for call admission control), then the remote sites may be connected to any of the main sites because all the main sites should be configured as location Hub_None. This configuration still maintains hub-and-spoke topology for purposes of call admission control.
  – If you are using the Multiprotocol Label Switching (MPLS) Virtual Private Network (VPN) feature, all sites in Unified CM locations and the remote sites may register with any of the main sites.
  – If bandwidth is limited between the main sites, call admission control must be used between sites, and all remote sites must register with the main site that is configured as location Hub_None. This main site is considered the hub site, and all other remote sites and clustering-over-the-WAN sites are spokes sites.
- During a software upgrade, all servers in the cluster should be upgraded during the same maintenance period, using the standard upgrade procedures outlined in the software release notes. The software upgrade time will increase for higher round-trip delay time over the IP WAN. Lower bandwidths such as 1.544 Mbps (T1 link) can also cause the software upgrade process to take longer to complete, in which case additional bandwidth above 1.544 Mbps might be required if a faster upgrade process is desired.

Unified CM Provisioning for Local Failover

Provisioning of the Unified CM cluster for the local failover model should follow the design guidelines for capacities outlined in the chapter on Call Processing, page 8-1. If voice or video calls are allowed across the WAN between the sites, then you must configure Unified CM *locations* in addition to the default location for the other sites, to provide call admission control between the sites. If the bandwidth is over-provisioned for the number of devices, it is still best practice to configure call admission control based on locations. If the locations-based call admission control rejects a call, automatic failover to the PSTN can be provided by the automated alternate routing (AAR) feature.

To improve redundancy and upgrade times, Cisco recommends that you enable the Cisco Trivial File Transfer Protocol (TFTP) service on two Unified CM servers. More than two TFTP servers can be deployed in a cluster, however this configuration can result in an extended period for rebuilding all the TFTP files on all TFTP servers.

You can run the TFTP service on either a publisher or a subscriber server, depending on the site and the available capacity of the server. The TFTP server option must be correctly set in the DHCP servers at each site. If DHCP is not in use or if the TFTP server is manually configured, you should configure the correct address for the site.

Other services, which may affect normal operation of Unified CM during WAN outages, should also be replicated at all sites to ensure uninterrupted service. These services include DHCP servers, DNS servers, corporate directories, and IP phone services. On each DHCP server, set the DNS server address correctly for each location.

IP phones may have shared line appearances between the sites. During a WAN outage, call control for each line appearance is segmented, but call control returns to a single Unified CM server once the WAN is restored. During the WAN restoration period, there is additional traffic between the two sites. If this situation occurs during a period of high call volume, the shared lines might not operate as expected during that period. This situation should not last more than a few minutes, but if it is a concern, you can provision additional prioritized bandwidth to minimize the effects.

Gateways for Local Failover

Normally, gateways should be provided at all sites for access to the PSTN. The device pools should be configured to register the gateways with the Unified CM servers at the same site. Call routing (route patterns, route lists, and route groups) should also be configured to select the local gateways at the site as the first choice for PSTN access and the other site gateways as a second choice for overflow. Take special care to ensure emergency service access at each site.

You can centralize access to the PSTN gateways if access is not required during a WAN failure and if sufficient additional bandwidth is configured for the number of calls across the WAN. For E911 requirements, additional gateways might be needed at each site.

Voicemail for Local Failover

Cisco Unity Connection or other voicemail systems can be deployed at all sites and integrated into the Unified CM cluster. This configuration provides voicemail access even during a WAN failure and without using the PSTN. Using Voice Mail Profiles, you can allocate the correct voicemail system for the site to the IP phones in the same location. You can configure a maximum of four voicemail systems per cluster that use the SMDI protocol, that are attached directly to the COM port on a subscriber, and that use the Cisco Messaging Interface (CMI).

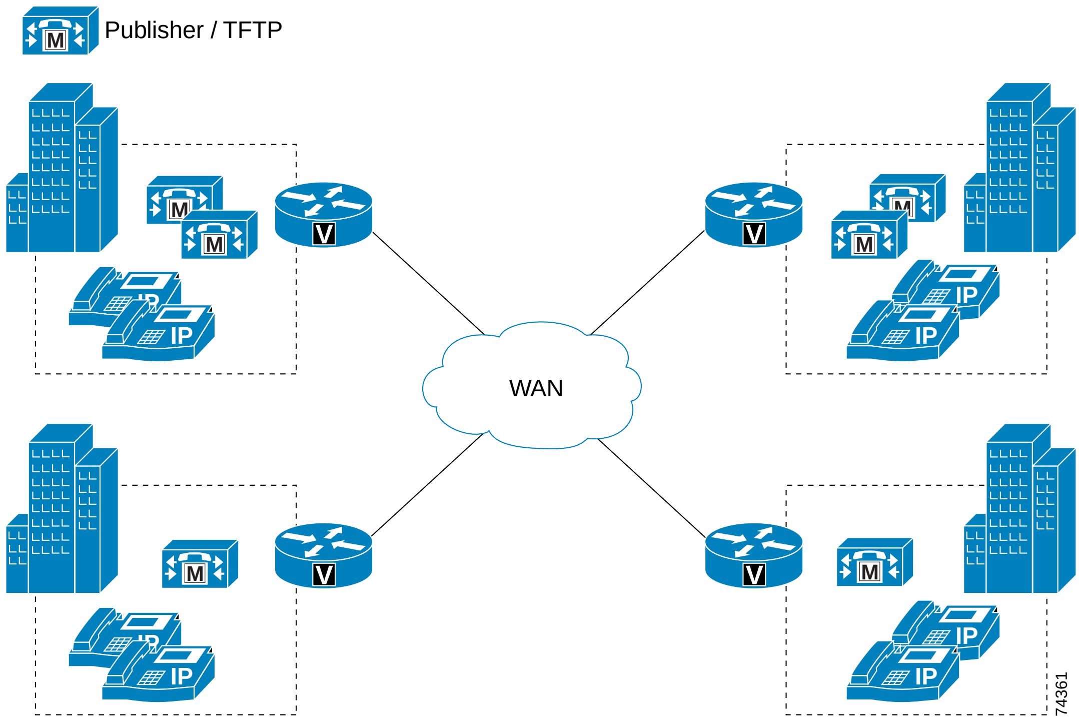
Music on Hold and Media Resources for Local Failover

Music on hold (MoH) servers and other media resources such as conference bridges should be provisioned at each site, with sufficient capacity for the type and number of users. Through the use of media resource groups (MRGs) and media resource group lists (MRGLs), media resources are provided by the on-site resource and are available during a WAN failure.

Remote Failover Deployment Model

The remote failover deployment model provides flexibility for the placement of backup servers. Each of the sites contains at least one primary Unified CM subscriber and may or may not have a backup subscriber. This model allows for multiple sites, with IP phones and other devices normally registered to a local subscriber when using 1:1 redundancy and the 50/50 load balancing option described in the chapter on Call Processing, page 8-1. Backup subscribers are located across the WAN at one or more of the other sites. (See Figure 5-11.)

Figure 5-11 Remote Failover Model with Four Sites



When implementing the remote failover model, observe all guidelines for the local failover model (see Local Failover Deployment Model, page 5-37), with the following modifications:

- Configure each site to contain at least one primary Unified CM subscriber and an optional backup subscriber as desired. If a backup subscriber over the IP WAN is not desired, a Survivable Remote Site Telephony (SRST) router may be used as a backup call processing agent.
- You may configure Unified CM *groups* and *device pools* to allow devices to register with servers over the WAN as a second or third choice.
- Signaling or call control traffic requires bandwidth when devices are registered across the WAN with a remote Unified CM server in the same cluster. This bandwidth might be more than the ICCS traffic and should be calculated using the bandwidth provisioning calculations for signaling, as described in Bandwidth Provisioning, page 3-45.

Note You can also combine the features of these two types of deployments for disaster recovery purposes. For example, Unified CM groups permit configuring up to three servers (primary, secondary and tertiary). Therefore, you can configure the Unified CM groups to have primary and secondary servers that are located at the same site and the tertiary server at a remote site over the WAN.

Cisco Business Edition 6000 Clustering over the WAN

Cisco Business Edition 6000 may be deployed using the clustering-over-the-WAN call processing local failover model. In this type of deployment, two Business Edition 6000 server nodes are deployed at each of two sites to provide geographic redundancy for the Unified CM call processing application. The two Business Edition 6000 server nodes may both be UCS C200 Rack-Mount Servers, or alternatively one of the servers may be a regular Cisco Media Convergence Server (MCS).

Business Edition 6000 call processing clustering over the WAN deployments must observe the same guidelines and requirements as with regular Unified CM clustering over the WAN and as described earlier. Observe the following guidelines when clustering Business Edition 6000 over the WAN with the local failover model:

- Configure Unified CM groups and device pools to allow devices within each site to register with only the servers at that site under all conditions.
- Cisco highly recommends that you replicate key services (TFTP, DNS, DHCP, LDAP, and IP Phone Services), all media resources (conference bridges and music on hold), and gateways at each site to provide the highest level of resiliency.
- Under a WAN failure condition, the site without access to the publisher database will lose some functionality. For example, system administration at the secondary site will not be able to add, modify, or delete any part of the configuration. However, users can continue to access the user-facing features listed in the section on Unified CM Publisher, page 5-35.
- Under WAN failure conditions, calls made to phone numbers that are not currently communicating with the subscriber placing the call, will result in either a fast-busy tone or a call forward (possibly to voicemail or to a destination configured under Call Forward Unregistered).
- The maximum allowed round-trip time (RTT) between the two Business Edition 6000 server nodes at the two sites is 80 ms.
- 1.544 Mbps (T1) bandwidth is required for Intra-Cluster Communication Signaling (ICCS) busy hour call attempts (BHCA) between the two sites that are clustered over the WAN. This is a bandwidth requirement for call control traffic, and it applies to deployments where directory numbers are not shared between sites that are clustered over the WAN.   This call control traffic is classified as priority traffic. Priority ICCS traffic is marked with IP Precedence 3 (DSCP 24 or PHB CS3).
- In addition to the bandwidth required for Intra-Cluster Communication Signaling (ICCS) traffic, an additional 1.544 Mbps (T1) of bandwidth is required for database and other traffic between the two Business Edition 6000 server nodes.

More than two UCS C200 Rack-Mount Servers may be clustered for a Business Edition 6000 deployment to provide additional geographic redundancy beyond two sites with the remote failover model for clustering over the WAN (see Remote Failover Deployment Model, page 5-43). However, the total number of users across the Business Edition 6000 cluster may not exceed 1,000 and the total number of configured devices across the cluster may not exceed 1,200. A deployment of UCS C200 Rack-Mount Servers in a cluster exceeding 1,000 users and 1,200 configured devices is considered a regular Unified CM cluster, and as such it is bound by all requirements and design guidance for regular Unified CM clusters.

In deployments of Business Edition 6000 with more than two UCS C200 Rack-Mount Servers in the remote failover model for clustering over the WAN, the following additional guidelines must be observed:

- 1.544 Mbps (T1) bandwidth is required for Intra-Cluster Communication Signaling (ICCS) busy hour call attempts (BHCA) between each site that is clustered over the WAN. This is a bandwidth requirement for call control traffic.
- In addition to the bandwidth required for Intra-Cluster Communication Signaling (ICCS) traffic, an additional 1.544 Mbps (T1) of bandwidth is required for database and other inter-server traffic between any server nodes remote from the Business Edition 6000 publisher node.

Clustering over the WAN for Business Edition 6000 Co-Resident Applications

In addition to clustering call processing services over the WAN, Cisco Business Edition 6000 co-resident applications (Cisco Unity Connection, Cisco IM and Presence, and Cisco Unified Contact Center Express) may also be clustered over the WAN provided that these deployments adhere to the same guidelines and restrictions as apply to these applications running on separate systems.

Each co-resident application must adhere strictly to its maximum delay and bandwidth requirements. Furthermore, it is important to understand that, while maximum delay budget will apply to all applications, the WAN bandwidth required for each clustered application (including call processing) must be added together to derive the appropriate WAN bandwidth requirement.

Observe the following general guidelines when clustering co-resident Cisco Business Edition 6000 applications and services:

- Round-trip delay across the WAN must not exceed 80 milliseconds because this is the maximum round-trip delay supported across all applications, including call processing.
- The bandwidth requirement on the WAN is based on the total of each application's bandwidth requirement for clustering over the WAN. For example, if all applications (Cisco Unified CM, IM and Presence, Unity Connection, and Unified Contact Center Express) are clustered over the WAN, the total bandwidth required on the WAN would be calculated as follows:

  (Total required WAN bandwidth) = (Unified CM required bandwidth) + (IM and Presence required bandwidth) + (Unity Connection required bandwidth) + (Unified Contact Center Express required bandwidth)

For information on clustering delay and bandwidth requirement for each co-resident application, refer to the following information:

- For Cisco IM and Presence, see Clustering Over the WAN, page 23-21.
- For Cisco Unity Connection, see Cisco Unity Connection Redundancy and Clustering Over the WAN, page 21-17.
- For Cisco Unified Contact Center Express, see Clustering Over the IP WAN, page 26-9.

Deploying Unified Communications on Virtualized Servers

Cisco Unified Communications applications can run in a virtualized environment as virtual machines using the VMware ESXi hypervisor. Two hardware options are available:

- Tested Reference Configurations (TRC), which are selected hardware configurations based on Cisco Unified Computing System (UCS) platforms
- Specification-based hardware that provides more hardware flexibility and that, for example, adds support for other Cisco UCS, Hewlett-Packard, and IBM platforms listed in the VMware Hardware Compatibility List (available at (http://www.vmware.com/resources/compatibility/search.php)

This section presents a short introduction of the Cisco Unified Computing System (UCS) architecture, Hypervisor Technology for Application Virtualization, and Storage Area Networking (SAN) concepts, with a simple overview of where each product fits in a Cisco Virtualized Unified Communications solution for enterprises. It also includes design considerations for deploying Unified Communications applications over virtualized servers.

This description is not meant to replace or supersede product-specific detailed design guidelines available at the following locations:

- http://www.cisco.com/en/US/products/ps10265/index.html
- http://www.cisco.com/go/uc-virtualized

For sizing aspects of Unified Communications systems on virtualized servers, use the Cisco Unified Communications Sizing Tool, available to Cisco partners and employees (with valid login authentication) at

http://tools.cisco.com/cucst

Cisco Unified Computing System

Unified Computing is an architecture that integrates computing resources (CPU, memory, and I/O), IP networking, network-based storage, and virtualization, into a single highly available system. This level of integration provides economies of power and cooling, simplified server connectivity into the network, dynamic application instance repositioning between physical hosts, and pooled disk storage capacity.

The Cisco Unified Computing System is built from many components. But from a server standpoint, the UCS architecture is divided into the following two categories:

- Cisco UCS B-Series Blade Servers, page 5-47
- Cisco UCS C-Series Rack-Mount Servers, page 5-50

For more details on the Cisco Unified Computing System architecture, refer to the documentation available at

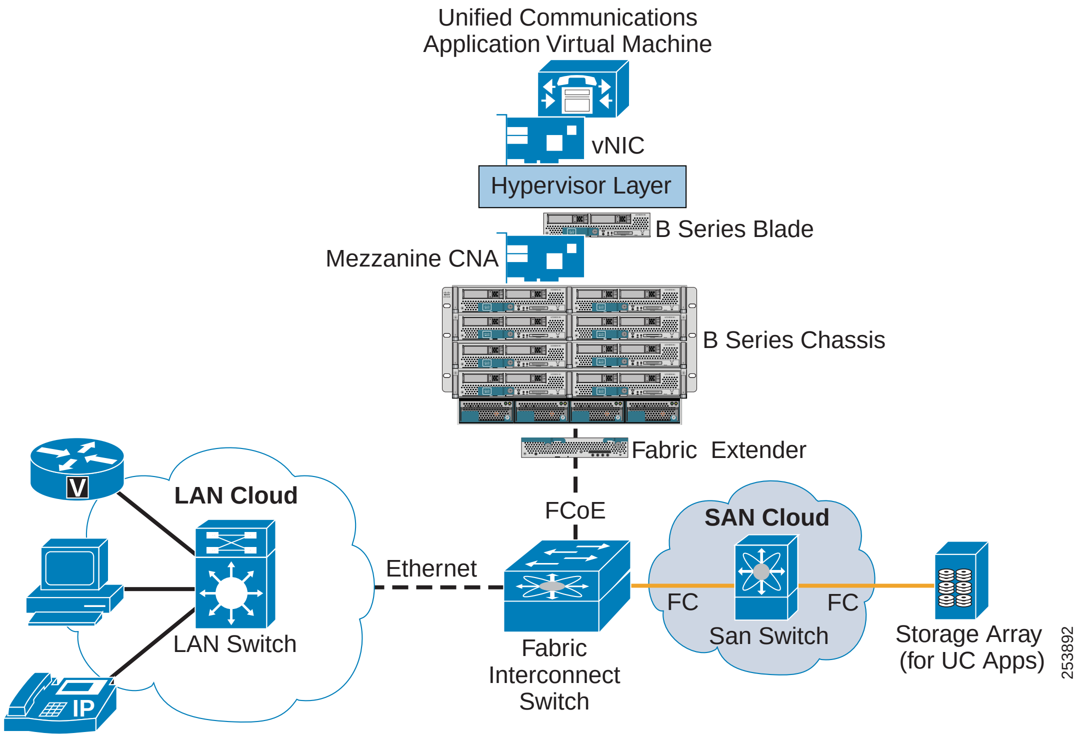
http://www.cisco.com/en/US/netsol/ns944/index.html

Cisco UCS B-Series Blade Servers

The Cisco Unified Computing System (UCS) features blade servers based on x86 architecture. Blade servers provide computing resources (memory, CPU, and I/O) to operating systems and applications. Blade servers have access to the unified fabric through mezzanine form factor Converged Network Adapters (CNA).

The architecture uses a unified fabric that provides transport for LAN, storage, and high-performance computing traffic over a single infrastructure with the help of technologies such as Fibre Channel over Ethernet (FCoE). (See Figure 5-12.) Cisco's unified fabric technology is built on a 10-Gbps Ethernet foundation that eliminates the need for multiple sets of adapters, cables, and switches for LANs, SANs, and high-performance computing networks.

Figure 5-12 Basic Architecture of Unified Communications on Cisco UCS B-Series Blade Servers



This section briefly describes the primary UCS components and how they function in a Unified Communications solution. For details about the Cisco UCS B-Series Blade Servers, refer to the model comparison at

http://www.cisco.com/en/US/products/ps10280/prod_models_comparison.html

Cisco UCS 5100 Series Blade Server Chassis

The Cisco UCS 5100 Series Blade Server chassis not only hosts the B-Series blade servers but also provides connectivity to the uplink Fabric Interconnect Switch by means of Cisco UCS Fabric Extenders.

Cisco UCS 2100 and 2200 Series Fabric Extenders

Cisco UCS 2100 and 2200 Series Fabric Extenders are inserted into the B-Series chassis, and they connect the Cisco UCS 5100 Series Blade Server Chassis to the Cisco UCS Fabric Interconnect Switch. The fabric extender can pass traffic between the blade server's FCoE-capable CNA to the fabric interconnect switch using Fibre Channel over Ethernet (FCoE) protocol.

Cisco UCS 6100 and 6200 Series Fabric Interconnect Switch

A Cisco UCS 6100 and 6200 Series Fabric Interconnect Switch is 10 Gigabit FCoE-capable switch. The B-Series Chassis (and the blade servers) connect to the fabric interconnect, and it connects to the LAN or SAN switching elements in the data center.

Cisco UCS Manager

Management is integrated into all the components of the system, enabling the entire UCS system to be managed as a single entity through the Cisco UCS Manager. Cisco UCS Manager provides an intuitive user interface to manage all system configuration operations.

Hypervisor

A hypervisor is a thin software system that runs directly on the server hardware to control the hardware, and it allows multiple operating systems (guests) to run on a server (host computer) concurrently. A guest operating system (such as that of Cisco Unified CM) thus runs on another level above the hypervisor. Hypervisors are one of the foundation elements in the cloud computing and virtualization technologies, and they consolidate applications onto fewer servers.

Storage Area Networking

Storage area networking (SAN) enables attachment of remote storage devices or storage arrays to the servers so that storage appears to the operating system to be attached locally to the server. SAN storage can be shared between multiple servers.

Design Considerations for Running Virtual Unified Communications Applications on B-Series Blade Servers

This section highlights some design rules and considerations that must be followed for running Unified Communications services on virtualized servers. Many Cisco Unified Communications applications support virtualization on a B-Series Blade server, such as:

- Cisco Unified Communications Manager (Unified CM)
- Cisco Unified CM Session Manager Edition
- Cisco Unity Connection
- Cisco IM and Presence
- Cisco Unified Contact Center Express
- Cisco Unified Contact Center Enterprise

For a full list of supported Cisco Unified Communications applications, refer to the documentation available at

http://www.cisco.com/go/uc-virtualized

Blade Server

The Cisco B-Series Blade Servers support multiple CPU sockets, and each CPU socket can host multiple multi-core processors. For example, one B200 blade has two CPU sockets that can host up to two multi-core processors. This provides the ability to run multiple Unified Communications applications on a single blade server.

Cisco Unified Communications applications should be run on dedicated blades that are not running any non-Unified Communications applications. Each Unified Communications application should be allotted dedicated processing and memory resources, so that the resources are not oversubscribed.

Hypervisor

The VMware ESXi Hypervisor is required to run virtual Unified Communications applications. The local hard drives attached to the Blade Server cannot be used to store virtual machines; they can be used only to install the ESXi hypervisor software. Unified Communications applications must follow the respective guidelines for their virtual machine template and configuration.

VMware vCenter is not mandatory when using a Tested Reference Configuration, but it is strongly recommended to manage multiple ESXi hosts for a large deployment.

For specific configuration and sizing requirements for virtual machines, refer to the respective product documentation available at

http://www.cisco.com/go/uc-virtualized

SAN and Storage Arrays

Tested Reference Configurations based on the Cisco UCS B-Series platform require the virtual machines to run from a Fibre Channel SAN storage array. The SAN storage array must satisfy the requirements of the VMware hardware compatibility list. Other storage options such as iSCSI, FCoE SAN, and NFS NAS are supported with the specification-based hardware support. For more details, refer to the documentation available at

http://www.cisco.com/go/uc-virtualized

Cisco UCS C-Series Rack-Mount Servers

Beside the B-Series Blade Servers, the Cisco Unified Computing System (UCS) also features general purpose rack-mount servers based on x86 architecture. The C-Series Rack-Mount Servers provide computing resources (memory, CPU, and I/O) and optional local storage to operating systems and applications. For more information on C-Series servers, refer to the documentation at

http://www.cisco.com/en/US/products/ps10493/index.html

Design Considerations for Running Virtual Unified Communications Applications on C-Series Rack-Mount Servers

Tested Reference Configurations are also available with Cisco UCS C-Series Rack Mount Servers such as the Cisco UCS C200, C210 and C260.

Many Cisco Unified Communications applications support virtualization on a C-Series Rack Mount Server, such as:

- Cisco Unified Communications Manager (Unified CM)
- Cisco Unified CM Session Manager Edition
- Cisco Unity Connection
- Cisco IM and Presence
- Cisco Unified Contact Center Express
- Cisco Unified Contact Center Enterprise

For a full list of supported Cisco Unified Communications applications, refer to the documentation available at

http://www.cisco.com/go/uc-virtualized

Unlike with the UCS B-Series, the Tested Reference Configurations based on the high-end UCS C-Series Rack Mount Servers (for example, C210 and C260) support storage for virtual machines either locally on the directly attached storage drives or on an FC SAN storage array. Multiple Unified Communications applications can reside on the same C-Series server. Low-end UCS C-Series Rack Mount Servers (for example, C200) allow only local storage of Cisco Unified Communications virtual machines.

UCS C210 servers support more user capacity than UCS C200 servers.

There are specific requirements that must be met in order to run Cisco Unified Communications applications as virtual servers on the UCS C-Series Rack-Mount Servers. These requirements are mentioned in the following document:

http://www.cisco.com/en/US/prod/collateral/voicesw/ps6790/ps5748/ps378/solution_overview_c22-597556.html

Impact of Virtual Servers on Deployment Models

Deploying Cisco Unified Communications applications on virtualized servers supports the same deployment models as when physical servers are used. The chapter on Network Infrastructure, page 3-1, offers some design guidance on how to integrate the QoS capabilities of Cisco UCS B-Series virtualized servers into the network. Also, the integration of physical servers (such as Cisco MCS servers) and Cisco UCS virtual servers is supported in many cases. As an example, music on hold (MoH) servers can run on Cisco MCS server platforms, while also being part of a cluster whose other member servers are run on Cisco UCS virtual servers.

All the call processing deployment models described in this chapter are supported on Cisco UCS virtual server platforms.

Design Considerations for Section 508 Conformance

Regardless of which deployment model you choose, you should consider designing your Cisco Unified Communications network to make the telephony features more accessible to users with disabilities, in conformance with Section 255 of the Telecommunications Act and U.S. Section 508.

Observe the following basic design guidelines when configuring your Cisco Unified Communications network to conform to Section 508:

- Enable Quality of Service (QoS) on the network.
- Configure only the G.711 codec for phones that will be connected to a terminal teletype (TTY) device or a Telephone Device for the Deaf (TDD). Although low bit-rate codecs such as G.729 are acceptable for audio transmissions, they do not work well for TTY/TDD devices if they have an error rate higher than 1% Total Character Error Rate (TCER).
- Configure TTY/TDD devices for G.711 across the WAN, if necessary.
- Enable (turn ON) Echo Cancellation for optimal performance.
- Voice Activity Detection (VAD) does not appear to have an effect on the quality of the TTY/TDD connection, so it may be disabled or enabled.

- Configure the appropriate *regions* and *device pools* in Unified CM to ensure that the TTY/TDD devices always use G.711 codecs.
- Connect the TTY/TDD to the Cisco Unified Communications network in either of the following ways:
  - Direct connection (Recommended method)

    Plug a TTY/TDD with an RJ-11 analog line option directly into a Cisco FXS port. Any Cisco voice gateway with an FXS port will work. Cisco recommends this method of connection.
  - Acoustic coupling

    Place the IP phone handset into a coupling device on the TTY/TDD. Acoustic coupling is less reliable than an RJ-11 connection because the coupling device is generally more susceptible to transmission errors caused by ambient room noise and other factors.
- If stutter dial tone is required, use an analog phone in conjunction with an FXS port on the Cisco VG224 or ATA 187. In addition, most Cisco IP Phones support stutter dial tone, which is sometimes referred to as audible message waiting indication (AMWI).

Call Routing and Dial Plan Distribution Using Call Control Discovery for the Service Advertisement Framework

When multiple call processing agents are present in the same system, each can be configured manually to be aware of the others. This configuration can be time consuming and error prone. Call routing between the various call processing agents requires the configuration of static routes on the call agents and updating them when changes occur.

Instead, the Cisco Service Advertisement Framework (SAF) can be used to share call routing and dial plan information automatically between call agents. SAF allows non-Cisco call agents (such as TDM PBXs) to partake in the Service Advertisement Framework when they are interconnected through a Cisco IOS gateway.

The Service Advertisement Framework (SAF) enables networking applications to advertise and discover information about networked services within an IP network. SAF consists of the following functional components and protocols:

- SAF Clients — Advertise and consume information about services.
- SAF Forwarders — Distribute and maintain SAF service availability information.
- The SAF Client Protocol — Used between SAF Clients and SAF Forwarders.
- The SAF Forwarder Protocol — Used between SAF Forwarders.

The nature of the advertised service is unimportant to the network of SAF Forwarders. The SAF Forwarder protocol is designed to dynamically distribute information about the availability of services to SAF client applications that have registered to the SAF network.

Services that SAF Can Advertise

In theory, any service can be advertised through SAF. The first service to use SAF is Cisco Unified Communications Call Control Discovery (CCD). CCD uses SAF to distribute and maintain information about the availability of internal directory numbers (DNs) hosted by call control agents such as Cisco Unified CM and Unified CME. CCD also distributes the corresponding number prefixes that allow these internal directory numbers to be reached from the PSTN ("To PSTN" prefixes).

The dynamic nature of SAF and the ability for call agents to advertise the availability of their hosted DN ranges and To PSTN prefixes to other call agents in a SAF network, provides distinct advantages over other static and more labor-intensive methods of dial plan distribution.

This chapter discusses the deployment of Call Control Discovery (CCD) in SAF-enabled Unified Communications networks. For more information on SAF itself, see Service Advertisement Framework (SAF), page 3-69.

The following Cisco products support the Call Control Discovery (CCD) service for SAF:

- Cisco Unified Communications Manager (Unified CM) Release 8.0(1) or higher
- Cisco Unified Communications Manager Express (Unified CME) on a Cisco Integrated Services Router (ISR)
- Survivable Remote Site Telephony (SRST) on a Cisco ISR platform
- Cisco Unified Border Element on a Cisco ISR platform
- Cisco IOS Gateways on a Cisco ISR platform

CCD is supported on Cisco ISR platforms running Cisco IOS Release 15.0(1)M or higher. For more information on Cisco IOS Release 15.0(1)M, refer to the following websites:

- http://wwwin.cisco.com/ios/release/15mt
- http://www.cisco.com/en/US/products/ps10621/index.html

For information on the use of CCD with Unified CM, refer to the *Cisco Unified Communications Manager Features and Services Guide*, available at

http://www.cisco.com/en/US/products/sw/voicesw/ps556/prod_maintenance_guides_list.html

SAF Service IDs

CCD is the first SAF service. SAF services are identified to a network of SAF Forwarders and Clients by their SAF Service ID. CCD for Unified Communications uses a SAF Service ID of 101:2:*x.x.x.x*, where:

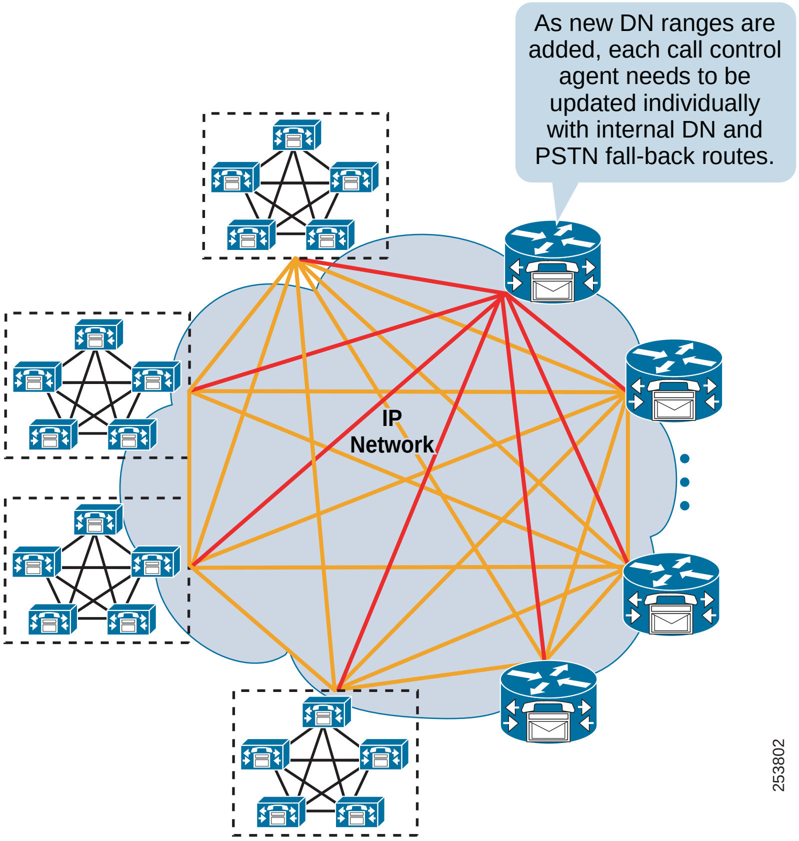
- Service ID 101 = Unified Communications
- Sub-Service ID 2 = CCD
- Instance ID *x.x.x.x* = ID of Unified CM cluster (PKID) or Cisco IOS device

Deploying SAF CCD Within Your Network

The SAF CCD service allows information about the location and availability of directory number ranges hosted by call control agents, such as Unified CM and Unified CME, to be propagated dynamically within a SAF-enabled Unified Communications network.

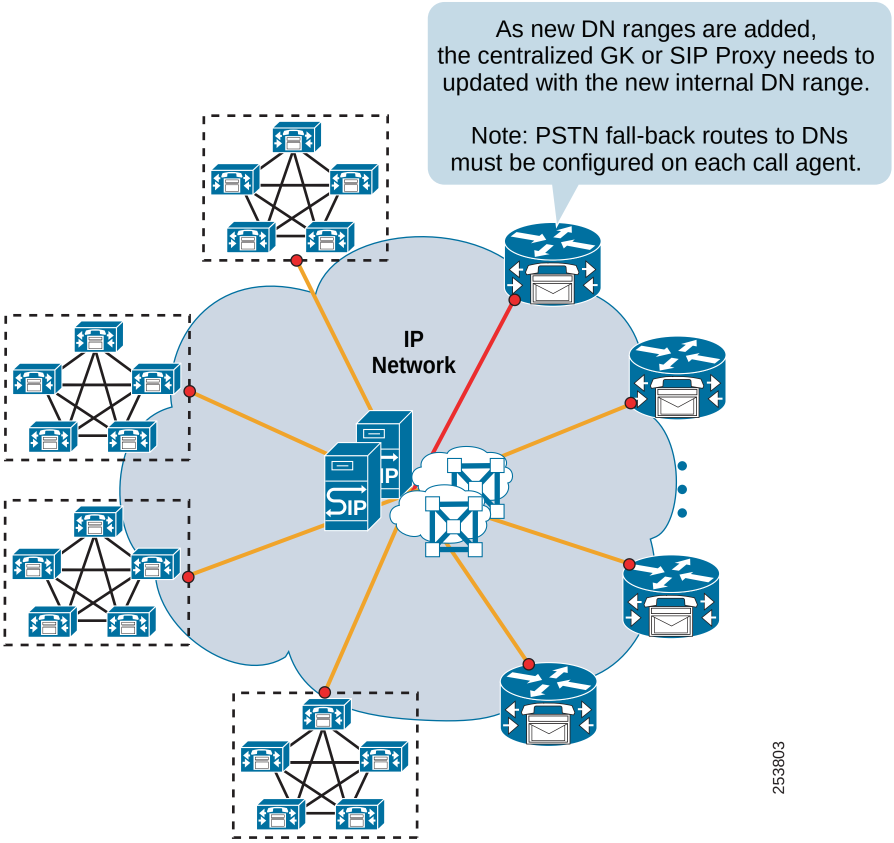
The advantages of deploying SAF to distribute and maintain DN information can be understood by considering the management of the dial plan in an example Unified Communications network consisting of four Unified CM clusters and 40 Unified CMEs. In a statically configured network, as new directory number ranges are introduced within the Unified Communications system, details of how those new number ranges can be reached must be made available to all other call control applications within the Unified Communications network. In the worst case, with a full mesh of connections between all call control applications, each call control application must be updated with information about each new number range and how it can be reached (see Figure 5-13). This cascade of configuration changes is time consuming, error prone, and requires significant ongoing management.

Figure 5-13 A Full Mesh of Connections Between Call Control Applications


As new DN ranges are added, each call control agent needs to be updated individually with internal DN and PSTN fall-back routes.
IP Network
253802

The dial plan can be centralized on a Session Management Edition cluster, H.323 gatekeeper, or SIP proxy (see Figure 5-14). This reduces configuration overhead, but it allows only the internal dial plan to be centralized. If access to the centralized dial plan is unavailable, alternative routes such as PSTN routes can be used only if they are configured as backup routes in each call control application.

Figure 5-14 A Centralized Internal Dial Plan


As new DN ranges are added, the centralized GK or SIP Proxy needs to updated with the new internal DN range.

Note: PSTN fall-back routes to DNs must be configured on each call agent.
IP
Network
IP
SIP
253803

SAF CCD enables each call control application to advertise its directory number ranges and their corresponding "To PSTN" prefixes to all other call control applications in the SAF network. (See Figure 5-15 and Figure 5-16.) In doing so, SAF CCD removes the following restrictions:

- The need for a centralized application that hosts the internal system-wide dial plan.
- The requirement to configure each call control application individually as new DN ranges and their corresponding "To PSTN" prefixes are added to the Unified Communications network.

Furthermore, SAF CCD is dynamic rather than static in nature. When DN ranges are deleted or IP connectivity is lost to the call control application, the SAF network automatically updates all other call control applications by withdrawing the routes to the unavailable DNs. Likewise, when connectivity is reestablished (or DN ranges are reconfigured), the SAF network updates all other call control applications, thus reinstating the routes to the DN ranges.

Figure 5-15 Advertising Unified CM Internal DN Ranges and Corresponding "To PSTN" Prefixes to the SAF Network

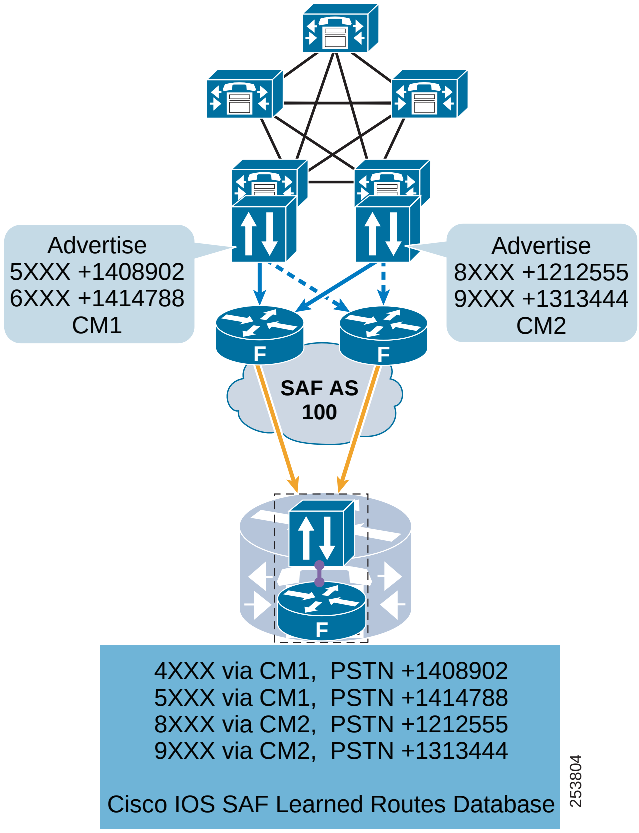
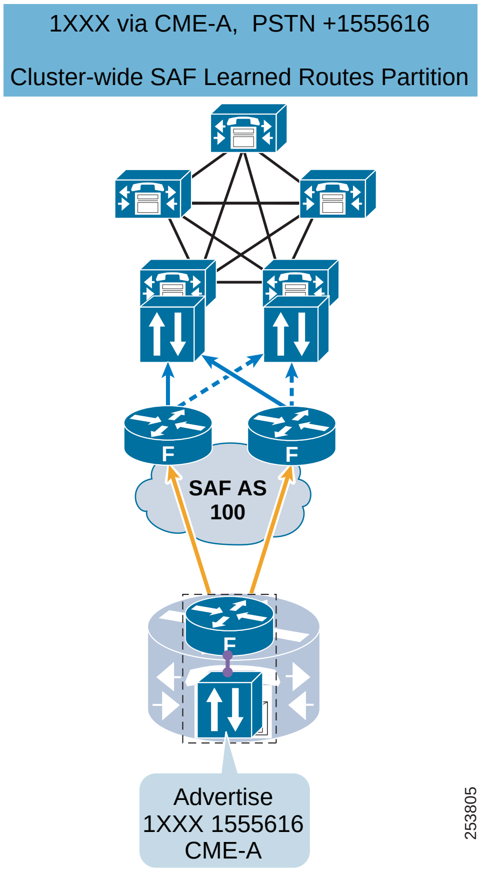

Advertise
5XXX +1408902
6XXX +1414788
CM1
Advertise
8XXX +1212555
9XXX +1313444
CM2
F
F
SAF AS
100
F
4XXX via CM1, PSTN +1408902
5XXX via CM1, PSTN +1414788
8XXX via CM2, PSTN +1212555
9XXX via CM2, PSTN +1313444
Cisco IOS SAF Learned Routes Database
253804

Figure 5-16 Advertising Unified CME Internal DN Ranges and Corresponding "To PSTN" Prefixes to the SAF Network

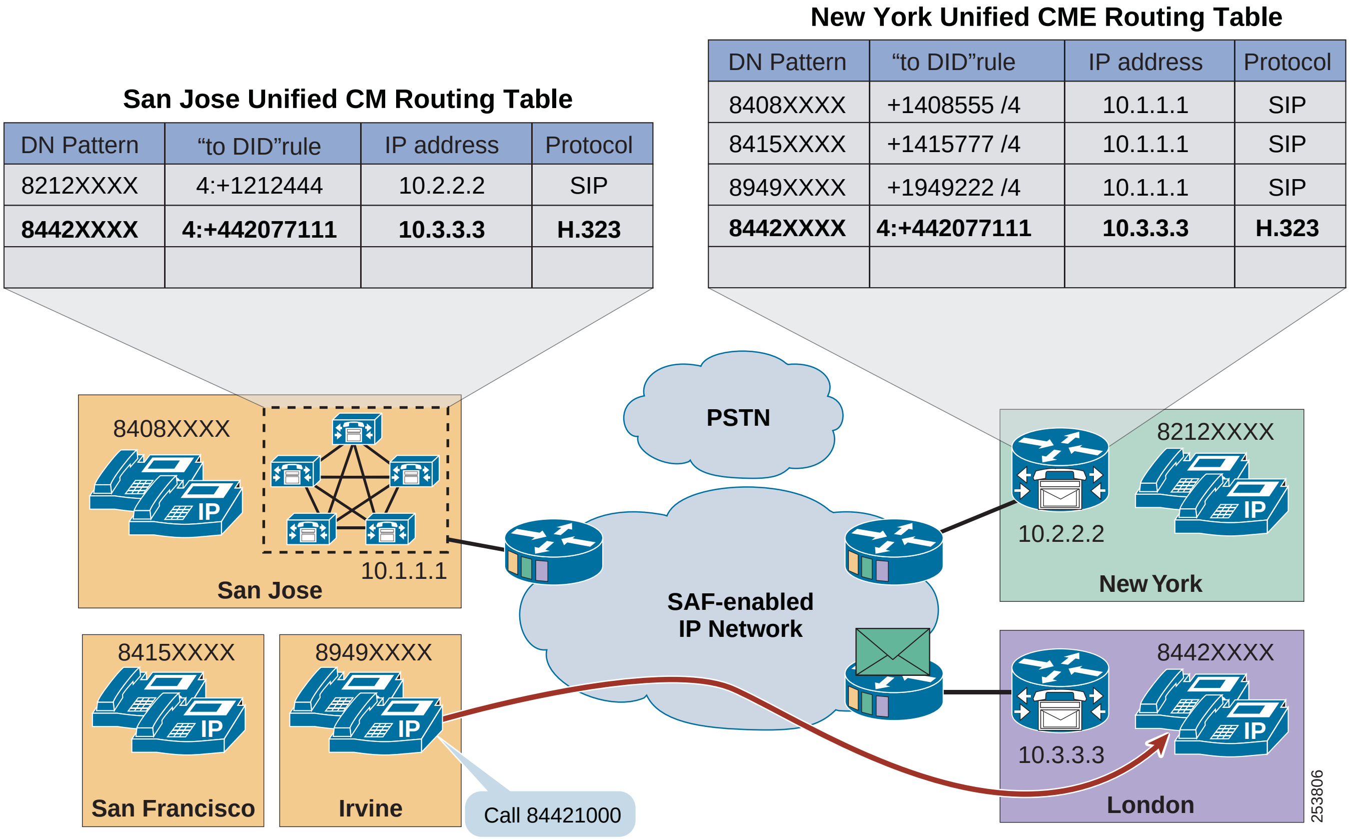

1XXX via CME-A, PSTN +1555616
Cluster-wide SAF Learned Routes Partition
F
F
SAF AS
100
F
Advertise
1XXX 1555616
CME-A
253805

Comparison of SAF CCD Operation and Standard Unified CM Call Routing

Call routing using SAF CCD is fundamentally different than standard Unified CM call routing, which uses route patterns, route lists, and route groups that are not used by SAF CCD. Instead, the directory numbers, directory number ranges, and "To PSTN" prefixes to remote endpoints are learned dynamically by a SAF CCD-enabled cluster rather than being configured statically (see Figure 5-17). With SAF CCD, each Unified CM cluster (or other SAF-enabled call control application) configures which directory numbers, DN ranges, and so forth, that it wishes to advertise to the SAF network. SAF CCD also advertises the means by which to reach these numbers, by advertising the IP addresses and port numbers of the SAF-enabled SIP or H.323 trunks in the cluster.

Each SAF-enabled cluster also listens for advertisements from other clusters about their DNs, DN ranges, associated "To PSTN" Prefixes, and trunk information. These SAF learned routes are placed into a single partition. Any device that has access to this partition can reach any device advertised within SAF. Cisco recommends SAF CCD for the distribution of internal DN ranges only and their To PSTN routes.

Figure 5-17 Dynamic Call Routing with SAF CCD



San Jose Unified CM Routing Table

DN Pattern	"to DID"rule	IP address	Protocol
8212XXXX	4:+1212444	10.2.2.2	SIP
8442XXXX	**4:+442077111**	**10.3.3.3**	**H.323**

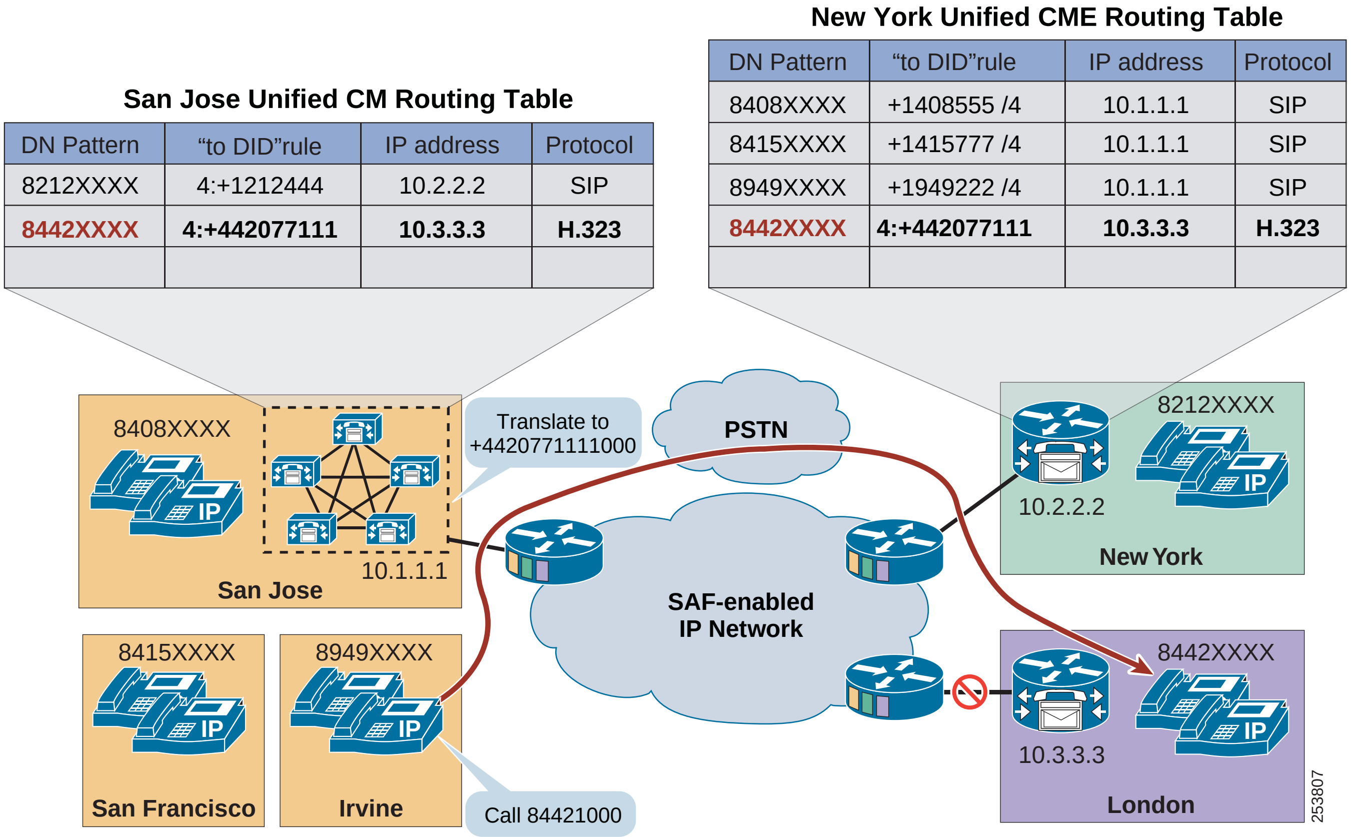
New York Unified CME Routing Table

DN Pattern	"to DID"rule	IP address	Protocol
8408XXXX	+1408555 /4	10.1.1.1	SIP
8415XXXX	+1415777 /4	10.1.1.1	SIP
8949XXXX	+1949222 /4	10.1.1.1	SIP
8442XXXX	**4:+442077111**	**10.3.3.3**	**H.323**

Any call made using SAF learned routes has automatic PSTN failover if the IP path to the called number is not available (see Figure 5-18). The call is routed according to the following order:

- Take the selected IP path to reach the called number.
- If the IP path is not available, use the PSTN prefix to modify the called number and route the call through the PSTN.

Figure 5-18 Automatic PSTN Failover with SAF CCD



San Jose Unified CM Routing Table

DN Pattern	"to DID"rule	IP address	Protocol
8212XXXX	4:+1212444	10.2.2.2	SIP
8442XXXX	**4:+442077111**	**10.3.3.3**	**H.323**

New York Unified CME Routing Table

DN Pattern	"to DID"rule	IP address	Protocol
8408XXXX	+1408555 /4	10.1.1.1	SIP
8415XXXX	+1415777 /4	10.1.1.1	SIP
8949XXXX	+1949222 /4	10.1.1.1	SIP
8442XXXX	**4:+442077111**	**10.3.3.3**	**H.323**

SAF CCD is different than standard call routing in that only a single IP route can be chosen for a given SIP or H.323 call, whereas with standard call routing, multiple IP paths may be defined and consecutively attempted for a single call by using route lists and route groups.

CCD and Unified CM

CCD enables Unified CM to advertise multiple directory numbers, directory number ranges, and their corresponding "To PSTN" prefixes to a SAF-enabled network. CCD introduces several new configurable components in Unified CM:

- SAF Forwarder Configuration (the external SAF Client on Unified CM)
- SAF Enabled Trunks
- Hosted DN Patterns
- Hosted DN Groups
- CCD Advertising Service
- CCD Requesting Service

SAF Forwarder Configuration (External SAF Client on Unified CM)

The SAF Forwarder Configuration on Unified CM represents the configuration of the External SAF Client to a SAF Forwarder in a Unified Communications network. The Unified CM SAF Forwarder configuration defines the following items:

- The destination IP address and port number of the remote SAF Forwarder
- The Security Profile (username and password) used to authenticate with the SAF Forwarder
- The Client Label

  This is a string that the SAF Forwarder uses to map the Unified CM external client into a specific SAF Autonomous System. Cisco IOS supports bulk provisioning of the Client Label, whereby a client-label string that ends with an @ is considered as a base name or label. A base label configured on a router will accept any character following the @ in the base name as a valid client-label to identify a client in the REGISTER message sent by an external client.

For example, Unified CM cluster A can use CUCM-A as the base name for the cluster and can append a number after the @ following the base name for each configured SAF Forwarder (external SAF Client in Unified CM). By defining the external client CUCM-A as a base name in Cisco IOS, the Cisco IOS forwarder will accept any client label beginning with CUCM-A@, such as any of the following labels:

- CUCM-A@Client-1
- CUCM-A@Client-2
- CUCM-A@Client-3
- CUCM-A@Client-4

This allows SAF Clients 1 through 4 to register with the same SAF Forwarder and SAF autonomous system (AS).

External SAF Client Instance Creation and Activation within the Unified CM Cluster

By default, an instance of the external SAF client configured through the SAF Forwarder Configuration page in Unified CM is created on every call processing node within the cluster (see Figure 5-19). The external SAF client is activated only if an instance of a CCD Advertising Service or the CCD Requesting Service is also active on the call processing node. The activation of Advertising and Requesting Services on call processing nodes is determined by the SAF trunks associated with each service. (For details, see CCD Advertising and Requesting Services, page 5-63.)

Figure 5-19 Single SAF Forwarder Defined in Unified CM

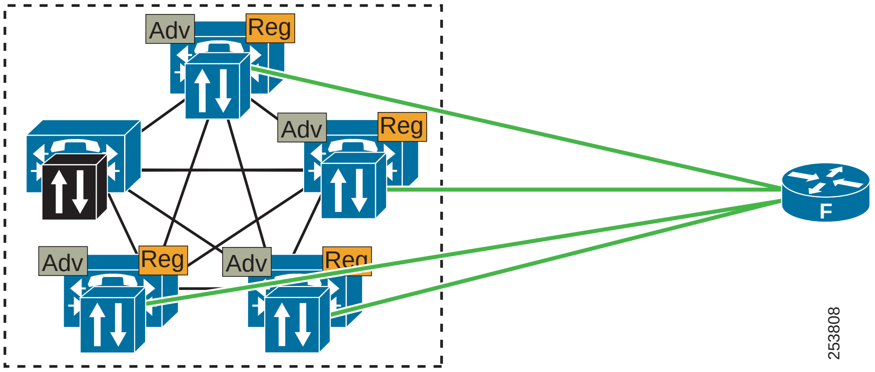
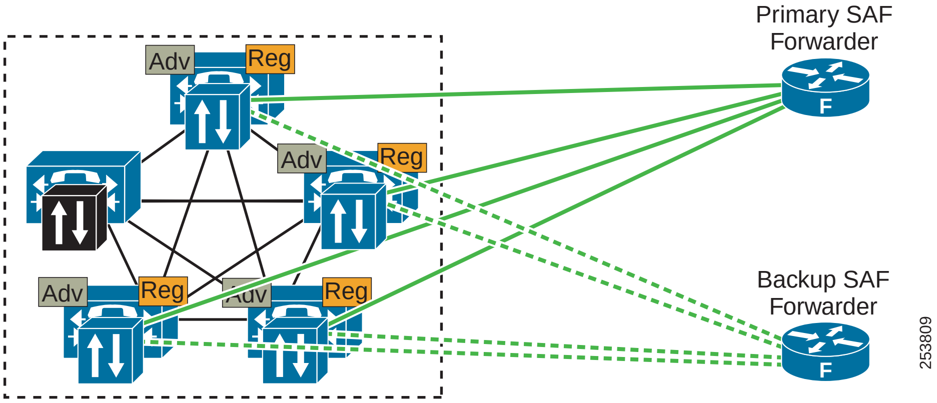


Figure 5-19 shows four active External SAF Clients connecting to a single SAF Forwarder. (The greyed-out SAF client is not activated because there is no active Advertising or Requesting Service associated with that Unified CM node). Each active External SAF client establishes a connection to the SAF Forwarder, registers with the SAF network, publishes its associated Services, and subscribes to the SAF CCD service active in the SAF AS. Such duplication can be useful for resilience and redundancy, but it can also create overhead within the cluster and the SAF Forwarder. By carefully selecting where the Advertising and Requesting Services run within the cluster, you can fine-tune this duplication and redundancy. For more information, see the CCD Advertising and Requesting Services, page 5-63.

Multiple SAF Forwarders

You can configure multiple SAF Forwarders within a cluster for redundancy. The SAF Client establishes a secure connection to the primary and the backup SAF Forwarders, registers with the SAF Forwarders, and sends a publish request for the HostedDN service to the primary SAF Forwarder. The SAF Client makes an arbitrary decision on selecting one SAF Forwarder as primary and another as backup at system startup time, based on the first SAF Forwarder to respond to a registration request from the client. The SAF Client publishes and subscribes services to the primary SAF Forwarder only. The SAF Client maintains the connection to the SAF Forwarder by sending keepalives to the SAF Forwarder at regular intervals. If the connection to the primary SAF Forwarder fails, the SAF Client switches to the backup SAF Forwarder, sending all the publish and subscription requests to the backup SAF Forwarder that it had sent to the primary SAF Forwarder.

Figure 5-20 Two SAF Forwarders Defined in Unified CM



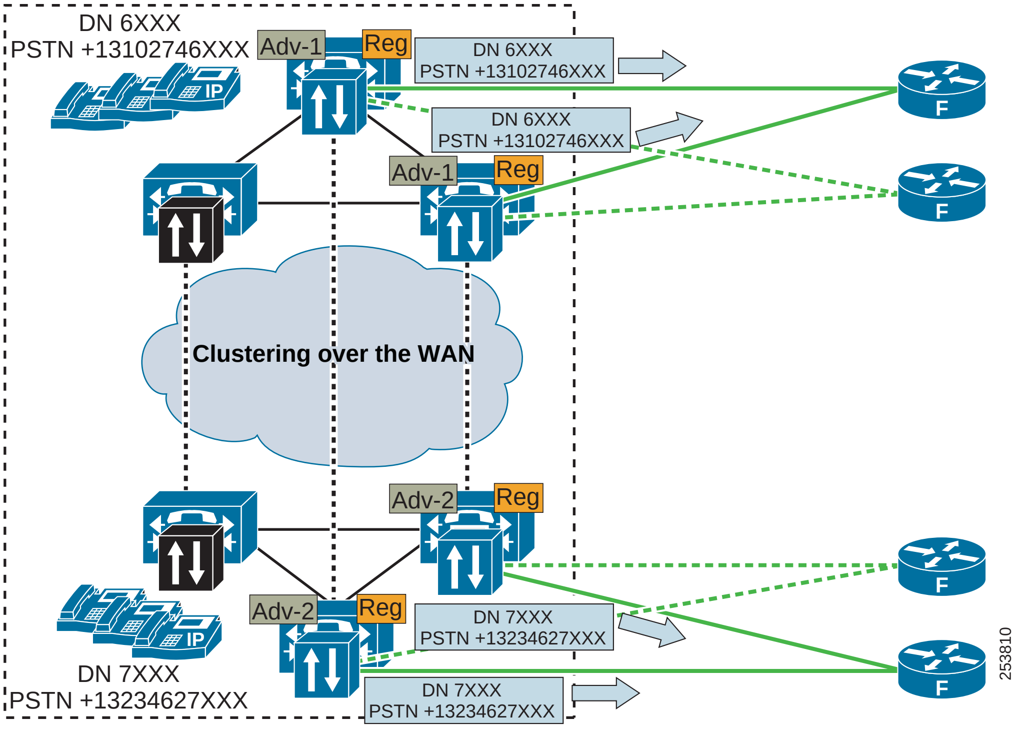
Advanced SAF Client Configuration

By default, for each configured SAF Forwarder a corresponding instance of the SAF client is created on every call processing node within the Unified CM cluster. Using the advanced SAF Forwarder configuration option, the administrator can create the SAF client on selected call processing nodes within the cluster. This configuration option enables the administrator to create the SAF client on specific nodes within the cluster and to configure SAF CCD with spatial distribution of CCD services for systems that employ clustering over the WAN.

SAF CCD and Clustering over the WAN

By creating multiple SAF client instances and multiple Advertising Services and associating them with specific Unified CM nodes within a cluster that uses clustering over the WAN, you can advertise CCD Hosted Directory Number ranges into the SAF network, with a geographical association to their local Unified CM trunks and nodes within the cluster.

Figure 5-21 SAF CCD-Selected SAF Client Configuration for Clustering over the WAN



SAF-Enabled Trunks

SAF-enabled trunks are used solely to route calls between SAF-enabled call control applications. They cannot be used with standard route patterns, route lists, and route groups. You cannot configure the destination address of a SAF-enabled trunk because this destination address is learned through SAF; however, you can configure all other trunk parameters.

You can enable SAF on the following trunk types:

- SIP trunks — Enabled by selecting **Call Control Discovery** as the Trunk Service Type when creating a new SIP trunk.
- H.323 Non-Gatekeeper controlled intercluster trunks — Enabled by checking the **Enable SAF** check box on the Trunk configuration page.

Both of these trunk types may be used between Unified CM clusters and between Unified CM and Cisco IOS gateways.

CCD uses SAF-enabled trunks for two purposes:

- To originate calls — These SAF-enabled trunks are associated with the CCD Requesting Service.
- To accept incoming calls — These SAF-enabled trunks are associated with the CCD Advertising Service. The IP addresses and port numbers of these SAF-enabled trunks are published with the DN ranges associated with the Advertising Service.

A SAF-enabled trunks can be used by both the Advertising and Requesting Service.

When the CCD Advertising Service publishes the trunk details for a hosted DN range, it sends the IP address and port number of each Unified CM node in the SAF trunk's Cisco Unified Communications Manager Group in separate SAF advertisements. For example, to advertise hosted DN range 5*XXX* from SIP trunk A, which has CUCM1 and CUCM2 in its Cisco Unified Communications Manager Group, the CCD Advertising Service would publish two advertisements:

- 5XXX via SIP trunk IP address (CUCM1) port number 5060
- 5XXX via SIP trunk IP address (CUCM2) port number 5060

The Requesting Service of the cluster receiving this advertisement would place two routes to 5XXX in its SAF learned routes partition:

- 5XXX via SIP trunk IP address (CUCM1) port number 5060
- 5XXX via SIP trunk IP address (CUCM2) port number 5060

Calls to 5XXX from this cluster would select the two available SIP trunk destinations in round-robin order.

SAF trunks support TCP or UDP transport protocols. Because a SAF trunk can accept incoming calls from multiple call control applications, TLS-based Signalling Authentication and Encryption is not supported over SAF-enabled trunks.

Hosted DN Patterns and Hosted DN Groups

Hosted DN groups represent groups of hosted DN patterns. The hosted DN patterns in a hosted DN group typically represent the range of directory numbers associated with a physical site. Digit strip and prepend information for "To PSTN" failover routing can be configured for each hosted DN group. The same DN pattern cannot be associated with multiple hosted DN groups.

A hosted DN pattern can define a single directory number (for example, 5000), or a range of directory numbers (for example, 5*XXX*). Every DN pattern must be unique. Each hosted DN pattern can be configured with digit strip and prepend information for PSTN failover routing. The PSTN failover configuration on the hosted DN pattern takes precedence over the PSTN failover configuration at the hosted DN group level.

CCD Advertising and Requesting Services

CCD uses two Unified CM services to communicate with the SAF network: the Advertising Service, which is used to publish DN ranges and their associated trunks to the SAF network, and the Requesting Service, which is used to learn about the reachability of DN ranges from other call agents in the SAF network. The following sections describe these two services.

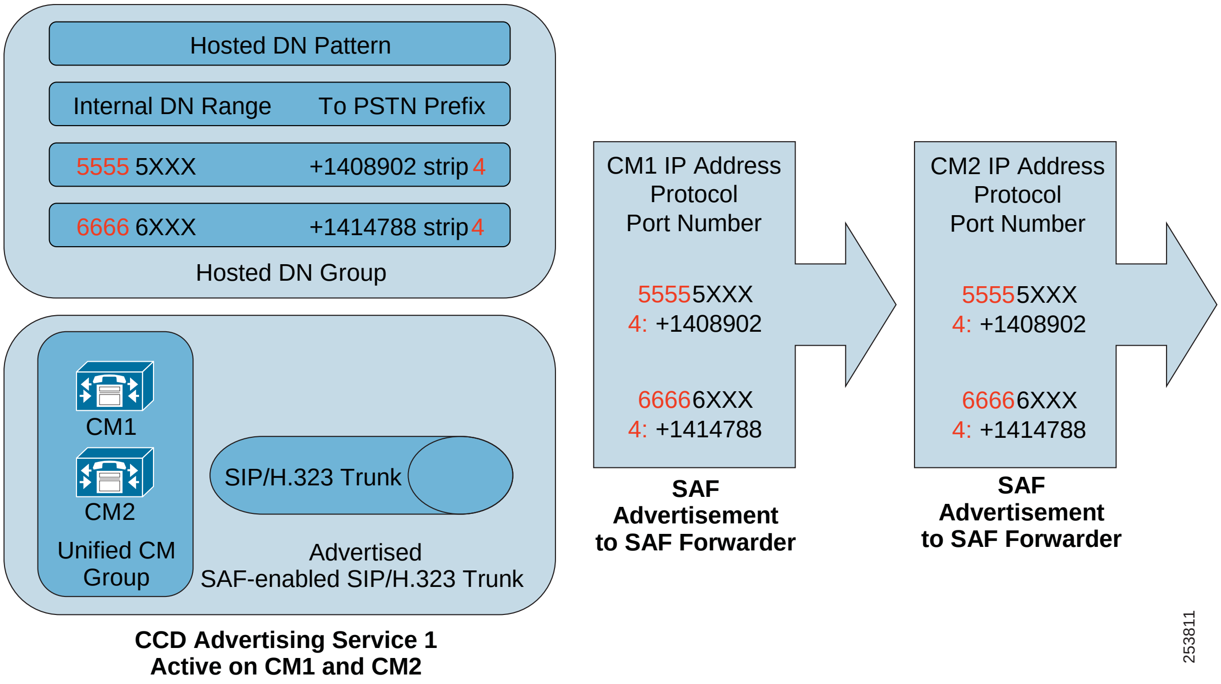
CCD Advertising Service

The CCD Advertising Service associates one hosted DN group with a SAF-enabled SIP and/or H.323 trunk. The Advertising Service is created and activated on each server in the Cisco Unified Communications Manager Group (Unified CM Group) of its associated SAF-enabled trunk(s). The

Advertising Service uses the SAF Client on each of the servers in the Unified CM Group of each trunk to publish information about the group of hosted DNs and associated trunk nodes to the client's SAF Forwarder. (See Figure 5-22.)

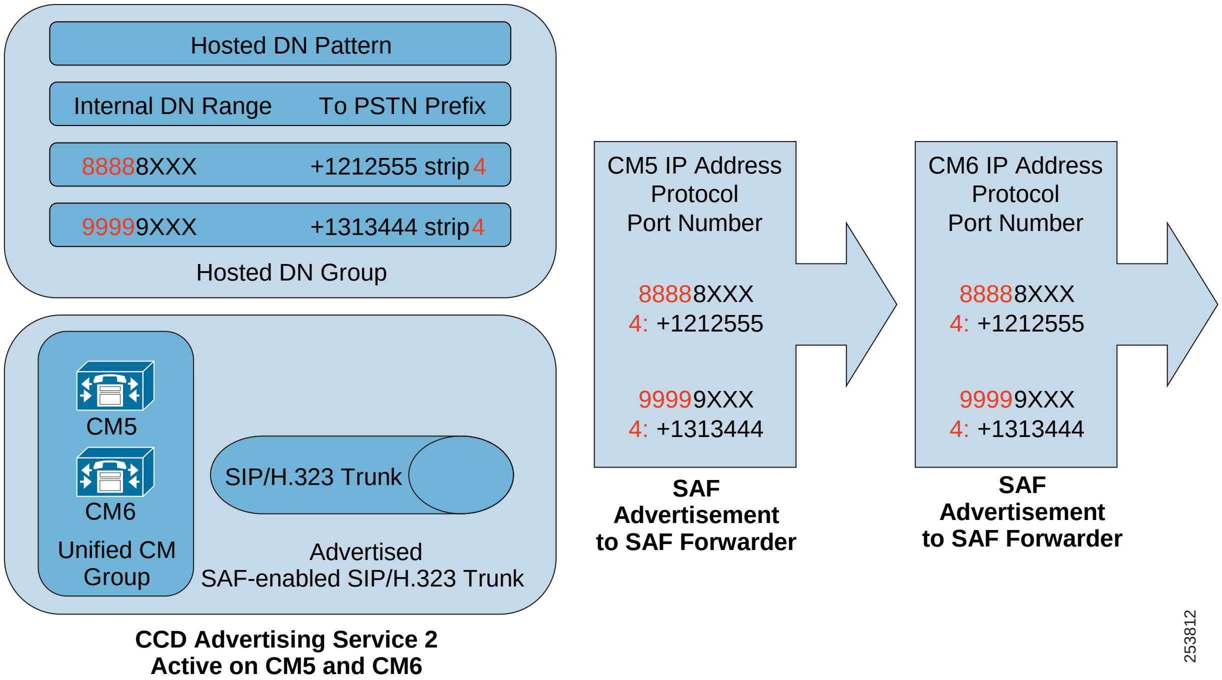
Because SIP and H.323 trunks support different feature sets (for example, H.323 trunks support QSIG over Annex M1), it is typical to select only one trunk type per Advertising Service. If both an H.323 and a SIP trunk are selected, calls to the hosted DN ranges associated with this Advertising Service will be distributed in a round-robin fashion across both the SIP and H.323 trunks.

Figure 5-22 CCD Advertising Service 1 Active on CM1 and CM2



You can create multiple advertising services within a Unified CM cluster. An Advertising Service can use the same (or different) SAF-enabled trunks as other Advertising Services. However, each Advertising Service must be associated with a unique hosted DN group, and the same hosted DN pattern cannot be advertised by multiple Advertising Services within a cluster. Creating multiple Advertising Services allows inbound calls to be distributed by DN range across multiple trunk servers within a cluster. (See Figure 5-23.)

Figure 5-23 CCD Advertising Service 2 Active on CM5 and CM6

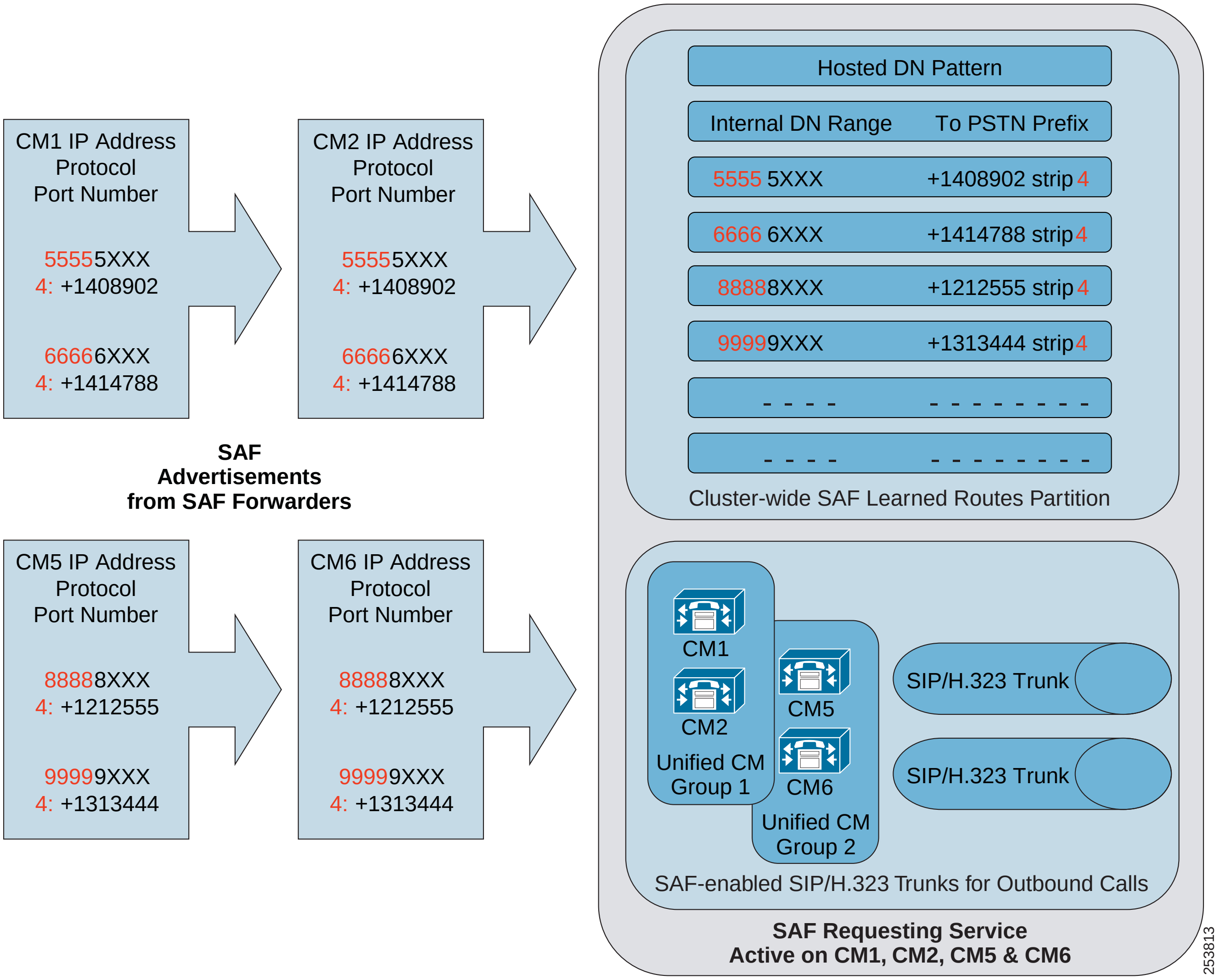


CCD Requesting Service

The CCD Requesting Service collects information about hosted DN routes advertised in the SAF AS and places them into a partition for SAF learned routes. (See Figure 5-24.) The Requesting Service is also used to select which SAF trunks will be used to initiate outbound SAF calls. More than one SAF-enabled trunk can be selected. If multiple trunks are selected, these SAF trunks and their corresponding Unified CM Group server nodes are selected on a round-robin basis for outbound calls. Similar to the Advertising Service, trunks of the same protocol type are usually associated to the Requesting Service. The Requesting service also allows digits to be prefixed to learned DN patterns and learned "To PSTN" patterns.

Only a single Requesting Service can be configured in the Unified CM cluster, and the Requesting Service is activated on all of the nodes in the Unified CM Groups of its associated SAF trunks.

Figure 5-24 Unified CM CCD Requesting Service



Blocking CCD Learned Patterns

Unified CM enables the SAF CCD administrator to purge and block learned route information from the SAF CCD learned routes partition. Routes can be blocked based on whether they match one or more of the following entries:

- Learned Pattern (for example, 500*X*)
- Learned Pattern Prefix (for example, +1408)
- Remote Call Control Entity Name (This is the Unified CM Cluster ID in Enterprise Parameters.)
- Remote Call Control IP Address (This could be the address of a Cisco IOS SAF CCD router or one or more Unified CM servers in a Unified CM cluster.)

If required, these entries can be used in a logical AND combination such as the following:

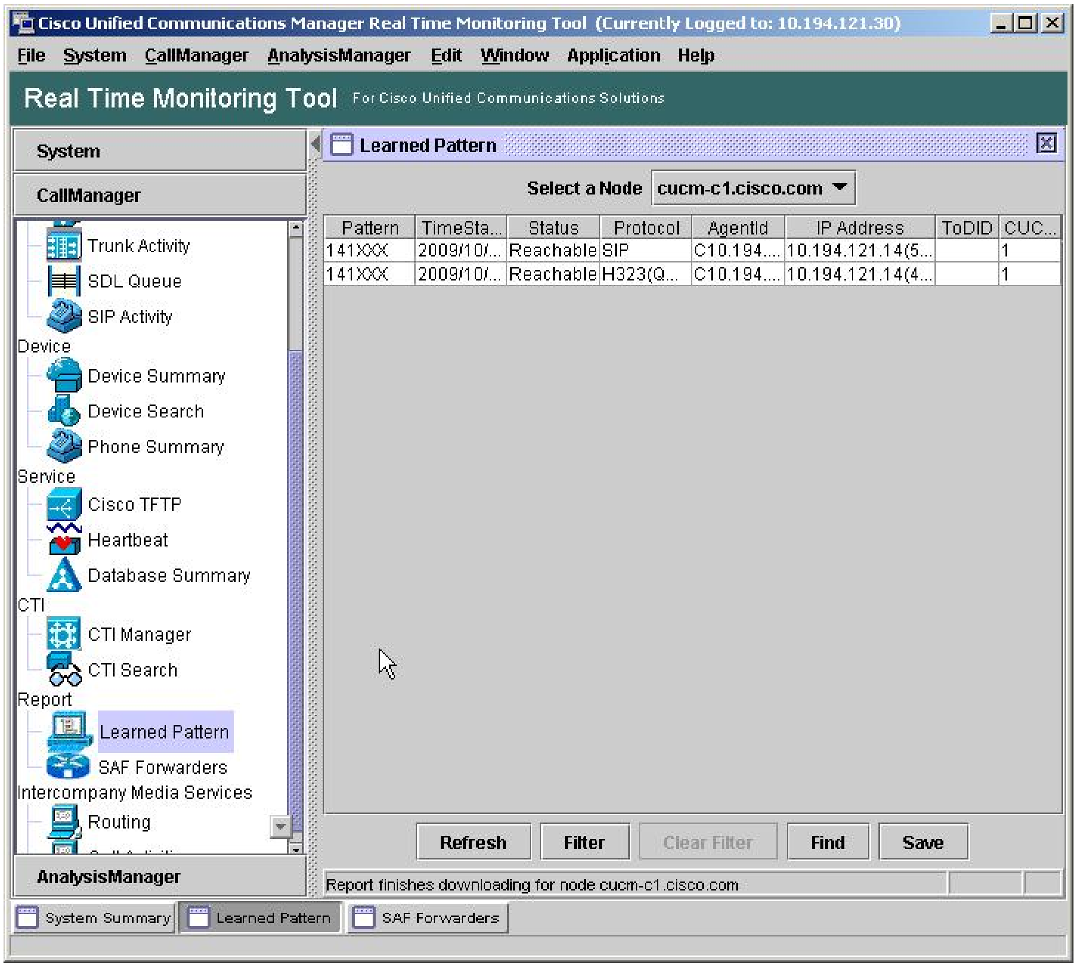
Pattern = "5XXX" AND Prefix = "+1408" AND Remote Call Control Address = "10.10.1.1"

Blocking CCD learned patterns can be particularly useful in SAF CCD deployments where a Unified CM cluster connects to multiple SAF ASs and wishes to advertise DN route information to an AS but does not wish to receive some or all of the DN route information being sent by the AS.

Displaying SAF Learned Routes in Unified CM

Because SAF learned routes are dynamic in nature, they are not held in the Unified CM database but are stored in memory. Use the Cisco Unified Communications Manager Real-Time Monitoring Tool (RTMT) to display SAF learned routes and to monitor SAF Forwarders (see Figure 5-25).

Figure 5-25 Real-Time Monitoring Tool (RTMT) for SAF CCD



Select a Node cucm-c1.cisco.com

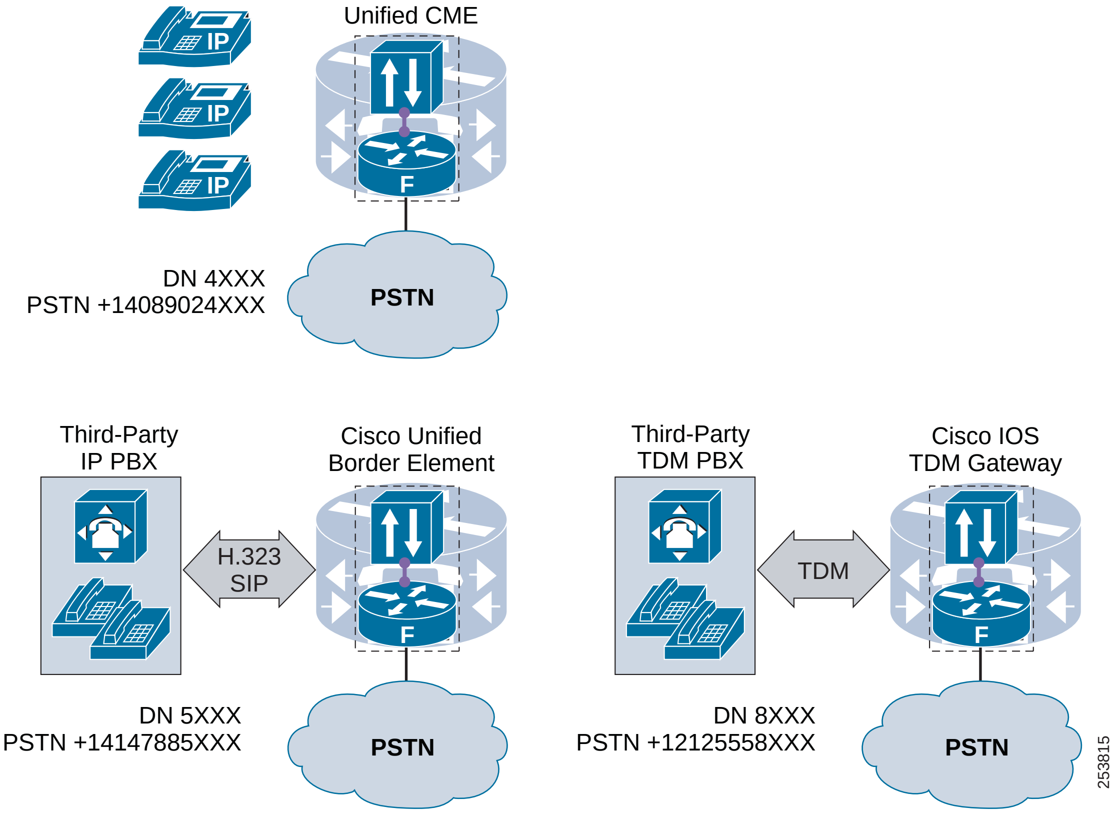
Pattern	TimeSta...	Status	Protocol	AgentId	IP Address	ToDID	CUC...
141XXX	2009/10/...	Reachable	SIP	C10.194....	10.194.121.14(5...		1
141XXX	2009/10/...	Reachable	H323(Q...	C10.194....	10.194.121.14(4...		1

Report finishes downloading for node cucm-c1.cisco.com

Cisco IOS-Based SAF CCD

Cisco IOS-based SAF CCD is supported by Unified CME, SRST, Cisco Unified Border Element, and Cisco IOS Gateways on the Integrated Services Router (ISR) platform with Cisco IOS Release 15.0(1)M. (See Figure 5-26.) The configuration of Cisco IOS SAF CCD is the same across all of these products. SRST, however, is a special case of CCD and is discussed in the section on SAF CCD and SRST, page 5-71.

Figure 5-26 Cisco IOS-Based SAF CCD Call Agents



For Unified CME, Cisco IOS TDM gateways, and Cisco Unified Border Element, SAF CCD can be used to advertise the internal directory number ranges and "To PSTN" prefixes of the endpoints associated with each of these products and also to subscribe to SAF advertisements from other SAF CCD-enabled call control applications.

For both Cisco IOS and Unified CM, Cisco does *not* recommend using SAF CCD to advertise external PSTN number ranges (for example, for tail-end hop off) for the following key reasons:

- SAF CCD provides no information about the capacity of IP, PSTN, or TDM trunks. (For example, an ISDN BRI with two DS0s and a T1 TDM interface with 24 DS0s would be weighted equally by SAF CCD.)
- All SAF CCD routes are placed into a single partition. This means that any SAF CCD user has access to all learned SAF CCD routes and that no SAF CCD classes of service can be created.

Although the principles of Cisco IOS SAF CCD configuration are the same as those for Unified CM, the naming conventions and commands are different.

Internal SAF Clients

For Cisco IOS-based SAF CCD applications, the SAF Client and Forwarder are co-resident within Cisco IOS. Configuration and authentication is not required between the internal SAF Client and internal SAF Forwarder.

External SAF Clients

To enable the authentication of an external SAF Client to a Cisco IOS SAF Forwarder, use the **external-client** Cisco IOS command to define the external client's label or base name, username, password, and keepalive timer.

SAF-Enabled Trunks

SAF trunks are defined under the **profile trunk-route** Cisco IOS command. The trunk-route profile defines the IP address, port number, protocol (SIP or H.323), and transport protocol (UDP or TCP) for the SAF trunk.

DN Patterns, DN Blocks, and DN Service

The definition and configuration of directory numbers, DN ranges and "To PSTN" prefixes is slightly different in Cisco IOS when compared with Unified CM configuration. Cisco IOS uses the concept of DN blocks to group DN numbers and DN ranges. A DN block can contain more than one DN pattern. The "To PSTN" failover rules for stripping and prefixing digits are also defined at the DN block command line. The PSTN failover rule is known as an **alias** in Cisco IOS. (The PSTN failover rule is applied to the concatenated Site Code and Extension DN Pattern.) The following example shows the Cisco IOS configuration for a DN block:

```
profile dn-block 1 alias 1408902 strip 3
    pattern 1 extension 5xxx
    pattern 2 extension 6xxx
```

Call Control Profile, DN Service, and Site Code

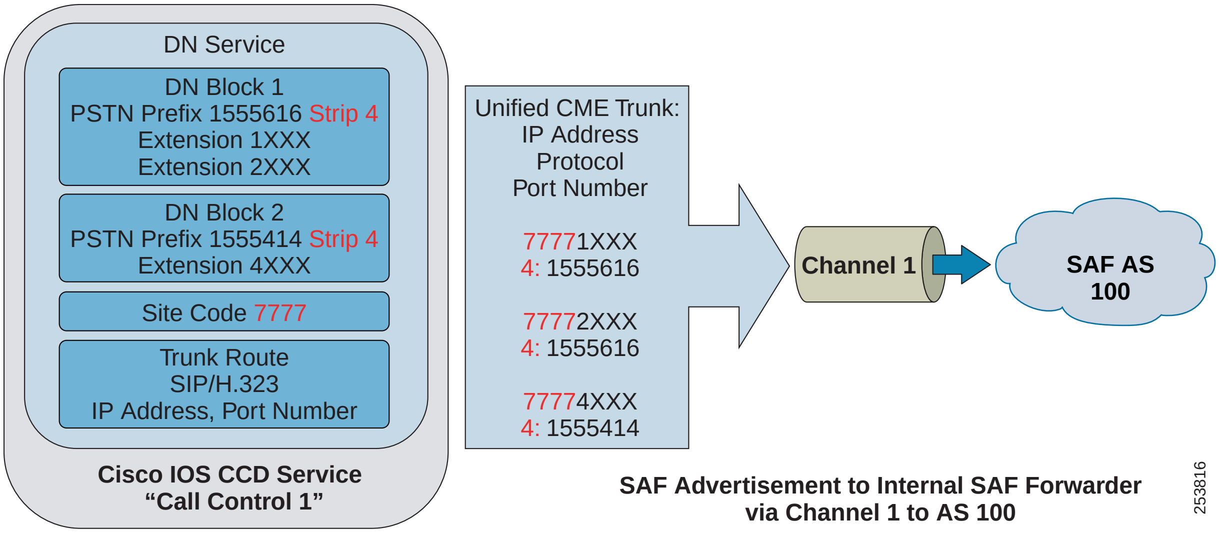
The CCD call control profile is associated with a DN service. A DN service in Cisco IOS can be considered to be equivalent to an Advertising Service in Unified CM. The DN service is used to group one or more DN blocks, one trunk route, and one site code. If present, the site code consists of one or more digits that are prepended to the advertised extension DN patterns.

Multiple call control profiles can be created. The same DN blocks, trunk routes, and site codes can be reused in multiple call control profiles, but only one profile can be associated with a SAF AS.

Publishing and Subscribing to SAF Services within a SAF AS

Call control profiles advertise their associated DN ranges, "To PSTN" failover rules, and trunk route to one SAF AS by means of a configured SAF "channel." A SAF channel can publish the CCD service information contained in only one call control profile to a single SAF AS. (See Figure 5-27.)

Figure 5-27 Cisco IOS CCD Service Call Control 1 Advertising Through Channel 1 to SAF AS 100



A SAF Channel can subscribe to all CCD services within a SAF AS using a wildcard service ID, or up to two selected SAF CCD services that are identified by the instance values in the SAF service ID. (The instance value for Unified CM is the cluster PKID.) For example:

Wildcard SAF Service ID =

Service:	Sub-service:	Instance.	Instance.	Instance.	Instance.
101:	2:	FFFFFFFF.	FFFFFFFF.	FFFFFFFF.	FFFFFFFF.

Tip Use the Cisco IOS command **show eigrp service-family ipv4 [***AS number***] events** to display the Service ID for the Cisco IOS SAF CCD service on the router. The Service ID will be displayed as "connected" (for example, 101:2:59F8412.0.0.6F0100).

Outbound SAF CCD Calls in Cisco IOS

Cisco IOS adds SAF as a configurable session target to standard Cisco IOS voice dial peers. Dial peers can also be assigned a preference setting to control the order in which standard and SAF dial peers are selected.

SAF CCD and SRST

SRST CCD is a special type of SAF deployment. SRST CCD does not advertise any number ranges into SAF; it only listens to the advertisements from other SAF CCD services such as Unified CM, Unified CME, and so forth. SRST CCD does not use SAF learned IP routes at any time; only PSTN routes are used and only when the router and associated phones are in SRST mode.

You can use SAF for SRST CCD to avoid the labour-intensive task of updating every SRST router with a new number expansion rule every time a new SRST router is added to the Unified Communications network.

With standard (non-SAF) SRST operation, if Unified CM becomes unavailable, phones register their extension numbers to their SRST reference router. (See Figure 5-28.) In SRST mode, calls can be made to other phones registered to the SRST router by dialing their extension number as normal. When a phone in SRST mode is used to call a phone in another site, the PSTN number of the called phone must be dialed. (See Figure 5-29.) The number expansion command in Cisco IOS, much like the PSTN failover rule in SAF CCD, allows the dialed extension number to be expanded to the full PSTN number in SRST mode.

In a Unified Communications deployment with many SRST routers, when a new SRST router is added to the Unified Communications network, every SRST router must add a number expansion rule that corresponds to the PSTN access prefix for this new SRST site.

SAF for SRST CCD allows the PSTN failover rules for every SRST site to be distributed to every SRST router within the SAF AS.

Figure 5-28 Normal (Unified CM) Operation of a Unified CM Deployment with SAF SRST CCD

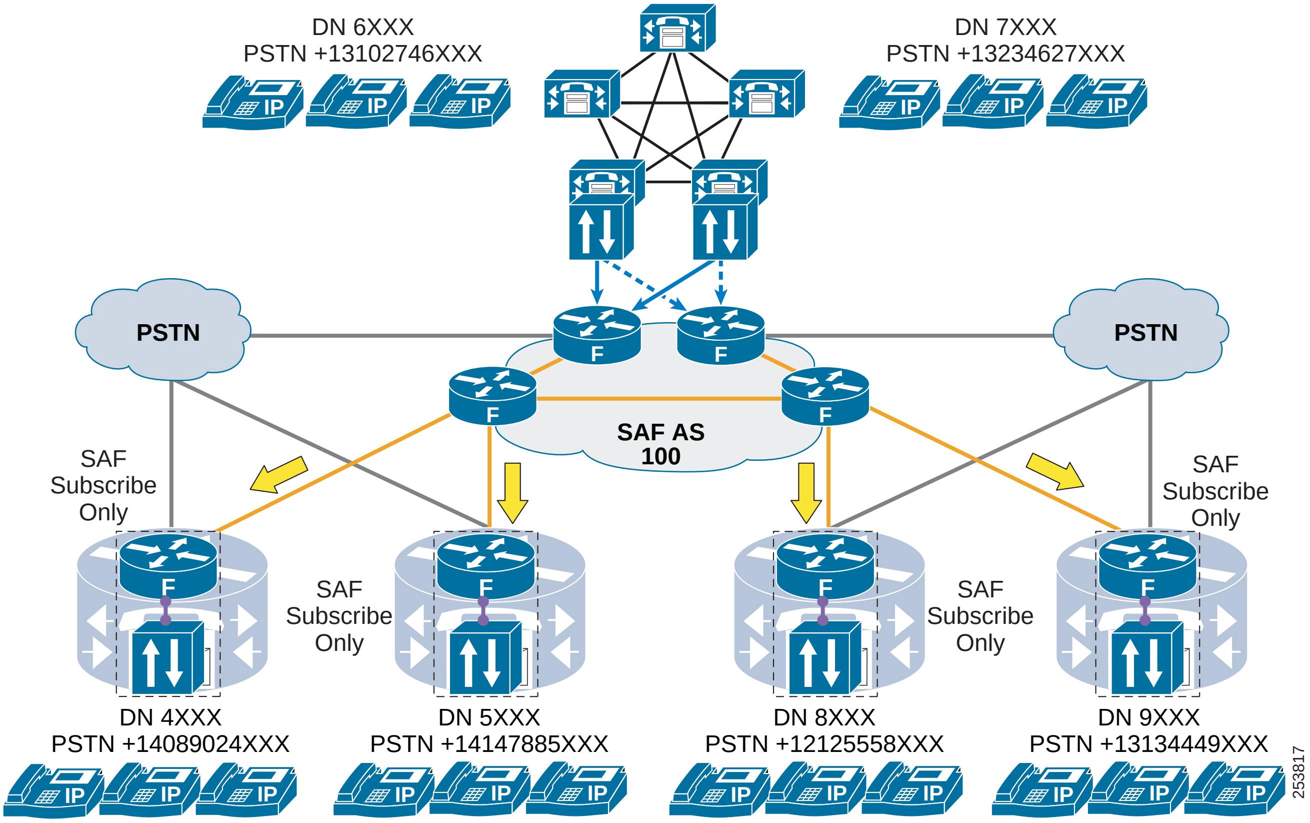
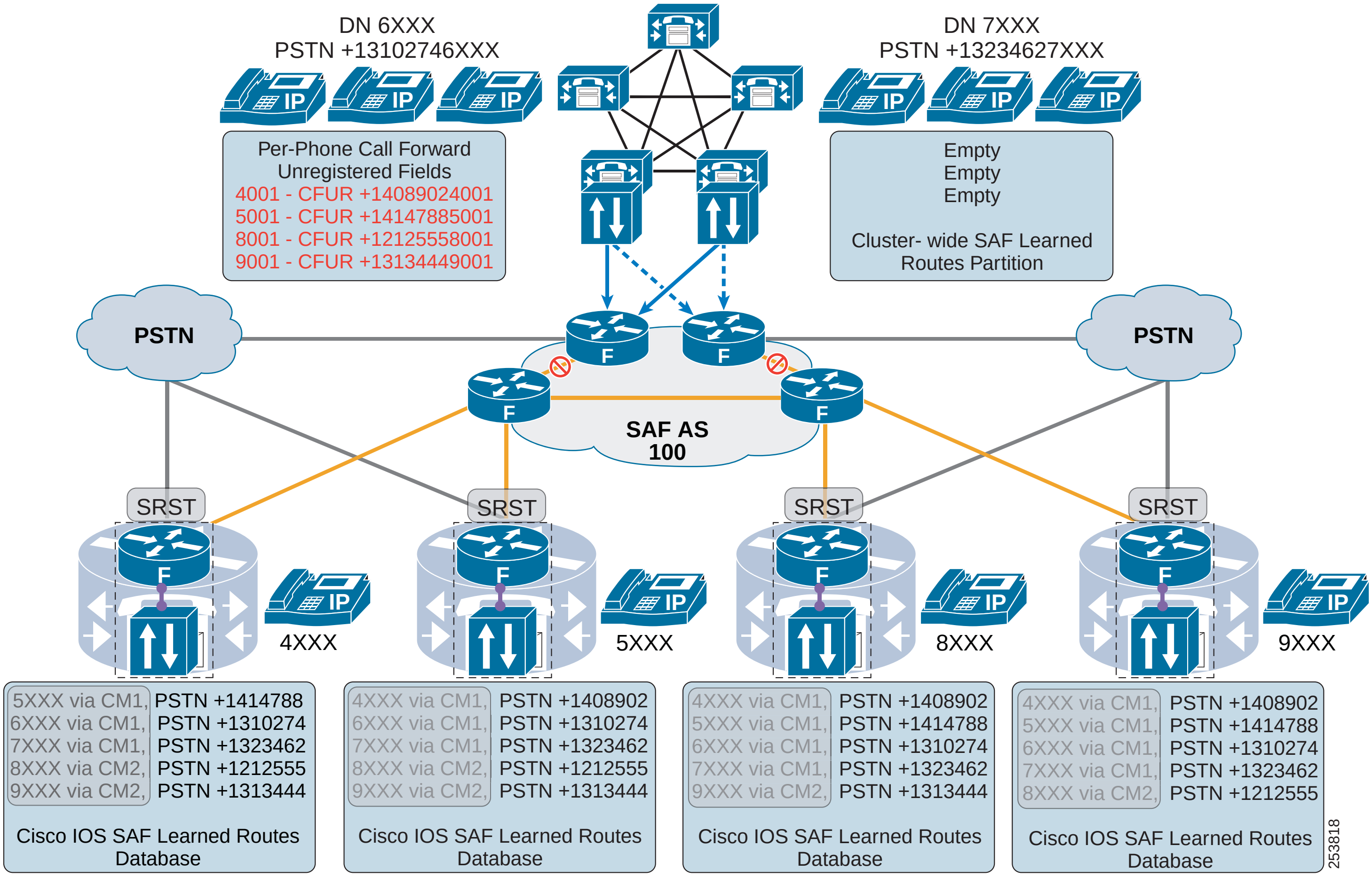


Figure 5-29 SRST Operation of a Unified CM Deployment with SAF SRST CCD


DN 6XXX
PSTN +13102746XXX
IP
Per-Phone Call Forward
Unregistered Fields
4001 - CFUR +14089024001
5001 - CFUR +14147885001
8001 - CFUR +12125558001
9001 - CFUR +13134449001
DN 7XXX
PSTN +13234627XXX
Empty
Empty
Empty
Cluster- wide SAF Learned
Routes Partition
PSTN
PSTN
F
SAF AS
100
SRST
SRST
SRST
SRST
4XXX
5XXX
8XXX
9XXX
5XXX via CM1, PSTN +1414788
6XXX via CM1, PSTN +1310274
7XXX via CM1, PSTN +1323462
8XXX via CM2, PSTN +1212555
9XXX via CM2, PSTN +1313444
Cisco IOS SAF Learned Routes
Database
4XXX via CM1, PSTN +1408902
6XXX via CM1, PSTN +1310274
7XXX via CM1, PSTN +1323462
8XXX via CM2, PSTN +1212555
9XXX via CM2, PSTN +1313444
Cisco IOS SAF Learned Routes
Database
4XXX via CM1, PSTN +1408902
5XXX via CM1, PSTN +1414788
6XXX via CM1, PSTN +1310274
7XXX via CM1, PSTN +1323462
9XXX via CM2, PSTN +1313444
Cisco IOS SAF Learned Routes
Database
4XXX via CM1, PSTN +1408902
5XXX via CM1, PSTN +1414788
6XXX via CM1, PSTN +1310274
7XXX via CM1, PSTN +1323462
8XXX via CM2, PSTN +1212555
Cisco IOS SAF Learned Routes
Database
253818

Typical SAF CCD-Based Unified Communications Deployments

Figure 5-30 show a typical SAF CCD network deployment.

Figure 5-30 A Global SAF Network with Regional Call Agents and SAF Clients and Forwarders

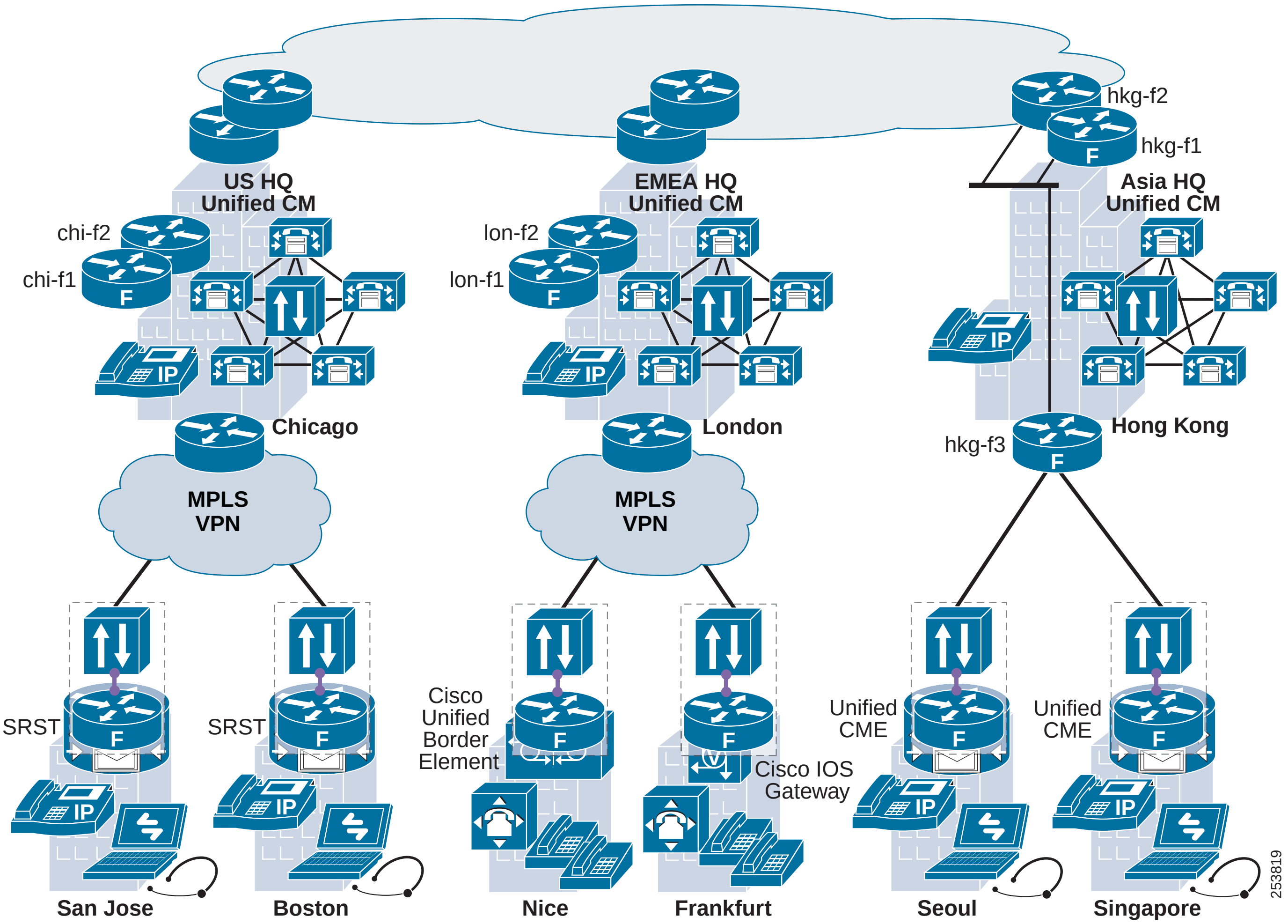
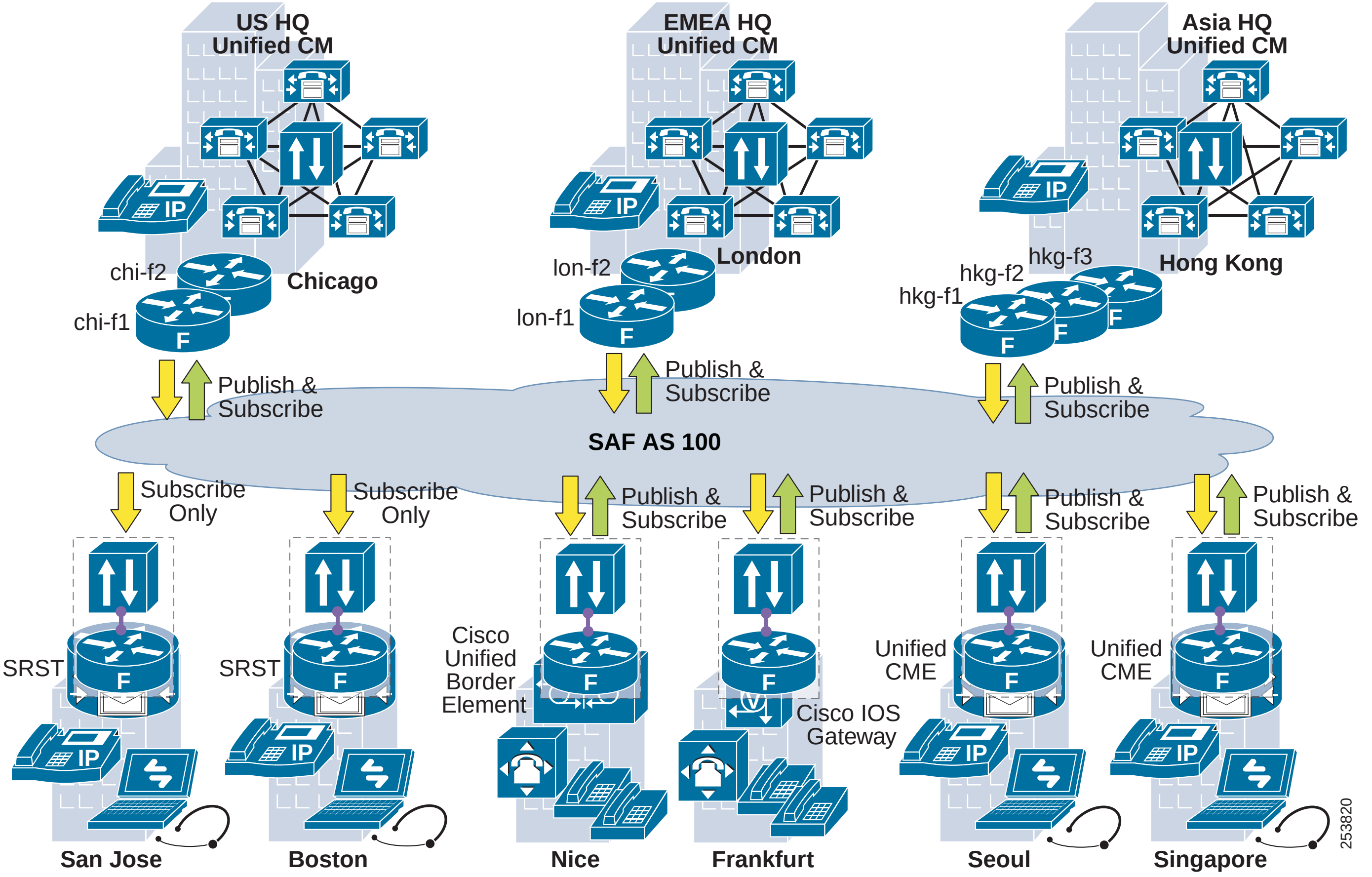


Figure 5-31 shows a logical diagram of the same global SAF network with regional call agents and SAF Clients and Forwarders

Figure 5-31 Logical Representation of Global SAF Network with Regional Call Agents and SAF Clients and Forwarders



SAF CCD Deployment Considerations

Migration to SAF CCD is relatively risk free. The SAF CCD network can be built and tested for basic operation and scalability before any devices that use SAF are enabled in the network. Unified CM users can be given the capability to use the SAF CCD network by adding the SAF Learned Routes Partition to their device or profile. In Cisco IOS the preference for SAF dial peers can be prioritized above standard dial peers. This allows SAF to be enabled incrementally throughout the network.

The following scalability limits apply to Unified CM and Cisco IOS SAF CCD products:

- Up to 2,000 advertised DN patterns per Unified CM cluster
- Up to 20,000 learned DN patterns per Unified CM cluster
- Up to 125 advertised DN patterns per Unified CME, Cisco Unified Border Element, or Cisco IOS Gateway
- Up to 6,000 learned DN patterns per Unified CME, Cisco Unified Border Element, Cisco IOS Gateway, or SRST (platform-dependant)

Note For SAF deployments using a single SAF AS and consisting of Cisco Unified CM and Cisco IOS SAF CCD systems, SAF CCD system-wide scalability is limited to 6,000 learned DN patterns.

In very large SAF CCD networks, multiple SAF ASs can be used to limit the distribution of SAF advertised DN patterns. Unified CM and/or Cisco Unified Border Element may also be used to manually summarize SAF advertisements from one SAF AS and statically advertise them into another SAF AS.

SAF CCD Port Numbers

SAF CCD uses the following port numbers:

- SAF EIGRP — IP Protocol 88
- Unified CM SAF Client to Cisco IOS SAF Forwarder — TCP port 5050 (configurable)
- Advertised SIP trunks — port 5060
- Advertised H.323 — ephemeral port number

Note The Cisco Adaptive Security Appliance (ASA) firewall uses standard SIP inspection and fix-up to open pinholes in the firewall for the RTP media streams of SAF enabled SIP trunk calls. H.323 inspection and fix-up of SAF-enabled H.323 trunk calls are not supported.

Cisco Intercompany Media Engine

Cisco Intercompany Media Engine (IME) is another variation of a multisite deployment with distributed call processing; however, with IME the sites are separate enterprise organizations. The term *boundary-less* Unified Communications is used to describe this technology because it allows for the business-to-business extension of Unified Communications capabilities such as high-fidelity codecs, enhanced caller ID, and video telephony outside the corporate networks. The solution learns routes in a dynamic, secure manner and provides for secure communications between organizations across the internet. Organizations that work closely together and have high levels of intercompany communications will benefit most from the enhanced communications offered by IME. This section discusses the components of the solution and the high-level architecture, with relevant design considerations for deploying IME.

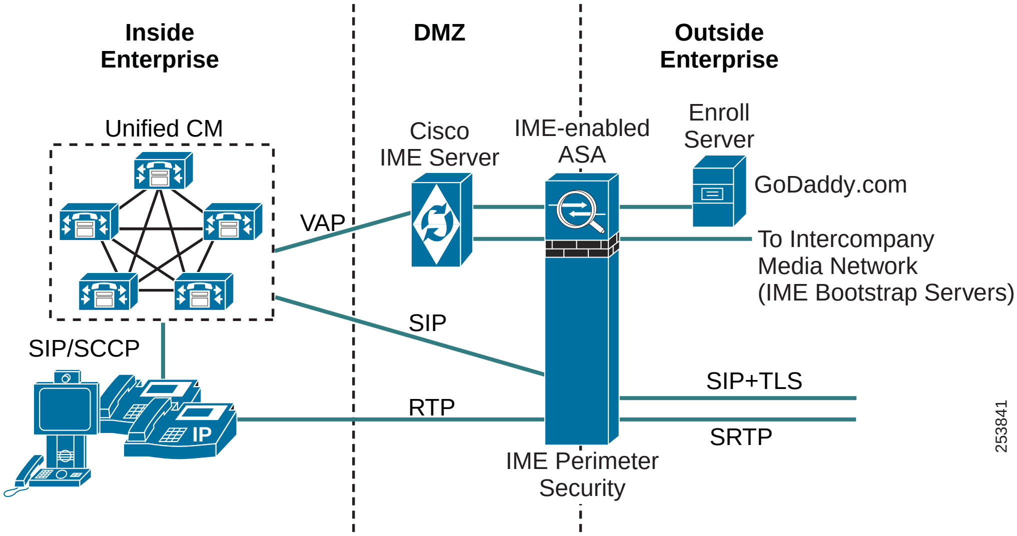
IME Components

The IME solution consists of several components to allow for the dynamic learning of IME routes and the secure encryption of call signaling and media between organizations. Two elements of the solution are hosted on the internet: the GoDaddy.com Enrollment Servers and the Intercompany Media Engine Bootstrap Servers, hosted by GoDaddy.com and Cisco, respectively. The following additional integral components are deployed on-premises:

- Cisco Intercompany Media Engine Server
- Cisco Unified Communications Manager (Unified CM)
- Cisco Adaptive Security Appliance (ASA)

Figure 5-32 illustrates a high-level view of the deployed components.

Figure 5-32 Cisco Intercompany Media Engine Components



GoDaddy.com Enrollment Server

The GoDaddy.com Enrollment Server validates all IME servers before they enter the ring of IME Servers formed over the Internet. Only IME servers installed and enrolled with the proper GoDaddy.com certificates are allowed to participate in the ring. This enrollment server is accessed only prior to entry into the ring or when certificates expire and an IME server must re-enroll.

Intercompany Media Engine Bootstrap Servers

The IME bootstrap servers are a collection of globally accessible IME servers owned and operated by Cisco. Each IME server participating in the ring (also known as the distributed cache ring) joins the network by first connecting to an IME bootstrap server. The peer-to-peer certificate obtained in the enrollment process is used for all peer-to-peer TLS connections, including the initial connection to the bootstrap server.

Intercompany Media Engine Servers

Each organization owns and operates one or more IME servers on their network. The IME server is responsible for publishing directory numbers owned by the organization to the distributed cache ring, validating call records, learning routes to remote enterprises, and pushing IME learned routes to Unified CM. It is involved only with the IME learning cycle of the solution and does not play a role in the real-time signaling or media communications.

Unified Communications Manager and Session Management Edition

Cisco Unified CM 8.*x* or Unified CM Session Management Edition 8.*x* is required for any organization to participate in IME. Unified CM communicates with IME servers to upload the IME designated directory numbers to the distributed cache ring and sends call records to IME for PSTN calls made by these directory numbers. Unified CM also receives IME learned routes that are validated by IME servers and initiates dynamic SIP trunk calls to the remote directory numbers in these IME learned routes. SIP trunk signaling always flows through an IME-enabled Adaptive Security Appliance (ASA).

Adaptive Security Appliance

All IME calls must flow through an IME-enabled Adaptive Security Appliance (ASA), which provides perimeter security for the solution. The IME-enabled ASA is responsible for receiving SIP signaling communications (outbound from Unified CM or inbound from remote enterprises), validating IME tickets, performing address translation, and providing SIP to SIP+TLS conversion for secure signaling across the Internet. Audio and video media between organizations also flow through the IME-enabled ASA, where it provides RTP-to-Secure RTP (sRTP) conversion and voice quality monitoring of the audio stream incoming from the Internet. There are off path and basic (inline) deployment options. For more information on these deployment options, see ASA Intercompany Media Engine Proxy, page 4-25.

IME Architecture

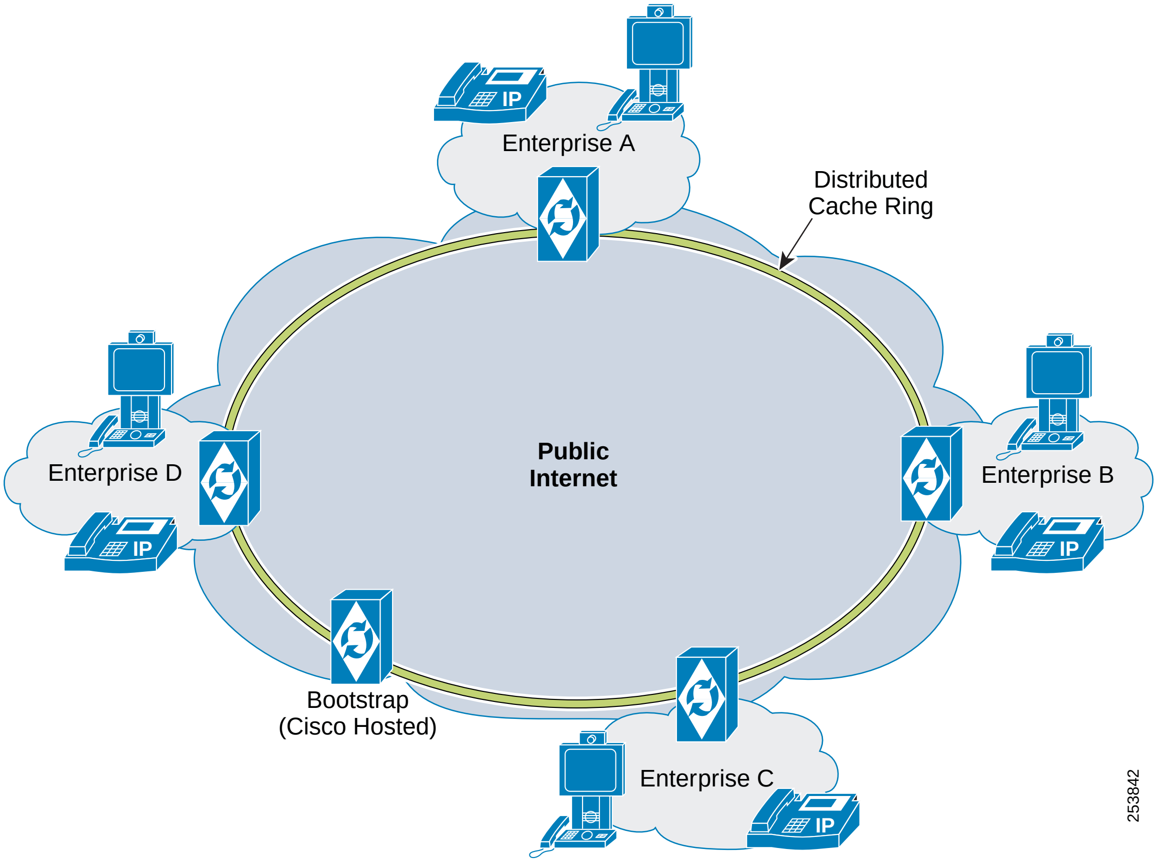
The architecture of IME is reflected in the way IME operates. The operation of IME involves the following high-level phases:

- IME Learned Routes, page 5-77
- IME Call Processing, page 5-80

IME Learned Routes

After enrolling with the GoDaddy.com Enrollment Server and being validated by the IME Bootstrap Server, an IME server becomes an active server on the peer-to-peer ring. IME servers from all organizations participating in IME join the ring on the Internet and communicate using a secure peer-to peer technology based on the Resource Location And Discovery (RELOAD) protocol. The IME servers create a distributed hash table that stores one IME-specific piece of information: a one-way hash of all published +E.164 directory numbers and the IME server peer ID that owns them. This information is distributed across all IME servers, and the architecture of the peer-to-peer technology is such that IME servers can dynamically join or leave the ring without degradation of the ring's functionality. Establishing the IME server on the ring and publishing the enterprise's IME-enrolled directory numbers is the first step toward learning IME routes. Figure 5-33 shows a logical view of the distributed cache ring (DCR).

Figure 5-33 Intercompany Media Engine Distributed Cache Ring

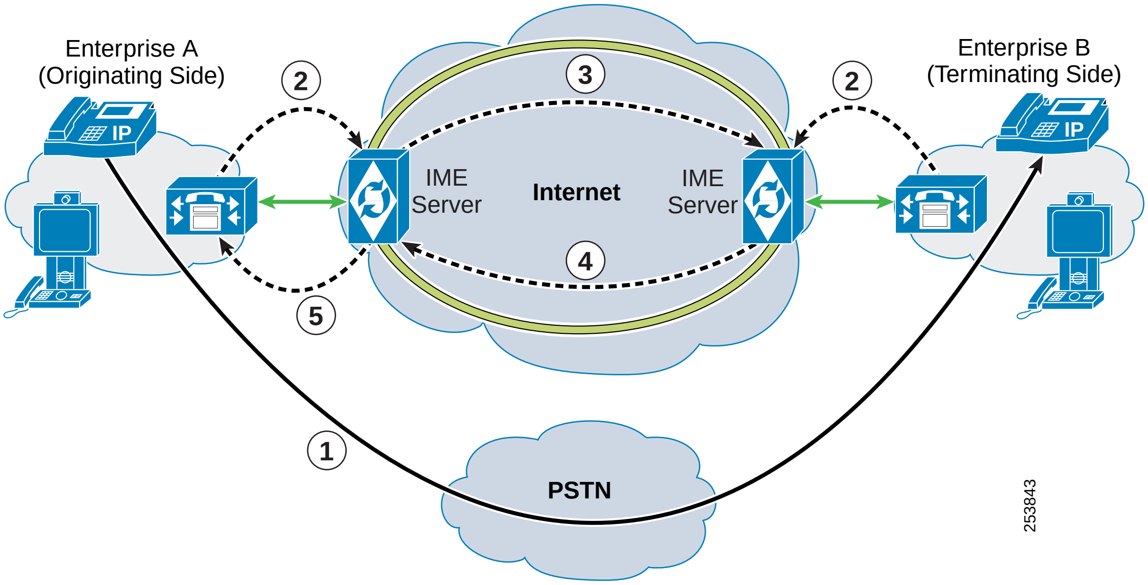


Note A draft has been submitted to the IETF governing body to standardize the IME server peer-to-peer protocol. More information is available at http://datatracker.ietf.org/doc/draft-rosenberg-dispatch-vipr-reload-usage/.

Note IME requires all directory numbers associated with the Intercompany Media Network to be in E.164 format, including the international + prefix (such as +14085551212). This is referred to as +E.164 format throughout this document.

Figure 5-34 illustrates the IME Learned Route process.

Figure 5-34 Intercompany Media Engine Learned Route Process



After the IME solution is deployed in an organization, select directory numbers can be enrolled administratively for IME, and these +E.164 numbers will be published to the distributed cache ring. The first call from an IME directory number uses the PSTN as it did before (Step 1 in Figure 5-34). Because it is an IME directory number, after the call is completed, information about that call in the form of a Voice Call Record (VCR), a CDR-like record specific to IME, is uploaded to the IME server by means of Validation Access Protocol (VAP) (Step 2 in Figure 5-34).

Voice call records contain information such as the called and calling numbers in +E.164 format and the start and stop time of the call. At some point later (it is not real time), the IME server of the enterprise that originated the call will query its peers on the DCR in an attempt to find the enterprise that owns this +E.164 called number (Step 3 in Figure 5-34). When the owner of this called number is discovered (which implies that this directory number has been enrolled in IME by another enterprise), the validation process begins. All communication between IME servers occurs over 128-bit AES TLS. The terminating-side IME server verifies that the called/calling number and start/stop times of the originating IME server's VCR match a corresponding VCR on the terminating side. If verified, the terminating IME server sends a successful reply to the originating IME server that includes a “ticket” (a security hash that only the terminating-side ASA can decipher, as described in IME Call Processing, page 5-80) and the external IP address to which IME SIP trunk calls should be directed for this +E.164 number (Step 4 in Figure 5-34). This constitutes an IME learned route. The originating IME server receives this learned route and at some point later publishes it to Unified CM by means of VAP (Step 5 in Figure 5-34). When Unified CM receives this IME learned route, it inserts the route into the Unified CM database. At this point, when any IME-enabled directory number in the originating enterprise makes a call to a number in the IME learned routes list, it will be an IME call. Note there are no real-time communications involved in learning IME routes. For a detailed example of learned routes, refer to the *Cisco Intercompany Media Engine Installation and Configuration Guide*, available at

http://www.cisco.com/en/US/products/ps10669/prod_maintenance_guides_list.html

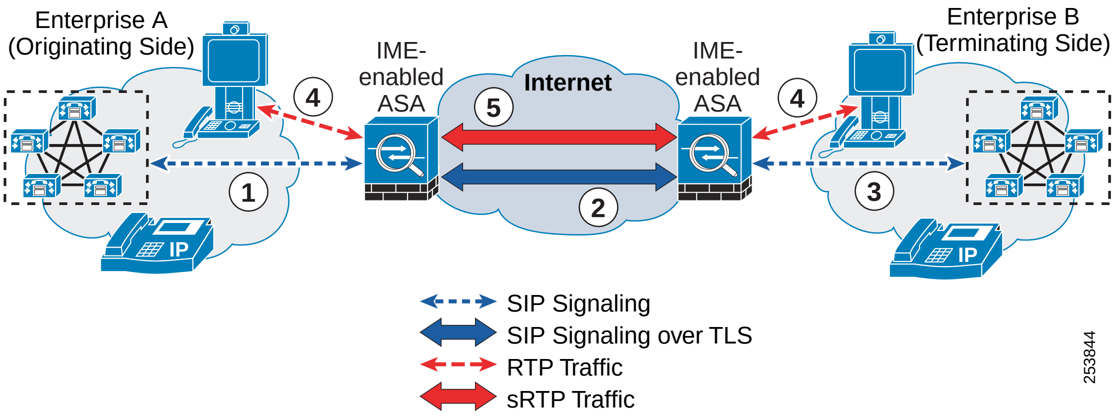
Note Unified CM provides a facility for blocking IME learned routes insertion into the Unified CM database for defined prefixes or domains.

Note Unified CM has a method of transforming called and calling numbers to a globalized +E.164 format specifically for IME VCRs, even if a globalized dial plan is not implemented. For more information on +E.164 transformations for VCRs, see Dial Plan Considerations for the Intercompany Media Engine, page 9-33.

IME Call Processing

Once an IME learned route exists in the Unified CM database, the information in the route is used to set up an IME call. However, the IME server itself is no longer involved in the call processing phase. Figure 5-35 illustrates a high-level view of IME call processing.

Figure 5-35 Intercompany Media Engine Call Processing



Note IME also supports the use of secure SIP signaling via TLS between IME-enabled ASA and Unified CM.

To initiate an IME call, the called number must match an IME learned route pattern in the database and the directory number of the calling endpoint must be enrolled in IME. If these criteria are met, Unified CM dynamically invokes an IME SIP trunk addressed to the external IP address or fully qualified domain name (FQDN) of the terminating enterprise that was included in the IME learned route. IME learned route patterns are in +E.164 format; however, even if a globalized Unified CM dial plan is not deployed, the same process available for converting called numbers to +E.164 format for IME-specific VCRs is used to analyze outgoing called numbers on gateways and to convert them to +E.164 format for IME-specific analysis only. For more information on E.164 transformation profiles, see Dial Plan Considerations for the Intercompany Media Engine, page 9-33.

An IME-enabled ASA serves as a proxy for all IME communications with remote organizations. The ASA provides network address translation (NAT) and SIP application layer gateway (ALG) functionality to translate addressing inside the SIP messaging itself. There are two deployment options for the IME-enabled ASA: basic (inline) or offpath. Offpath is the recommended method because it provides the capability for Unified CM to direct IME traffic to an IME-enabled ASA in a DMZ. This allows use of an existing ASA already deployed in the network that all Unified CM internet-bound traffic would otherwise travel through. For more information about basic and offpath ASA deployments, see ASA Intercompany Media Engine Proxy, page 4-25.

The originating Unified CM initiates a SIP Invite that reaches the IME-enabled ASA (Step 1 in Figure 5-35) and that includes the security hash ticket from the learned route as an attribute in the SIP header. The ASA will fix-up the packets at the SIP level so that its externally facing IP address appears as the source of the Invite and extends it to the external IP address of the remote enterprise over a secure (256 bit AES) TLS connection (Step 2 in Figure 5-35). The external IP address listed in an IME learned route correlates with the outside address of the IME-enabled ASA that receives inbound SIP signaling on behalf of a Unified CM cluster. The terminating ASA receives the SIP Invite, decrypts it, and validates the ticket. Any request without a valid ticket is blocked. After the ticket as been verified, the ASA performs NAT and ALG functions before forwarding the ticket on to the terminating Unified CM (Step 3 in Figure 5-35).

Note The IME Server and IME-enabled ASA do not have direct communications; however, they are both configured with an identical **epoch ticketpassword**, which allows for successful ticket validations.

Once successful SIP signaling has been negotiated, each IME-enabled ASA instructs its respective Unified CMs to have the endpoints stream RTP media directly to its internal media termination address (Step 4 in Figure 5-35). The ASAs take in this RTP stream, encrypt it, perform NAT, and send it across to the remote ASA as sRTP sourced from its external media termination address, including audio and video media (Step 5 in Figure 5-35). At this point, the two endpoints have an active IME call.

The IME solution also provides a mechanism to allow calls to fall-back to the PSTN if the voice quality of the audio stream degrades below an acceptable level. Advanced features such as video are lost, but the audio portion of the call remains intact and the change is otherwise unapparent to the user.

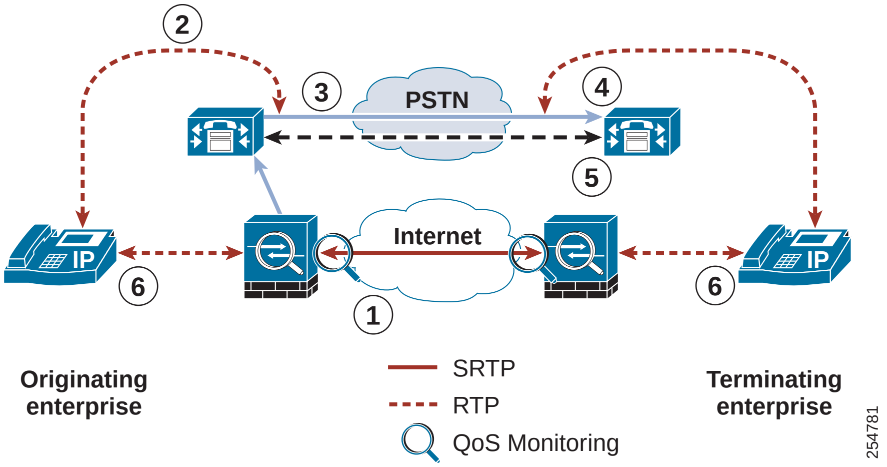
For more details regarding the IME fallback feature and IME-enabled ASA configuration, refer to the information on configuring Cisco Intercompany Media Engine Proxy in the *Cisco ASA 5500 Series Configuration Guide using the CLI, 8.3*, available at

http://www.cisco.com/en/US/docs/security/asa/asa83/configuration/guide/config.html

PSTN Failover

Cisco IME uses the public internet to carry business-to-business traffic. To avoid negative user experience caused by possible packet loss in the public network, the IME-enabled ASA involved in the call continuously monitors the inbound SRTP stream and calculates RTP statistics in real-time. The IME-enabled ASA looks for random packet loss, small bursts of lost packets, and bursts of large packet loss. All three packet impairment conditions are used to decide whether the call meets the defined minimum quality limits. If the measured quality as defined by the calculated real-time RTP statistics falls below certain pre-defined quality limits, the IME-enabled ASA initiates PSTN fallback for the affected call as shown in Figure 5-36. The QoS management algorithm maintains five sensitivity levels with varying packet impairment thresholds to allow granular control of the PSTN fallback sensitivity. The Fallback QoS Sensitivity Level can be set on a global basis or per IME Enrolled Group.

Figure 5-36 IME PSTN Failover



After an IME call is established, the ASA inspects RTP packets as they flow through it from outside to inside. The ASA inspects the sequence numbers and timestamps, and based on the observed packet loss, an algorithm decides whether a fallback is required. The algorithm uses the fallback QoS sensitivity levels to create a set of packet loss thresholds for each QoS sensitivity level. If the algorithm indicates a fallback is needed (step 1 in Figure 5-36), the ASA sends an out-of-dialog REFER to Cisco Unified Communications Manger, asking for it to fallback to the PSTN (step 2 in Figure 5-36).

As the terminating Unified CM receives the REFER, it issues a mid-dialog REFER to the originating Unified CM over the existing dialog. This REFER is required to inform the originating Unified CM of the required fallback. The PSTN call required for the PSTN fallback is always initiated by the Unified CM originating the IME call, regardless of whether the IME-enabled ASA on the originating or terminating side triggers the fallback

The originating Unified CM then places a PSTN call to the Fallback Directory E.164 Number advertised by the terminating Unified CM as part of the SIP call setup (step 3 in Figure 5-36). The Fallback Directory E.164 Number can be configured both in the global Fallback Feature Configuration Settings and in the Fallback Profile Configuration Settings, thus allowing different Fallback E.164 Numbers per IME Enrolled Group.

The call to the Fallback Directory E.164 Number by default is routed using the calling device´s AAR calling search space. The global Fallback Feature Configuration Settings and the Fallback Profile Configuration Settings also allow you to use the calling device´s Reroute Calling Search Space.

When a PSTN call is routed to a configured Fallback Directory E.164 Number, Unified CM has to associate the incoming call with the correct IME call. The first step is to match the caller ID of this PSTN call against the caller ID signalling from the originating Unified CM in the SIP INVITE of the initial VoIP call (step 4 in Figure 5-36). If sufficient digits (as defined by the Number of Digits for Caller ID Partial Match in the global Fallback Feature Configuration Settings or in the Fallback Profile Configuration Settings) are matched, the Unified CM terminating the PSTN fallback call informs the Unified CM originating the PSTN call by sending a single DTMF digit 1. The originating Unified CM immediately splits the VoIP call, connects the PSTN leg with the phone, and terminates the VoIP leg. If the caller IDs do not match, the terminating Unified CM sends a single DTMF digit 2 (step 5 in Figure 5-36).

The originating Unified CM waits for DTMF digits. If a 1 is received, the originating Unified CM immediately splits the VoIP call, connects the PSTN leg with the phone, and terminates the VoIP leg (step 6 in Figure 5-36).

If a 2 is received, indicating that the terminating Unified CM was not able to associate the PSTN fallback call with a unique existing IME VoIP call, the originating Unified CM out-pulses a DTMF sequence uniquely identifying the call. This DTMF sequence is learned from the terminating Unified CM as part of the SIP exchange during the initial IME VoIP call establishment. After sending the DTMF sequence, the originating Unified CM splits the VoIP call, connects the PSTN leg with the phone, and terminates the VoIP leg (step 6, in Figure 5-36).

Unified CM expects to receive the DTMF digits received as part of the PSTN fallback procedure to be delivered out-of-band.

Capacity Planning

IME servers are sized according to how many enrolled DIDs will be published on them. Table 5-5 provides the current supported capacity limits per platform.

Table 5-5 IME Server Supported Capacities

Platform	Maximum Number of Enrolled DIDs
Cisco MCS 7825-H2/I2 and 7825-H4/I4	20,000
Cisco MCS 7845-H2/I2 and 7845-I3	40,000

Because all IME call media (audio and video) flow through the IME-enabled ASA, capacity depends on the type and number of calls flowing through it. The IME-enabled ASA monitors only the audio stream incoming from the internet for voice quality. The video media is not monitored for voice quality, but it does flow through the IME-enabled ASA for RTP-to-sRTP conversion, and the bandwidth of the video directly affects the number of sessions each can handle. Table 5-6 provides capacity limits for the ASA-5550 and ASA-5580. Performance limits of other ASA models have not been validated yet.

Table 5-6 Maximum Number of Calls per Type and ASA Model

ASA Model	Voice G.711	Video 300 kbps	Video 800 kbps	Video 1 Mbps
ASA-5550 4 GB	480 calls	240 calls	120 calls	80 calls
ASA-5580-20 4 GB	900 calls	600 calls	300 calls	200 calls

Unified CM does not have a limit on the number of IME calls it can handle, but IME calls should be factored into the overall call capacity provided by the cluster. Your Cisco Partner or Cisco Systems Engineer should use the Cisco Unified Communications Sizing Tool (http://tools.cisco.com/cucst) to validate all designs that handle large call traffic volumes. The Sizing Tool can accurately determine the number of servers or clusters required to meet your design criteria.

High Availability

The IME route learning phase involves several aspects of high availability. The distributed cache ring (DCR) itself has a high degree of redundancy built into the peer-to-peer technology, where the information stored on the DCR peers is adjusted as IME servers join or leave the ring. Multiple IME bootstrap servers are also hosted by Cisco to guarantee that valid IME servers can join the ring at any time. These aspects are inherent to the solution.

In Unified CM, each IME Service (which defines a set of enrolled DIDs, excluded DIDs, and IME servers, among other parameters) can consist of a primary and secondary IME Server. Both servers are up and active, and Unified CM uploads the enrolled DIDs and any terminating call VCRs to both servers. However, originating call VCRs are uploaded to the primary IME server only; therefore, only the primary will initiate validation requests, while either can process validation requests received from other enterprises because both have terminating call VCRs. There is no direct communication between primary and secondary IME servers regarding VCRs, so originating call VCRs stored for validation on the primary would be lost in the event of an outage. A recommended option is to split the enrolled DIDs into two ranges and to create two IME Services whereby the primary IME server for service A is the secondary IME server for service B, and vice versa. This balances the originating call validation load across the IME Servers and further minimizes the number of originating VCRs that are lost in the event of an outage.

On the Unified CM side, once an IME Service is configured, the Unified CM responsible for initiating VAP communications with the primary IME server is determined by the device pool of the IME SIP trunk associated with the IME Service. The Unified CM Group attribute associated with the device pool determines the primary, secondary, and tertiary Unified CM responsible for the service. In the event that the primary is down, the secondary Unified CM picks up VAP communications with the active IME server.

With respect to call processing, the Unified CM Group associated with the IME SIP trunk in the IME Service also determines which Unified CM subscribers initiate IME calls. This allows for IME calls to continue in the event that the primary Unified CM for the IME SIP trunk is offline. For receiving calls, each IME Service can configure external IP address and port pairs for Unified CM call processing subscribers in the cluster. Each external IP address and port pair is actually an IP address and port that is configured on the IME-enabled ASA and that has a 1:1 correlation to a Unified CM call processing node. When there are multiple external IP addresses and ports in an IME route, Unified CM sends calls for this IME route in a circular fashion so that calls are load-balanced across Unified CM servers at the remote enterprise. If a remote Unified CM is offline, the originating Unified CM tries the next external IP address and port in the list. If no response is received and this list is exhausted, the call is sent to the PSTN as it would have been without IME.

Dual IME-enabled ASAs may also be deployed in active-standby mode; however, this does not provide stateful failover. In the event of a failover, active calls are disconnected, but subsequent IME calls connect through the standby (now active) ASA. For deployments with an offpath IME-enabled ASA, the IME Service configuration in Unified CM allows for a single IME firewall to be associated. Multiple IME-enabled ASAs can be deployed to handle IME calls for different enrolled DID ranges, thus offering a mechanism of load balancing IME calls in addition to increasing capacity.

Note Active-active failover mode is not supported for the IME-enabled ASA.

While an IME call is connected, the IME-enabled ASA is capable of monitoring the quality of the call. If the quality falls below a certain sensitivity level, the call is moved back to the PSTN. For more information, see ASA Intercompany Media Engine Proxy, page 4-25.

Design Considerations

The IME solution requires that IME servers and the IME-enabled ASA have publicly reachable IP addressing; therefore, they are most commonly placed in an organization's DMZ. This may require close coordination between groups responsible for security and Unified Communications within the organization. It is important for both the security and Unified Communications teams to be involved from the early design stages of an IME project. In addition, observe the following guidelines and considerations when designing an IME solution for your enterprise:

- Cisco requires the use of Network Time Protocol for all Unified CM servers, IME servers, and IME-enabled ASAs. They must be synchronized to a dependable, high-stratum clock source. It is vital to Voice Call Record start and stop times during the IME route learning phase.
- A hosted IME solution deployment model is also supported. In a hosted IME deployment, an IME server publishes enrolled directory numbers and validates VCRs on behalf of multiple Unified CM or Unified CM Session Management Edition clusters. For more information, refer to the information on hosted IME solutions in the *Cisco Intercompany Media Engine Installation and Configuration Guide*, available at

  http://www.cisco.com/en/US/products/ps10669/prod_maintenance_guides_list.html

IME Servers

- By default, VAP communications between Unified CM and the IME server are authenticated only. When an IME server is located in the DMZ, Cisco recommends configuring VAP communications as authenticated and encrypted, which will force the communications to occur over TLS. This requires additional configuration to share security certificates.

Unified CM and Unified CM Session Management Edition

- The Intercompany Media Network requires all published numbers to be in +E.164 format to ensure their global uniqueness. Calling and called numbers must be converted to +E.164 format so that, when IME-specific Voice Call Records (VCRs) are uploaded to IME servers, they are in the proper format. Unified CM provides a facility to transform calling and called numbers to +E.164 format solely for IME purposes, which will not affect normal dial plan digit analysis. For more information, refer to the E.164 transformation profile information in the *Cisco Intercompany Media Engine Installation and Configuration Guide*, available at

  http://www.cisco.com/en/US/products/ps10669/prod_maintenance_guides_list.html

- Gateways or trunks used for PSTN connectivity must have the PSTN Access checkbox checked in order for calling and called numbers to be analyzed for IME participation. Upon upgrade to Unified CM 8.*x*, this parameter is enabled by default for all gateways and trunks. You can unchecked it if it is not required.
- Configuration settings in Unified CM for the regions between internal endpoints and the IME SIP trunk determine the audio and video capabilities allowed for IME calls.
- To limit capacity through the IME-enabled ASAs, Cisco recommends applying Unified CM locations-based call admission control to the IME SIP trunk to control the number of audio and video calls sent through the ASA. When the bandwidth limits are reached, subsequent calls will be routed through the PSTN as they were prior to IME deployment.
- Cisco recommends explicitly trusting the domains of remote enterprises with which your organization intends to communicate through IME. Once a trust group is configured, there is a default deny on all other domains that try to validate VCRs.
- Unicast music on hold (MoH) is supported during user-initiated hold and transfer scenarios. To work properly through the firewalls, the MoH full-duplex streaming service parameter must be enabled.
- Cisco recommends excluding analog and fax station directory numbers from the enrolled group of DIDs for an IME server because they will not benefit from enhanced Unified Communications and because fax calls are not supported over IME.

IME-Enabled ASA

- For more information on basic and offpath ASA deployments as well as security considerations for other firewalls in the network, see ASA Intercompany Media Engine Proxy, page 4-25.
- High-bandwidth video (greater than 384 kbps) is supported; however, it directly affects the capacity of calls flowing through the IME-enabled ASA.
- Fallback sensitivity levels should be left at the default settings for the initial IME deployment. Fallback should be monitored during the first few months of use and then adjusted accordingly. Cisco recommends viewing call detail records to find calls generated on behalf of IME or fallback. Appropriate fallback sensitivity levels will vary from enterprise to enterprise.
- When endpoints with IME-enrolled DIDs are remotely located with VPN connectivity into the enterprise, latency and jitter characteristics for calls with these endpoints will be amplified and could result in the IME-enabled ASAs triggering more frequent fallbacks to the PSTN. If fallbacks occur too frequently for a specific endpoint, it might be necessary either to configure these devices with a device pool that has a fallback profile with no fallback enabled, to lower the fallback sensitivity levels, or to remove the enrolled DID from IME.

CHAPTER 6

IP Telephony Migration Options

Revised: June 28, 2012; OL-27282-05

This chapter discusses how the individual Cisco Unified Communications family of applications can be migrated either from earlier releases or from existing third-party deployments. This chapter also describes several methods for migrating from separate standalone communication components to an integrated Cisco Unified Communications System. The topics discussed in this chapter are viewed from a customer or business viewpoint rather than a technical viewpoint that is based around which protocol to use or which features are required.

When considering any form of migration and/or upgrade, customers should first consult the *Cisco Unified Communications Compatibility Tool* (at http://tools.cisco.com/ITDIT/vtgsca/VTGServlet) to ensure a successful outcome. For example, some applications may start from such an early software release that a multi-step upgrade might be necessary. Similarly, server hardware along with software compatibility might require a combination of multi-step hardware and software upgrades.

Strong consideration should also be given to other Unified Communications applications because the system currently deployed might have multiple applications that have limited compatibility of interworking with each other.

What's New in This Chapter

Table 6-1 lists the topics that are new in this chapter or that have changed significantly from previous releases of this document.

Table 6-1 New or Changed Information Since the Previous Release of This Document

New or Revised Topic	Described in:	Revision Date
Migration from physical to virtual machines	Migrating from Physical Servers to Virtual Machines, page 6-5	June 28, 2012
Migration of Cisco IM and Presence	Cisco IM and Presence Migration, page 6-6	June 28, 2012
License migration and Cisco Enterprise License Manager (ELM)	Migrating Licenses to the Cisco Enterprise License Manager (ELM), page 6-7	June 28, 2012

Coexistence or Migration?

This is an important question that needs to be answered.

Coexistence typically means two or more systems coexisting for an extended period of time (for example, anything greater than six months). Under this scenario feature transparency, whether for PBX, voicemail, or other features, becomes a more significant consideration. Investment and/or upgrades to existing systems might be necessary in order to deliver the level of feature transparency required.

Migration typically occurs over a shorter period of time (for example, less than six months). Under this scenario users are more likely to tolerate a subset of existing features, knowing that the migration will be complete in a "short" period of time. Often existing system capabilities may be sufficient for this "short" period of time, therefore migration is often less costly when compared to coexistence.

Migration Prerequisites

Before implementing any Unified Communications service, customers should ensure that the underlying IP infrastructure is "UC ready," including redundancy, high availability, Quality of Service (QoS), in-line powered Ethernet ports, and so forth. For further details, refer to the chapter on Network Infrastructure, page 3-1.

Typically some kind of site or user-survey should be performed to ensure that all requirements (for example, fax/modems, environmental control systems, and so forth) are appropriately identified and accounted for.

Unified Communications Migration

There are two main methods for migrating to a Unified Communications system (or any individual Unified Communications service, for that matter):

Phased Migration

This method typically starts with a small trial focused around the Unified Communications service to be deployed. Once the customer is familiar with the Unified Communications service trial, then the migration starts by moving groups of users, one group at a time, to the production version of that Unified Communications service.

Parallel Cutover

This method begins similar to the phased approach; however, once the customer is satisfied with the progress of the trial, then a time and date are chosen for cutting-over all the users at once to the new Unified Communications service.

A parallel cutover has the following advantages over a phased migration:

- If something unexpected occurs, the parallel cutover provides a back-out plan that allows you to revert, with minimal effort, to the previous system, which is essentially still intact. For example, with phased migration from a PBX, service can be restored to the users simply by transferring the inbound PSTN trunks from the IP telephony gateway(s) back to the PBX.
- The parallel cutover allows for verification of the configuration of the Unified Communications service before the system carries live traffic. This scenario can be run for any length of time prior to the cutover of the Unified Communications service, thereby ensuring correct configuration of all user information such as phones, gateways, the dial plan, mailboxes, and so forth.

- Training can be carried out at a more relaxed pace by allowing subscribers to explore and use the Unified Communications service at their own leisure prior to the cutover.
- The system administrator does not have to make special provisions for "communities of interest." With a phased approach, you have to consider maintaining the integrity of features such as call pick-up groups, hunt groups, shared lines, and so forth. These associations can be easily accounted for when moving the complete Unified Communications service in a parallel cutover.

One disadvantage of the parallel cutover is that it requires the Unified Communications service, including the supporting infrastructure, to be fully funded from the beginning because the entire service must be deployed prior to bringing it into service. With a phased migration, on the other hand, you can purchase individual components of the system as and when they are needed, and this approach does not prevent you from starting with a small trial system prior to moving to full deployment.

Neither method is right or wrong, and both depend upon individual customer circumstances and preferences to determine which option is most suitable.

Example 6-1 Phased Migration for IP Telephony

This approach typically entails a small IP telephony trial that is connected to the main corporate PBX. The choice of which signaling protocol to use is determined by the required features and functionality as well as by the cost of implementation. Cisco Unified Communications Manager (Unified CM) can support either regular PSTN-type PRI or QSIG PRI as well as H.323 and SIP. Of these options, QSIG PRI typically provides the highest level of feature transparency between any two systems.

PSTN-type PRI provides for basic call connectivity as well as Automatic Number Identification (ANI). In some instances, the protocol also supports calling name information. This level of connectivity is available to all PBXs and therefore is considered to be the least costly option; that is, if the PBX can connect to the public network through PRI, then it can connect to Unified CM because Unified CM can be configured as the "network" side of the connection.

With either PSTN-type PRI or QSIG, the process for a phased migration is similar: move users from the PBX to Unified CM in groups, one group at a time, until the migration is complete.

The Cisco San Jose campus, consisting of some 23,000 users housed in approximately 60 buildings, was migrated to IP telephony in this manner and took just over one year from start to finish at the rate of one building per weekend. All users in the selected building were identified, and their extensions were deleted from the PBX on a Friday evening. At the same time, additions were made to the PBX routing tables so that anyone dialing those extension numbers would then be routed over the correct PRI trunk for delivery to Unified CM. During the weekend, new extensions were created in Unified CM for the users, and new IP phones were delivered to their appropriate office locations, ready for use by Monday morning. This process was repeated for each building until all users had been migrated.

Example 6-2 Parallel Cutover for IP Telephony

All IP phones and gateways are fully configured and deployed so that users have two phones on their desk simultaneously, an IP phone as well as a PBX phone. This approach provides the opportunity not only to test the system but also to familiarize users with their new IP phones. Outbound-only trunks can also be connected to the IP telephony system, giving users the opportunity to use their new IP phones to place external as well as internal calls.

Once the IP telephony system is fully deployed, you can select a time and date for bringing the new system into full service by transferring the inbound PSTN trunks from the PBX to the IP telephony gateways. You can also leave the PBX in place until such time as you are confident in the operation of the IP telephony system, at which point the PBX can then be decommissioned.

The Cisco San Jose campus voicemail service was provided by four Octel 350 systems serving some 23,000 users. Cisco Unity servers were installed and users' mailboxes were configured. Users had access to the their Unity mailbox by dialing the new access number, in order to allow them to record their name and greeting(s) as well as to allow them to familiarize themselves with the new Telephony User Interface (TUI). Approximately two weeks later, a Unified CM Bulk Administration Tool (BAT) update was carried out on a Friday evening to change the Call-Forward Busy and No-Answer (CFB/CFNA) numbers as well as the Messages button destination number for all users to the Unity system. Upon returning to work on Monday morning, users were serviced by Unity. The Octel 350 systems were left in place for one month to allow users to respond to any messages residing on those systems before they were decommissioned.

The Need for QSIG in Multisite Enterprises

While some enterprises consist of only one location, others consist of many sites, some of which may potentially be spread over large distances. PBX networks for multisite enterprises are usually connected using T1 or E1 PRI trunks (depending on location) running a proprietary protocol such as Avaya DCS, Nortel MCDN, Siemens CorNet, NEC CCIS, Fujitsu FIPN, or Alcatel ABC, among others. These proprietary networking protocols enable the PBXs to deliver a high level of feature transparency between end users.

QSIG was developed to enable the interconnection of PBXs from different vendors, thereby allowing similar levels of feature transparency.

By supporting QSIG, Unified CM can be introduced into a large enterprise network while also maintaining feature transparency between users. PBX locations can then be converted to IP telephony whenever convenient.

However, unless you already have QSIG enabled on your PBX or have a specific need for its additional features and functionality, the cost of upgrading the PBX might be hard to justify if it will be retired within a short period of time. For example, why spend $30,000 on enabling the PBX for QSIG if you plan to retire the PBX in two or three months?

Summary of IP Telephony Migration

Although both methods of IP telephony migration work well and neither method is right or wrong, the parallel cutover method usually works best in most cases. In addition, large enterprises can improve upon either migration method by using QSIG to enable Unified CM to become part of the enterprise network.

Cisco has a lab facility dedicated to testing interoperability between Unified CM and PBX systems. The results of that testing are made available as application notes, which are posted at

http://www.cisco.com/go/interoperability

The application notes are updated frequently, and new documents are continuously added to this website. Check the website often to obtain the latest information.

Centralized Unified Communications Deployment

In the case of an enterprise that has chosen to deploy Unified Communications in a centralized manner, two options exist:

- Start from the outside and work inward toward the central site (that is, smallest to largest).
- Start from the central site and work outward toward the edges.

The majority of customers choose the first option because it has the following advantages:

- It gives them the opportunity to fully deploy all the Unified Communications services and then conduct a small trial prior to rolling Unified Communications out to the remote locations.
- The rollout of Unified Communications can be done one location at a time, and subsequent locations can be migrated when convenient.
- This option is the lowest cost to implement once the core Unified Communications services are deployed at the central site.
- IT staff will gain valuable experience during migration of the smaller sites prior to migrating the central site.

The remote sites should be migrated by the parallel approach, whereas the central site can be migrated using either the parallel or phased approach.

Which Unified Communications Service First?

This choice is very much dependent on the customer's individual business needs, and the Cisco Unified Communications solution allows for most of its individual services to be deployed independently of the others; for example, IP telephony, voice messaging, contact center, and collaboration can all be deployed independently from each others.

This capability provides the customer with great flexibility. Consider a customer who is faced with a voicemail system that has since gone end-of-support and is suffering various issues leading to customer dissatisfaction. Cisco Unity can often be deployed and integrated with the current PBX, thereby solving this issue. Once the new voicemail system is operating appropriately, then attention can turn to the next Unified Communications service, namely IP telephony.

Migrating from Physical Servers to Virtual Machines

This type of migration refers to migrating from a system with Cisco Unified Communications Manager (Unified CM) deployed on a physical cluster of Cisco Media Convergence Servers (MCS) to a system with Unified CM deployed on Cisco Unified Computing System (UCS) virtual machines. The simplest way to perform this type of migration is through the disaster recovery service (DRS) method; however, some customer environments might require another approach involving what is termed a "dead net." A dead net allows the new virtual machine system to be set up and tested in an isolated environment prior to bringing it into service. This means that the user data obtained from the initial backup process is considered as frozen, so any changes that have occurred prior to the backup will need to be re-entered. On the other hand, the DRS/server replacement method is performed on a cluster that is still operational and therefore carries more risk.

The DRS backup-and-restore server replacement method is carried out by performing a backup operation on each physical server in the Unified CM cluster and then restoring the backup to the corresponding virtual machine. This is a serial process, and depending upon the number of servers involved and the available outage windows, it may take some time to complete. This approach is preferred over simply doing a re-import of the database through the Cisco Unified CM Bulk Administration Tool (BAT) and has the advantage that it fully captures the state of the cluster (for example, custom music-on-hold files, non-default TFTP firmware files, states of various Unified CM services, and so forth) at the time the backup is taken.

There are currently well defined processes for moving existing MCS severs to virtual machines, described as server replacement and IP readdressing. For details, refer to the latest version of *Replacing a Single Server or Cluster for Cisco Unified Communications Manager*, available at

http://www.cisco.com/en/US/products/sw/voicesw/ps556/prod_installation_guides_list.html

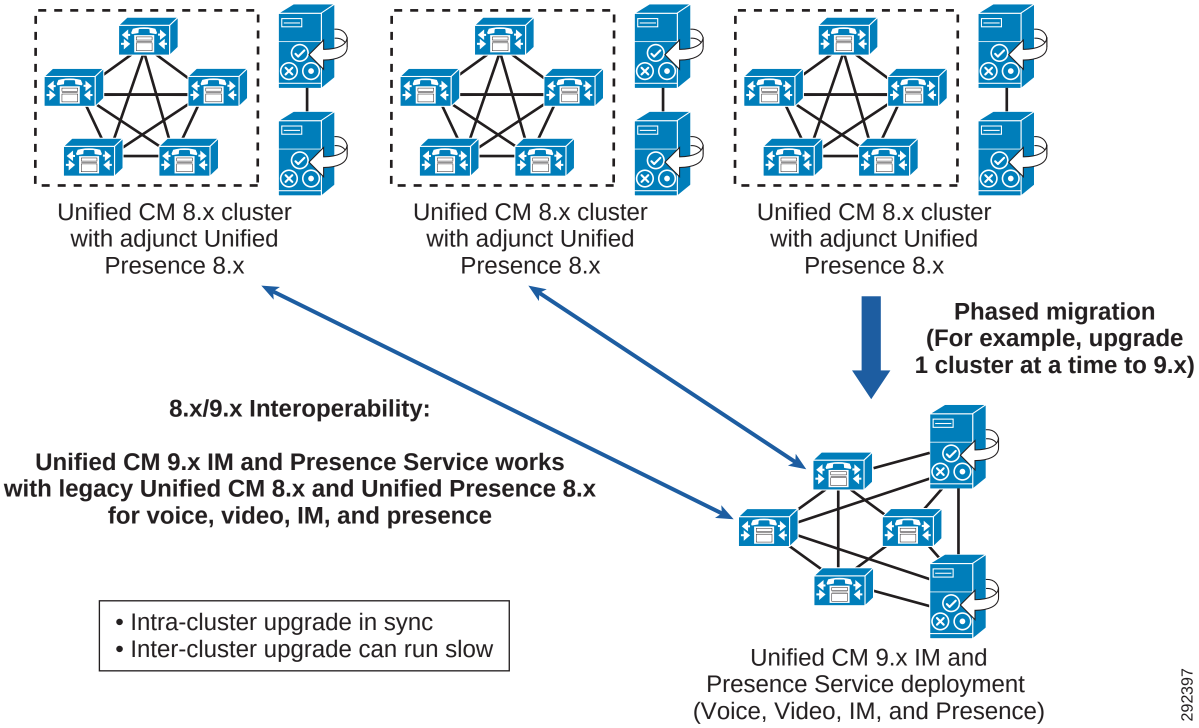
Cisco IM and Presence Migration

Migration of a Cisco Unified Presence deployment to Release 8.*x* is supported from Cisco Unified Presence 7.*x* only. User assignments on a Cisco Unified Presence cluster will be maintained on a per-server basis with migration from version 7.*x* to 8.*x*.

The following guidelines apply to migration of Cisco Unified Presence:

- Cisco Unified Presence 7.*x* deployments with more than one cluster require you to deactivate the presence engine on each server in the cluster before upgrading to Cisco Unified Presence 8.*x*. The presence engine should be reactivated only after you have completed all server upgrades to Cisco Unified Presence 8.*x*.
- For Cisco Unified Presence 7.*x* deployments with more than one cluster, Cisco recommends upgrading all clusters to Cisco Unified Presence 8.*x* at the same time.
- Migrating Cisco Unified Presence deployments to version 8.*x* brings in a second standard protocol in XMPP, with the addition of the Jabber XCP architecture. Cisco Unified Personal Communicator 7.*x*, inter-domain federation with Microsoft Office Communications Server, and third-party applications all create SIP/SIMPLE subscriptions. The number of active SIP/SIMPLE subscriptions can be viewed using the Real-Time Monitoring Tool (RTMT) Active Subscription counter. Cisco Unified Presence 8.*x* manages these subscriptions using the SIP Federation Connection Manager. With a large number of active SIP/SIMPLE subscriptions (above 20,000), or if subscriptions begin to fail after upgrade, Cisco recommends increasing the SIP Federation Connection Manager service parameter Pre-allocated SIP stack memory (bytes) to twice the amount of its current default value.
- For Cisco Unified Communications System 9.0 and later releases, Cisco Unified CM and the Cisco IM and Presence Service are required to run the same version at all times.
- The Cisco IM and Presence Service (9.0 and later releases) supports backward compatibility, thereby providing customers who have multiple Cisco Unified Presence systems with the opportunity to upgrade those systems over a period of time. (See Figure 6-1.)

Figure 6-1 Large Enterprise Migration with Backward Compatibility


Unified CM 8.x cluster
with adjunct Unified
Presence 8.x
Unified CM 8.x cluster
with adjunct Unified
Presence 8.x
Unified CM 8.x cluster
with adjunct Unified
Presence 8.x
Phased migration
(For example, upgrade
1 cluster at a time to 9.x)
8.x/9.x Interoperability:
Unified CM 9.x IM and Presence Service works
with legacy Unified CM 8.x and Unified Presence 8.x
for voice, video, IM, and presence
• Intra-cluster upgrade in sync
• Inter-cluster upgrade can run slow
Unified CM 9.x IM and
Presence Service deployment
(Voice, Video, IM, and Presence)
292397

Migrating Licenses to the Cisco Enterprise License Manager (ELM)

Cisco Unified Communications System 9.0 and later releases move away from the Device License Unit (DLU) concept and implement user-based licensing, thereby matching what a customer actually purchases. This new licensing model is also under the management of the Cisco Enterprise License Manager (ELM). For more details on ELM, see the section on Enterprise License Manager, page 8-9.

For those products that support ELM, customers can continue to fulfill their licenses from the Cisco Product License Registration portal at http://www.cisco.com/go/license, and can now import them into the ELM instead of the individual product instances. Customers who have already deployed Cisco Unified Communications can use the following process (illustrated in Figure 6-2) to migrate existing licenses to Cisco Unified Communications System 9.*x* licenses:

1. Fully license all Unified Communications products prior to conversion to 9.*x*.

   Fulfill any unfulfilled Product Activation Keys (PAKs), and install licenses on the Unified Communications products. Product Activation Keys come with your product. Installing unfulfilled licenses prior to your migration to version 9.*x* will allow these unfulfilled licenses to be converted to new Unified Communications 9.*x* equivalent licenses.

   This step is necessary for the migration of all pre-9.0 licenses and must be done *prior* to the migration to version 9.*x*. After a system has been upgraded to version 9.*x*, any unfulfilled PAKs can no longer be used because Unified Communications 9.*x* uses a different licensing infrastructure.

Go to Product License Registration at http://www.cisco.com/go/license.

2. Order the Cisco Unified Communications System 9.*x* software.

   This step is required only if the customer does not already have access to Unified Communications 9.*x* software through some alternate method (for example, Cisco Software Download website).

3. Receive the Cisco Unified Communications System 9.*x* software.

4. Upgrade the Unified Communications products (such as Cisco Unified Communications Manager and Cisco Unity Connection).

   ELM will be installed as part of the upgrade. The upgrade must be done before migrating the licenses because ELM must be installed to perform the license migration.

5. Add Unified Communications product instances to the ELM product inventory. This is a normal part of the setup for Cisco Unified Communications Manager 9.*x* licensing and is needed for license migration.

   When a product instance is added, current licensing information (including unused DLUs in the case of Unified CM) is automatically sent to ELM. This information is used by the ELM Upgrade Licenses (migration) utility to determine the types of user licenses to be migrated.

6. Use the ELM Upgrade Licenses (migration) utility to view and modify the licenses to be upgraded.

   This step determines the actual licenses that will be migrated to the new Unified Communications System 9.*x* user licenses. It allows you to view the existing licenses, plan for license migration, select license migration, and generate an upgrade license request.

   As mentioned earlier, this Upgrade Licenses (migration) conversion can be done only once; therefore it is very important to correctly determine the licenses to be migrated.

7. Generate a License Migration Request.

   The License Migration Request contains the license information determined from the previous step and also contains identifying information of the ELM to which the new Unified Communications System 9.*x* user licenses will be tied. This information generated from the request is then provided to the Cisco Product License Registration site to generate the new licenses.

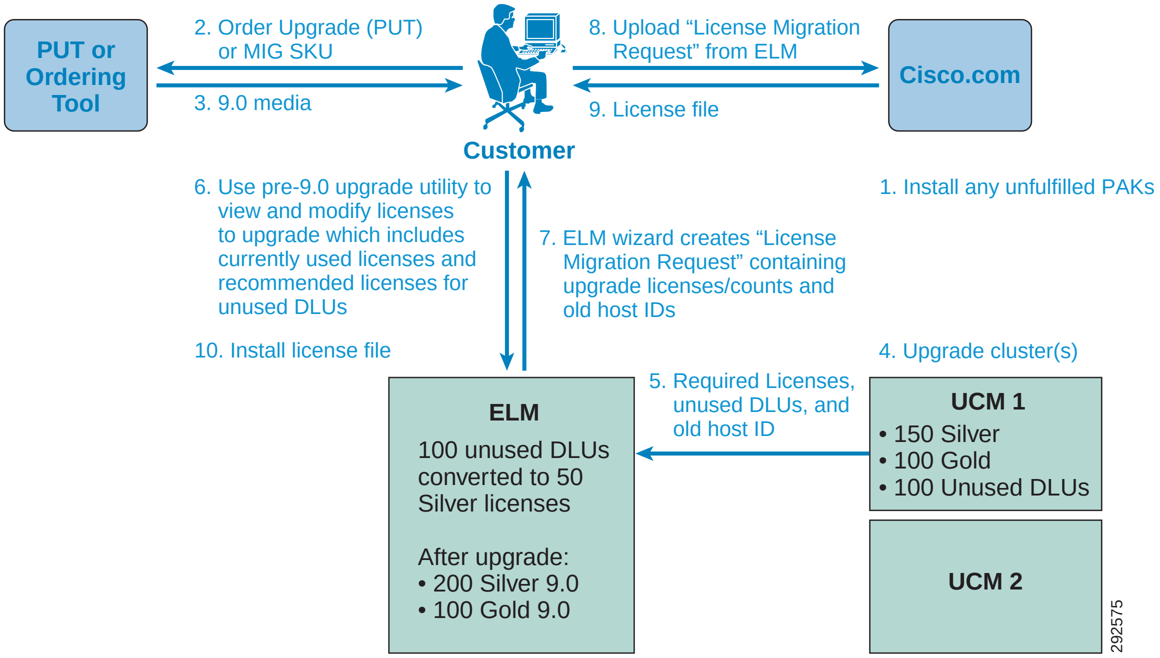
8. Submit the License Migration Request to Product License Registration. Go to Cisco Product License Registration site at http://www.cisco.com/go/license. Select **Get New**, and then **Migration License** and **Cisco Unified Communications 9.0**. Paste the contents of the License Migration Request from ELM into the request, and complete the request.

Note Once this request is submitted, no further changes can be made to the type and quantity of licenses, other than by purchasing new additional licenses.

9. Receive the license file. The license file is generated from the submitted request and sent by email from license@cisco.com.

10. Install the license file. Install the license file sent from Cisco onto ELM. From ELM License Management, go to **Licenses** > **Install License Fil**e. The pre-9.0 licenses have now been migrated and are available for use.

Figure 6-2 Process for Upgrading Pre-9.0 Cisco Unified Communications Systems


PUT or Ordering Tool
2. Order Upgrade (PUT) or MIG SKU
3. 9.0 media
Customer
8. Upload "License Migration Request" from ELM
9. License file
Cisco.com
6. Use pre-9.0 upgrade utility to view and modify licenses to upgrade which includes currently used licenses and recommended licenses for unused DLUs
1. Install any unfulfilled PAKs
7. ELM wizard creates "License Migration Request" containing upgrade licenses/counts and old host IDs
10. Install license file
4. Upgrade cluster(s)
5. Required Licenses, unused DLUs, and old host ID
ELM
100 unused DLUs converted to 50 Silver licenses
After upgrade:
• 200 Silver 9.0
• 100 Gold 9.0
UCM 1
• 150 Silver
• 100 Gold
• 100 Unused DLUs
UCM 2
292575

References

- Download software — Get the latest updates, patches, and releases of Cisco software.

  http://www.cisco.com/cisco/software/navigator.html

- Product License Registration

  http://www.cisco.com/go/license

PART 2

Unified Communications Call Routing

CHAPTER 7

Overview of Cisco Unified Communications Call Routing

Revised: February 29, 2012; OL-27282-05

Once the network infrastructure has been put in place for your Cisco Unified Communications System, call routing applications, components, and services can be layered on top of this infrastructure. There are numerous applications and features that can, and in some cases must, be deployed on the network infrastructure. In general, you should deploy the following call routing components, features, and services:

- Call processing agent — Provides telephony services and call routing capabilities.
- Dial plan — Provides endpoint numbering, dialed digits analysis, and classes of restriction to limit types of calls that a user can make.
- Call admission control — Provides mechanisms for preventing oversubscription of network bandwidth by limiting the number of calls that are allowed on the network at a given time based on overall call capacity of the call processing components and network bandwidth.
- Video telephony services — Provide the ability to provision and register video endpoints as well as to set up, route, and maintain video calls on the network.
- PSTN gateways and provider voice and data services — Provide access to voice and data networks outside the enterprise, including the PSTN, Internet, and service provider IP-based trunks.
- Remote site survivability — Provides continuation of basic telephony services at remote sites when the central-site telephony services are unavailable due to failed or flapping network connectivity.

The chapters in this part of the SRND cover the features, components, and services mentioned above. Each chapter provides an introduction to the component or service, followed by discussions surrounding architecture, high availability, capacity planning, and design considerations. The chapters focus on design-related aspects of the applications and services rather than product-specific support and configuration information, which is covered in the related product documentation.

This part of the SRND includes the following chapters:

- Call Processing, page 8-1

  This chapter examines the various types of call processing applications and platforms that facilitates IP telephony call routing. The chapter examines the call processing architecture, including hardware options, Unified CM clustering capabilities, high availability considerations for call processing, and capacity planning.

- Dial Plan, page 9-1

  This chapter explores dial plan features and functions that enable the call processing application to route calls to appropriate numbers. The chapter considers various aspects of dial plan services, including dial plan constructs, dial plan numbering options and design considerations, classes of restriction, inbound and outbound calling features, and dial plan and call routing redundancy mechanisms.

- Emergency Services, page 10-1

  This chapter discusses accessing emergency services through Public Safety Answering Points (PSAPs) on the PSTN from within the enterprise IP telephony environments, an important aspect of most deployments due to possible critical needs for medical, fire, and other emergency response services. The chapter provides an overview of the various emergency service components both inside and outside the enterprise. It also discusses planning, 911 network service providers, gateway interfaces, and number-to-location mapping.

- Call Admission Control, page 11-1

  This chapter examines the potential for oversubscribing IP links, which causes the voice quality for phone calls to become unacceptable. It also examines the use of call admission control to allow only a certain number of simultaneous calls on the network at a given time to prevent oversubscription. This chapter covers call admission control types, including location-based call admission control and RSVP, as well as design and deployment guidelines for successfully deploying admission control services.

- IP Video Telephony, page 12-1

  This chapter covers video telephony, an important and integral part of collaborative communication. The chapter discusses video telephony components, protocols and codecs, multipoint conferencing, and gatekeeper aspects for video call routing.

- Gateways, page 13-1

  This chapter explores voice and IP gateways, which are critical components of Unified Communications deployments because they provide the path for connecting to phones on the public telephone network. This chapter looks at gateway traffic types and patterns, protocols, capacity planning, and platform selection, as well as fax and modem support.

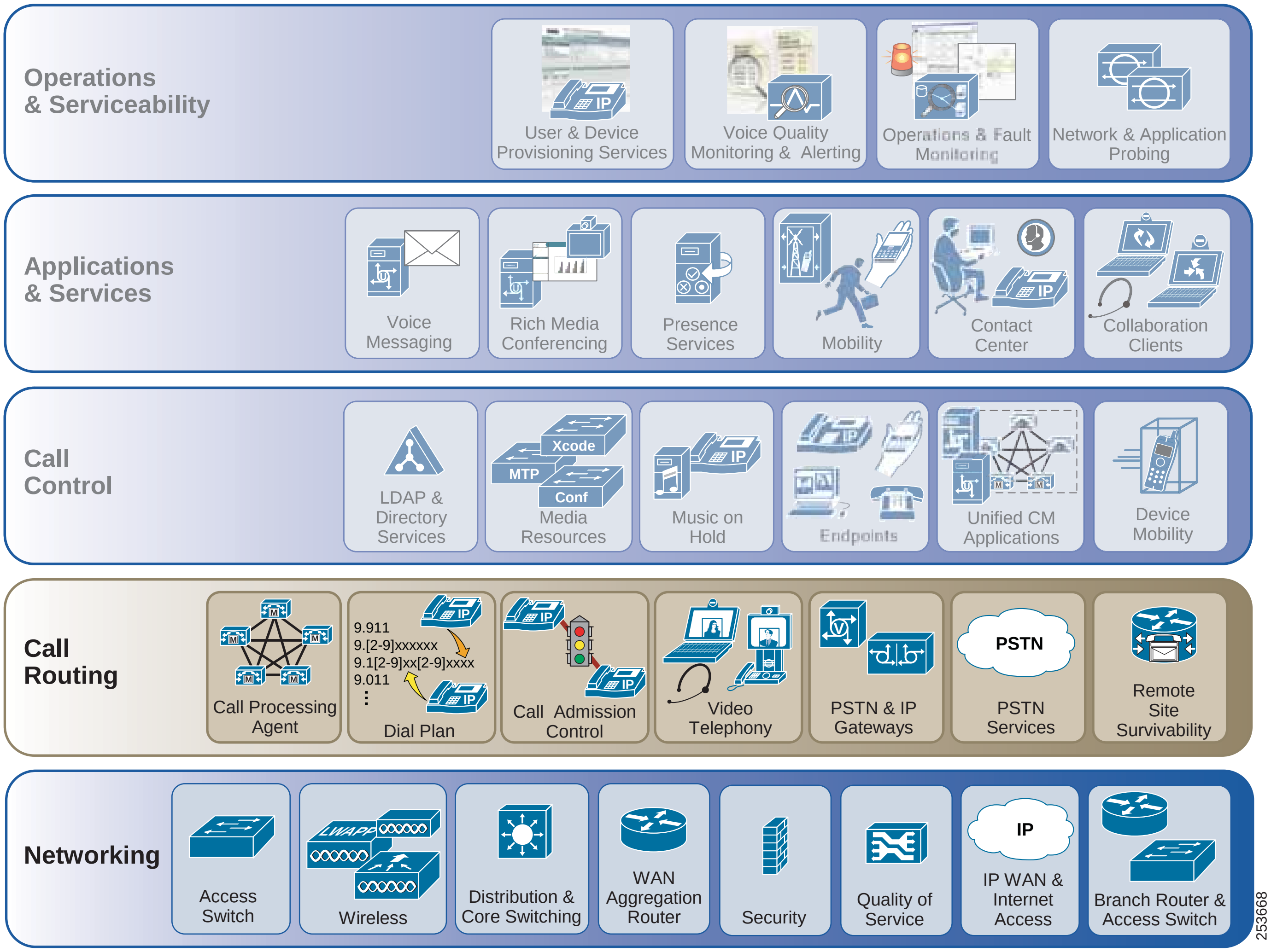
- Cisco Unified CM Trunks, page 14-1

  This chapter covers both intercluster and provider trunks, which provide the ability to router voice calls over IP and leverage various Unified Communications features and functions. This chapter discusses H.323 and SIP trunks, codecs, and supplementary services over these trunks, as well as sizing of trunks to accommodate network call load.

Architecture

Just as with other network and application technology systems, Unified Communications call routing components and services must be layered on top of the underlying network infrastructures. Figure 7-1 shows the logical location of call routing applications and services in the overall Cisco Unified Communications System architecture.

Figure 7-1 Cisco Unified Communications Call Routing Architecture



Unified Communications call routing components and services such as call processing agents and IP and PSTN gateways rely on the underlying network infrastructure for network connectivity and access. By connecting to the underlying network infrastructure, call routing components and features are able to leverage end-to-end network connectivity and quality of service to access both the enterprise and public telephone networks. In turn, call routing applications and services provide basic Unified Communications functions such as call control, dial plan, call admission control, and gateway services to other applications and services in the deployment. For example, a Unified CM cluster connects to the IP network through a switch in order to communicate with other devices and applications within the network as well as to access other devices and services in other locations. At the same time, the Unified CM cluster provides services such as phone registration and media resource provisioning and allocation to call control components and services such as IP phones.

Further, just as call routing components rely on the network infrastructure for network connectivity, call routing components and services are also often dependent upon each other for full functionality. For example, while Unified CM provides registration and call routing services to various IP endpoints within the network, it is completely dependent upon gateways and gateway services to route calls beyond the enterprise.

High Availability

As with the network infrastructure, critical Unified Communications call routing services should be made highly available to ensure that required features and functionality remain available if failures occur in the network or with individual call routing components. It is important to understand the various types of failures that can occur and the design considerations around those failures. In some cases, the failure of a single server or component (for example, a subscriber node in a Unified CM cluster) might have little or no impact due to the redundant nature of the Unified CM clustering mechanism. However, in other cases a single failure can impact multiple components or services. For example, the failure of a PSTN or IP gateway could result in loss of access to the public telephone network, and even though a call processing agent such as Unified CM is still available and able to provide most features and services, it cannot route calls to the PSTN because there is no path available if the gateway fails. To avoid these types of situations, you should deploy multiple PSTN gateways to provide redundant gateway services, and you should configure the call processing agent to handle call routing to both gateways as needed.

For features and services such as dial plan and call admission control, high availability considerations include temporary loss of functionality due to network connectivity or call processing agent application server failures, resulting in the inability of the call agent to route calls and therefore the inability for callers to make calls. Oversubscription of the network could also occur if call admission control services are not available to the endpoints initiating a call. For example, if RSVP call admission control is in use and an RSVP agent fails or loses connectivity to the network, the call may still go through but without the call admission control service being aware of the call, thus potentially resulting in poor quality. To avoid these types of scenarios, provide call admission control resiliency by deploying multiple RSVP agents so that a failed RSVP agent will not prevent another RSVP agent from providing the call admission control service.

High availability considerations are also a concern for components and services such as video endpoints and remote site survivability. For deployments with network-attached remote sites where devices are leveraging call processing services from an agent in a central site, remote site survivability using SRST, for example, can ensure that local phones within the remote site will still receive call processing services in the event of a connectivity failure to the central site. Likewise, to ensure that video endpoints are highly available, you can deploy more than one multipoint control unit (MCU) in case one fails.

Capacity Planning

The network infrastructures must be designed and deployed with consideration for the capacity and scalability of the individual components and the overall system. Similarly, deployments of call routing components and services must also be designed with attention to capacity and scalability considerations. When deploying various call routing applications and services, not only is it important to consider the scalability of the applications and services themselves, but you must also consider the scalability of the underlying network infrastructure. Certainly the network infrastructure must have available bandwidth and be capable of handling the additional traffic load that the call routing components will create. Similarly, the call routing infrastructure and its components must be capable of handling all the required device configurations and registrations as well as the call load or busy hour call attempts (BHCA),

For example, with call processing agents such as Unified CM, it is critical to assess the size of the deployment in terms of number of users, endpoints, and calls per user per hour, and to deploy sufficient resources to handle the required load. If a call processing agent is undersized and does not have sufficient resources, features and services will begin to fail as the load increases. Two of the chief considerations when attempting to size a call processing deployment are the call processing type and the call processing hardware. Both of these are critical for sizing the system appropriately given the number of users, locations, devices, and so on. As an example, Cisco Unified Communications Manager has a much higher capacity than Cisco Unified Communications Manager Express and should therefore be used for larger deployments. In addition, the server platform selected to run the call processing agent will, in many cases, determine the maximum load.

Capacity planning for remote site survivability is much the same in that it relies on backup call processing hardware. Selecting the appropriate Cisco IOS platform to provide backup or survivable call processing services typically begins with determining the number of devices or users that must be supported at that site in the event that connectivity to the central site is disrupted. Equally critical in this sizing exercise is the local PSTN gateway services. In the event of a central site connection failure, will the local PSTN gateway have sufficient circuits to be able to route all calls without blocking during the busiest hour? If the answer is no, adding additional gateways or trunks will be necessary to appropriately size the remote site for backup call processing.

PSTN and IP gateways must also be sized appropriately for a deployment, so that sufficient capacity is available to handle all calls in the busiest hour. In some cases, you might have to deploy multiple PSTN or IP gateways to provide enough resources.

When sizing call admission control, ensure that sufficient bandwidth is available over network connections to support the required number of calls. If sufficient bandwidth is not available, additional network capacity, gateways, and IP or telephony trunks may be required.

Sizing dial plan services is also important. However, in most cases dial plan capacity in terms of the number of endpoints or phone numbers, route patterns, or other dial plan constructs, is completely dependent upon the type of call processing agent and platform used.

For components and services such as video telephony, appropriate sizing is just as critical. Capacity planning considerations for video telephony center mainly on network bandwidth, available video ports, and MCU sessions. In most cases additional capacity can be added by increasing the number of application servers and MCUs or by upgrading server or MCU hardware with higher-scale models, assuming the underlying network infrastructure is capable of handling the additional load.

For a complete discussion of system sizing, capacity planning, and deployment considerations related to sizing, refer to the chapter on Unified Communications Design and Deployment Sizing Considerations, page 29-1.

CHAPTER 8

Call Processing

Revised: April 30, 2013; OL-27282-05

The handling and processing of voice and video calls is a critical function provided by IP telephony systems. This functionality is handled by some type of call processing entity or agent. Given the critical nature of call processing operations, it is important to design unified communications deployments to ensure that call processing systems are scalable enough to handle the required number of users and devices and are resilient enough to handle various network and application outages or failures.

This chapter provides guidance for designing scalable and resilient call processing systems with Cisco call processing products. These products include Cisco Unified Communications Manager (Unified CM), Cisco Business Edition, and Cisco Unified Communications Manager Express (Unified CME). In addition, this chapter provides coverage for gatekeeper functionality, which is another critical function for unified communications deployments in scenarios where multiple call processing systems or agents are deployed in parallel. In all cases, the discussions focus predominately on the following factors:

- Scale — The number of users, locations, gateways, applications, and so forth
- Performance — The call rate
- Resilience — The amount of redundancy

Specifically, this chapter focuses on the following topics:

- Call Processing Architecture, page 8-2

  This section discusses general call processing architecture and the various call processing hardware options. This section also provides information on Unified CM clustering.

- High Availability for Call Processing, page 8-14

  This section examines high availability considerations for call processing, including network redundancy, server or platform redundancy, and load-balancing.

- Capacity Planning for Call Processing, page 8-24

  This section provides an overview of sizing for call processing deployments and introduces the Unified Communications Sizing Tool. This tool provides guidance on sizing and required resources for various components of a Unified Communications deployment, and it should be used when planning an IP Telephony deployment.

- Design Considerations for Call Processing, page 8-28

  This section provides a summarized list of high-level design guidelines and best practices for deploying call processing.

- Computer Telephony Integration (CTI), page 8-30

  This section explains the Cisco Computer Telephony Integration (CTI) architecture and discusses CTI components and interfaces, CTI functionality, and CTI provisioning and capacity planning.

- Gatekeeper Design Considerations, page 8-37

  This section explains how gatekeepers can be used in a Cisco Unified Communications deployment. Cisco Gatekeeper may also be paired with other standby gatekeepers or may be clustered for higher performance and resilience. Gatekeepers may also be used for call routing and call admission control.

- Interoperability of Unified CM and Unified CM Express, page 8-44

  This section explains the H.323 and SIP integration between Cisco Unified CM and Cisco Unified Communications Manager Express (Unified CME) in a distributed call processing deployment.

What's New in This Chapter

Table 8-1 lists the topics that are new in this chapter or that have changed significantly from previous releases of this document.

Table 8-1 New or Changed Information Since the Previous Release of This Document

New or Revised Topic	Described in	Revision Date
Minor corrections and changes	Various sections	April 30, 2013
Minor updates for Cisco Business Edition	Various sections in this chapter	October 31, 2012
CTI Remote Device	Computer Telephony Integration (CTI), page 8-30	June 28, 2012
Enterprise License Manager	Enterprise License Manager, page 8-9	June 28, 2012

Call Processing Architecture

In order to design and deploy a successful Unified Communications system, it is critical to understand the underlying call processing architecture that provides call routing functionality. This functionality is provided by the following Cisco call processing agents:

- Cisco Unified Communications Manager Express (Unified CME)

  Cisco Unified CME provides call processing services for small single-site deployments, larger distributed multi-site deployments, and deployments in which a local call processing entity at a remote site is needed to provide backup capabilities for a centralized call processing deployment of Cisco Unified CM.

- Cisco Business Edition

  Cisco Business Edition provides call processing services for small single-site deployments or small distributed multisite deployments. There are three versions of Cisco Business Edition: Business Edition 3000, Business Edition 5000 and Business Edition 6000. The main differences between the three versions are as follows:

  – The hardware on which Cisco Business Edition is deployed and the number of applications and services that can be run co-resident. Business Edition 3000 and 5000 provide co-resident Cisco Unified CM call processing and Cisco Unity Connection messaging services. Business Edition 6000 supports up to five co-resident applications on a single UCS server. The supported applications are Cisco Unified CM, Cisco Unified Provisioning Manager, Cisco Unity Connection, Cisco IM and Presence, Cisco Unified Contact Center Express (Unified CCX), Cisco Unified Attendant Console Services, and Cisco TelePresence Video Communication Server (VCS).

  – The capacity of the system. Cisco Business Edition supports between 300 and 1,000 users and between 400 and 1,200 endpoints, depending on the version.

  – The install and upgrade procedure. Business Edition 3000 and 5000 use a single software image to install and/or upgrade Unified CM and Unity Connection natively on the supported Cisco MCS platforms. Business Edition 6000 uses discrete software images to install and/or upgrade each of the co-resident applications in VMware.

- Cisco Unified Communications Manager (Unified CM)

  Cisco Unified CM provides call processing services for small to very large single-site deployments, multi-site centralized call processing deployments, and/or multi-site distributed call processing deployments.

This section examines various call processing hardware options and then provides an overview of Unified CM clustering.

Call Processing Hardware

Three enterprise call processing types are supported on various types of platforms:

- Cisco Unified CME runs on Cisco Integrated Services Routers (ISR).
- Cisco Business Edition 3000 and Business Edition 5000 run as appliances directly on Cisco Media Convergence Servers (MCS). Cisco Business Edition 6000 runs as a virtual machine with the VMware Hypervisor on a Cisco Unified Computing System (UCS) C-Series Server.
- Cisco Unified CM runs either as an appliance directly on MCS (or equivalent) servers or as a virtual machine with the VMware Hypervisor. When Unified CM is deployed as a virtual machine, two hardware options are available: Tested Reference Configurations and specification-based hardware support. Tested Reference Configurations (TRC) are selected hardware configurations based on the Cisco UCS hardware. They are tested and documented for specific guaranteed performance, capacity, and application co-residency scenarios running "full-load" Unified Communications virtual machines. TRCs are intended for customers who want a packaged solution from Cisco that is pre-engineered for a specific deployment scenario and/or customers who are not experienced with hardware virtualization. For more information on the TRC, refer to the documentation at

  http://docwiki.cisco.com/wiki/Tested_Reference_Configurations_(TRC)

  Alternatively, more flexible hardware configurations are possible with the specification-based hardware support which, for example, adds support for Cisco UCS, Hewlett-Packard, and IBM platforms listed in the VMware Hardware Compatibility List (http://www.vmware.com/resources/compatibility/search.php), and for iSCSI, FCoE, and NAS (NFS) storage systems. Specification-based hardware support is intended for customers with extensive expertise in virtualization as well as server and storage sizing, who wish to use their own hardware standards. For more information on the specification-based hardware support, refer to the documentation at

  http://docwiki.cisco.com/wiki/Specification-Based_Hardware_Support

Table 8-2 provides a summary of the three enterprise call processing types, the types of servers or platforms on which these call processing applications reside, and the overall characteristics of those platforms. The table includes the Tested Reference Configurations for the Cisco UCS platforms but not the platforms that are supported with the specification-based hardware support.

Table 8-2 Types of Call Processing Platforms

Call Processing Type	Platform Type	Cisco Platform Model	Characteristics
Cisco Unified CME	Cisco IOS Router	2800, 2900, 3700, 3800, and 3900 Series[1]	• Single processor • Single or multiple power supplies, depending on model
Cisco Business Edition	Cisco Unified Computing System (UCS) C-Series Rack-Mount Servers (Business Edition 6000)	UCS C-Series: C200 M2 and C220 M3	• Multiple processors • Multiple SAS disk drives with RAID 10 support (running ESXi and also storing Cisco Unified Communications virtual machines on local disk drives) • No bare metal support[5]
	Standard Cisco Media Convergence Server (MCS) for Business Edition 5000	MCS 7828[2]	• Single processor • Single power supply • SATA controller with RAID 0/1 support • Dual IP interfaces
	Standard MCS for Business Edition 3000 version 8.5(1) and later	MCS 7816-I5	• Single processor • Single power supply • Non-RAID SATA hard disk • Dual IP interfaces
	Purpose-built appliance for Business Edition 3000 version 8.6(1) and later	MCS 7890-C1	• Single processor • Single power supply • Integrated voice gateway with 2 T1/E1 ports • On-board DSPs for media resources • Single IP interface

Table 8-2 Types of Call Processing Platforms (continued)

Call Processing Type	Platform Type	Cisco Platform Model	Characteristics
Cisco Unified CM	Standard MCS	MCS 7815, MCS 7816, or equivalent	• Single processor • Single power supply • Non-RAID SATA hard disk • Dual IP interfaces[3]
	Standard MCS with RAID	MCS 7825 or equivalent	• Single processor • Single power supply • SATA controller with RAID 0/1 support • Dual IP interfaces
	High-availability MCS	MCS 7835, MCS 7845, or equivalent	• One or multiple processors • Multiple power supplies • Multiple Serial Attached SCSI (SAS) drives with RAID 1 • Dual IP interfaces
	Unified Computing System (UCS) B-Series Blade Servers	UCS B-Series (for example, B200, B230, B440)	• Half-width or full-width blade • Multiple processors • Multiple power supplies • Multiple SAS disk drives (running ESXi)[4] or diskless blades • Cisco Unified Communications virtual machines stored on FC SAN Storage • No bare metal support[5]
	Unified Computing System (UCS) C-Series Rack-Mount Servers	Low-end UCS C-Series (for example, C200, C220 TRC#2)	• Multiple processors • One or multiple power supplies • Multiple SAS local disk drives • No bare metal support[5]
		High-end UCS C-Series (for example, C210, C220 TRC#1, C240, C260)	• Multiple processors • Multiple power supplies • Multiple SAS local disk drives running ESXi only, running ESXi and Unified Communications Virtual Machines, or diskless servers[6] • No bare metal support[5]

[1] This is not an exhaustive list of supported Cisco IOS platforms.

[2] The Cisco MCS 7828 supports only Business Edition 5000.

[3] The Cisco MCS 7815 platform has only a single IP interface

[4] UCS B-Series blade server disks are for virtual machine software (ESXi) only. Applications such as Unified CM are not installed and do not run on the on-blade drives.

[5] UCS B-Series and C-Series servers offer no bare metal support for Cisco Unified Communications applications. UCS B-Series and C-Series servers must run ESXi hypervisor software.

6. Supported options depend on the server model. For more details, refer to http://www.cisco.com/go/uc-virtualized.

For a complete list of supported MCS servers or equivalents, refer to the documentation available at

http://www.cisco.com/go/swonly

For a complete list of Tested Reference Configurations or for details on the specification-based hardware support, refer to the documentation available at

http://www.cisco.com/go/uc-virtualized

Determining the appropriate call processing type and platform for a particular deployment will depend on the scale, performance, and redundancy required. In general, Unified CM and the higher-end MCS and UCS servers provide more capacity and higher availability, while Cisco Unified CME and Cisco Business Edition provide lower levels of capacity and redundancy. For specifics regarding redundancy and scalability, see the sections on High Availability for Call Processing, page 8-14, and Capacity Planning for Call Processing, page 8-24.

Unified CM Cluster Services

Cisco Unified CME, Business Edition 3000 and 5000, and Unified CM running on an MCS-7815 or MCS-7816 are standalone call processing applications or entities. However, Unified CM running on all other server platforms involves the concept of clustering. The Unified CM architecture enables a group of physical servers to work together as a single call processing entity or IP PBX system. This grouping of servers is known as a *cluster*. A cluster of Unified CM servers may be distributed across an IP network, within design limitations, allowing for spatial redundancy and, hence, resilience to be designed into the Unified Communications System.

Within a Unified CM cluster, there are servers that provide unique services. Each of these services can coexist with others on the same physical server. For example, in a small system it is possible to have a single server providing database services, call processing services, and media resource services. As the scale and performance requirements of the cluster increase, many of these services should be moved to dedicated physical servers.

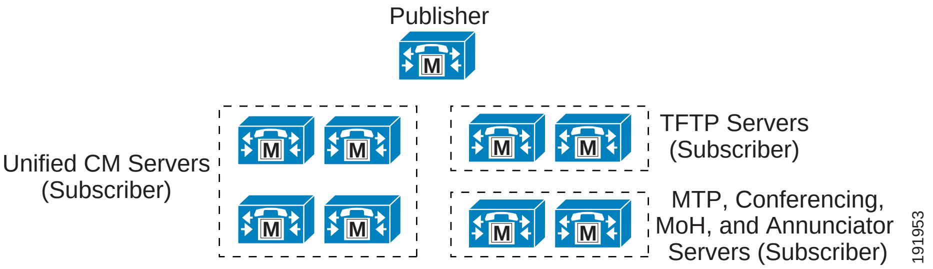
Note While Cisco recommends using the same server model for all servers in a cluster, mixing server models within a cluster is supported provided that all of the individual hardware versions are supported and that all servers are running the same version of Unified CM. However, differences in capacity between various server models within a cluster must be considered because the overall cluster capacity might ultimately be dictated by the capacity of the smallest server within the cluster. Mixing servers from different vendors within a cluster is also supported and does not have any adverse capacity implications, provided that all servers in the cluster are the same model type. For information on call processing capacity, see the section on Capacity Planning for Call Processing, page 8-24.

The following section describes the various functions performed by the servers that form a Unified CM cluster, and it provides guidelines for deploying the servers in ways that achieve the desired scale, performance, and resilience.

Cluster Server Nodes

Figure 8-1 illustrates a typical Unified CM cluster consisting of multiple server nodes. There are two types of Unified CM servers, publisher and subscriber. These terms are used to define the database relationship during installation.

Figure 8-1 Typical Unified CM Cluster



Publisher

The publisher is a required server in all clusters, and as shown in Figure 8-1, there can be only one publisher per cluster. This server is the first to be installed and provides the database services to all other subscribers in the cluster. The publisher server is the only server that has full read and write access to the configuration database.

On larger systems with more than 1250 users, Cisco recommends a dedicated publisher to prevent administrative operations from affecting the telephony services. A dedicated publisher does not provide call processing or TFTP services running on the server. Instead, other subscriber servers within the cluster provide these services.

The choice of hardware platform for the publisher should be based on the desired scale and performance of the cluster. Cisco recommends that the publisher have the same server performance capability as the call processing subscribers. Ideally the publisher should also be a high-availability server to minimize the impact of a hardware failure.

Subscriber

When the software is installed initially, only the database and network services are enabled. All subscriber nodes subscribe to the publisher to obtain a copy of the database information. However, in order to reduce initialization time for the Unified CM cluster, all subscriber servers in the cluster attempt to use their local copy of the database when initializing. This reduces the overall initialization time for a Unified CM cluster. All subscriber nodes rely on change notification from the publisher or other subscriber nodes in order to keep their local copy of the database updated.

As shown in Figure 8-1, multiple subscriber nodes can be members of the same cluster. Subscriber nodes include Unified CM call processing subscriber nodes, TFTP subscriber nodes, and media resource subscriber nodes that provide functions such as conferencing and music on hold (MoH).

Call Processing Subscriber

A call processing subscriber is a server that has the Cisco CallManager Service enabled. Once this service is enabled, the server is able to perform call processing functions. Devices such as phones, gateways, and media resources can register and make calls only to servers with this service enabled. As shown in Figure 8-1, multiple call processing subscribers can be members of the same cluster. In fact, Unified CM supports up to eight call processing subscriber nodes per cluster.

TFTP Subscriber

A TFTP subscriber or server node performs two main functions as part of the Unified CM cluster:

- The serving of files for services, including configuration files for devices such as phones and gateways, binary files for the upgrade of phones as well as some gateways, and various security files
- Generation of configuration and security files, which are usually signed and in some cases encrypted before being available for download

The Cisco TFTP service that provides this functionality can be enabled on any server in the cluster. However, in a cluster with more than 1250 users, other services might be impacted by configuration changes that can cause the TFTP service to regenerate configuration files. Therefore, Cisco recommends that you dedicate a specific subscriber node to the TFTP service, as shown in Figure 8-1, for a cluster with more than 1250 users or any features that cause frequent configuration changes.

Cisco recommends that you use the same hardware platform for the TFTP subscribers as used for the call processing subscribers.

Media Resource Subscriber

A media resource subscriber or server node provides media services such as conferencing and music on hold to endpoints and gateways. These types of media resource services are provided by the Cisco IP Voice Media Streaming Application service, which can be enabled on any server node in the cluster.

Media resources include:

- Music on Hold (MoH) — Provides multicast or unicast music to devices that are placed on hold or temporary hold, transferred, or added to a conference. (See Music on Hold, page 17-21.)
- Annunciator service — Provides announcements in place of tones to indicate incorrectly dialed numbers or call routing unavailability. (See Annunciator, page 17-20.)
- Conference bridges — Provide software-based conferencing for ad-hoc and meet-me conferences. (See Conferencing, page 17-6.)
- Media termination point (MTP) services — Provide features for H.323 clients, H.323 trunks, and Session Initiation Protocol (SIP) endpoints and trunks. (See Media Termination Point (MTP), page 17-12.)

Because of the additional processing and network requirements for media resource services, it is essential to follow all guidelines for running media resources within a cluster. Generally, Cisco recommends non-dedicated media resource subscribers for multicast MoH and annunciator services, but dedicated media resource subscribers as shown in Figure 8-1 are recommended for unicast MoH as well as large-scale software-based conferencing and MTPs unless those services are within the design guidelines detailed in the chapter on Media Resources, page 17-1.

Additional Cluster Services

In addition to the specific types of subscriber nodes within a Unified CM cluster, there are also other services that can be run on the Unified CM call processing subscriber nodes to provide additional functionality and enable additional features.

Computer Telephony Integration (CTI) Manager

The CTI Manager service acts as a broker between the Cisco CallManager service and TAPI or JTAPI integrated applications. This service is required in a cluster for any applications that utilize CTI. The CTI Manager service provides authentication of the CTI application and enables the application to monitor and/or control endpoint lines. CTI Manager can be enabled only on call processing subscribers, thus allowing for a maximum of eight nodes running the CTI Manager service in a cluster.

For more details on CTI Manager, see Computer Telephony Integration (CTI), page 8-30.

Unified CM Applications

Various types of application services can be enabled on Unified CM, such as Cisco Unified CM Assistant, Extension Mobility, and Web Dialer. For detailed design guidance on these applications, see the chapter on Cisco Unified CM Applications, page 19-1.

Enterprise License Manager

Cisco Unified Communications System 9.*x* incorporates an Enterprise License Manager (ELM) that administers software licenses based on users rather than devices. Customers purchase user licenses and add them to the ELM application. The ELM application then collects requirements from all the applications, aggregates them, and compares them with the total available entitlements.

The following Unified Communications applications use the ELM:

- Cisco Unified Communications Manager (Unified CM)
- Cisco Unity Connection

Licenses are purchased from Cisco, delivered by email, and then loaded into the ELM. Whenever a subscribing application requires licenses, they are deducted from the ELM license pool. Similarly, if licenses are no longer required, they are returned to the ELM license pool for future use.

A 60 day grace period allows administrators to add users even if insufficient licenses exist within the ELM license pool. If sufficient licenses are not available once the 60 day grace period expires, then the Unified Communications application(s) will no longer allow any further changes; however, the application(s) will continue to function with no loss of service.

For more information on Cisco Unified Communications licensing, refer to the information at

http://www.cisco.com/go/uclicensing

Deployment Scenarios

ELM can run either as a co-resident service, in which case it is automatically installed alongside any Unified Communications application that supports it, or it can be installed on a dedicated server or virtual machine. When operational, ELM consumes only a very small amount of resources and hence is considered to have no impact to server or virtual machine sizing. Furthermore, ELM is deployed as a non-redundant application. In the event that the ELM application becomes unavailable (for example, if the server that it resides on suffers a catastrophic hardware failure), the customer has the 60 day grace period within which the application needs to be restored before license enforcement occurs. When enforcement is invoked, bear in mind that applications continue to function without loss of service.

Deployment Recommendations:

- If you are installing only a single application on a single server or cluster, run ELM co-resident.
- If you are installing a very small number of application instances, you may:
  - Run ELM on a separate virtual machine or server. This is the recommended approach.
  - Run a different ELM on each application server if you do not need license pooling and do not desire centralized license management.
  - Run a single ELM co-resident with one application server if you want license pooling and/or centralized management, but you are unwilling to dedicate a virtual machine or server for running the ELM.
- If you have a medium to large deployment, run ELM on a separate server or virtual machine. The incremental impact on the number of required virtual machines or servers is minimal in this case, and the trade-off between operating expenses and capital expenditures is favorable.

The ELM may be deployed in any of the following ways:

- Enterprise or global

  As the description implies, one ELM instance can support an entire enterprise or global deployment. This model provides the most simplicity by utilizing one common license pool for all the Unified Communications applications subscribing to the ELM.
- Regional or line of business

  For an enterprise that has multiple Unified Communications deployments across the globe, multiple ELM instances can be configured per region (for example, one for North America, a second for EMEA, and a third for APAC). This model enables an enterprise to account more easily for the costs of licenses across differing fiscal boundaries.
- Individual Unified Communications application

  For those customers requiring even more granularity, an ELM instance can be configured for each Unified Communications application. For example, if a customer has three Cisco Unified CM clusters, three ELM instances can be configured. This scenario is useful for customers who operate along more granular accounting lines and prefer multiple smaller license pools in order to better manage operating costs and other expenses.

Intracluster Communications

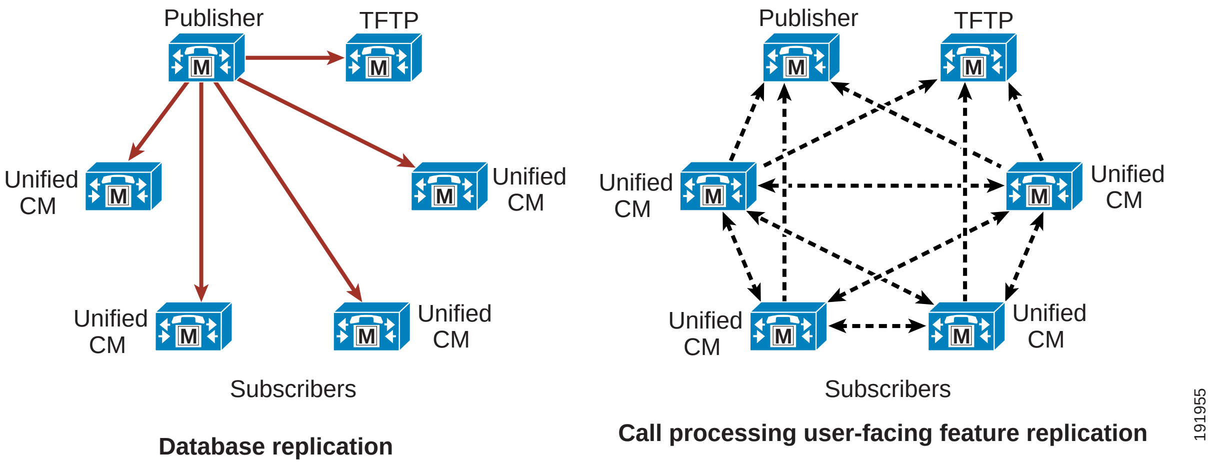
There are two primary kinds of intracluster communications, or communications within a Unified CM cluster (see Figure 8-2 and Figure 8-3.) The first is a mechanism for distributing the database that contains all the device configuration information (see “Database replication” in Figure 8-2). The configuration database is stored on a publisher server, and a copy is replicated to the subscriber nodes of the cluster. Most of the database changes are made on the publisher and are then communicated to the subscriber databases, thus ensuring that the configuration is consistent across the members of the cluster and facilitating spatial redundancy of the database.

Database modifications for user-facing call processing features are made on the subscriber servers to which an end-user device is registered. The subscriber servers then replicate these database modifications to all the other servers in the cluster, thus providing redundancy for the user-facing features. (See "Call processing user-facing feature replication" in Figure 8-2.) These features include:

- Call Forward All (CFA)
- Message waiting indicator (MWI)
- Privacy Enable/Disable

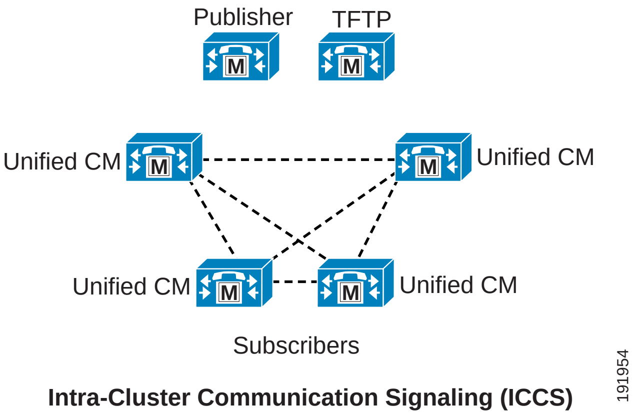
- Extension Mobility login/logout
- Hunt Group login/logout
- Device Mobility
- Certificate Authority Proxy Function (CAPF) status for end users and applications users
- Credential hacking and authentication

Figure 8-2 Replication of the Database and User-Facing Features



The second type of intracluster communication, called Intra-Cluster Communication Signaling (ICCS), involves the propagation and replication of run-time data such as registration of devices, locations bandwidth, and shared media resources (see Figure 8-3). This information is shared across all members of a cluster running the Cisco CallManager Service (call processing subscribers), and it ensures the optimum routing of calls between members of the cluster and associated gateways.

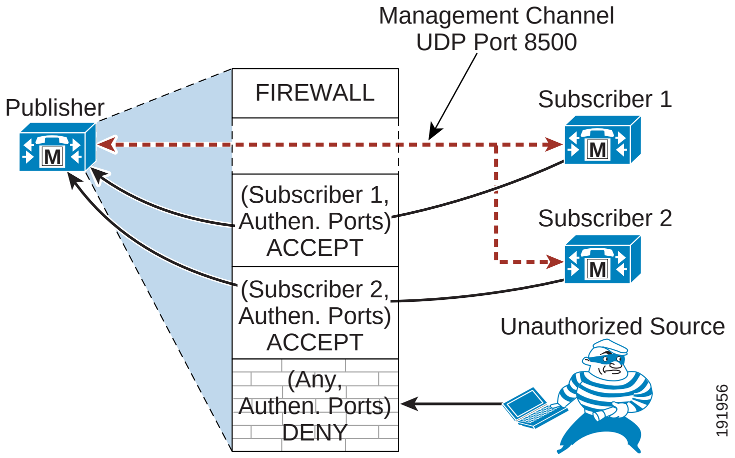
Figure 8-3 Intra-Cluster Communication Signaling (ICCS)



Intracluster Security

Each server in a Unified CM cluster runs an internal dynamic firewall. The application ports on Unified CM are protected by source IP filtering. The dynamic firewall opens these application ports only to authenticated or trusted servers. (See Figure 8-4.)

Figure 8-4 Intracluster Security



This security mechanism is applicable only between server nodes in a single Unified CM cluster. Unified CM subscribers are authenticated in a cluster before they can access the publisher's database. The intra-cluster communication and database replication take place only between authenticated servers. During the installation process, a subscriber node is authenticated to the publisher using a pre-shared key authentication mechanism. The authentication process involves the following steps:

1. Install the publisher server using a security password.
2. Configure the subscriber server on the publisher by using Unified CM Administration.
3. Install the subscriber server using the same security password used during publisher server installation.
4. After the subscriber is installed, the server attempts to establish connection to the publisher on a management channel using UDP 8500. The subscriber sends all the credentials to the publisher, such as hostname, IP address, and so forth. The credentials are authenticated using the security password used during the installation process.
5. The publisher verifies the subscriber's credentials using its own security password.
6. The publisher adds the subscriber as a trusted source to its dynamic firewall table if the information is valid. The subscriber is allowed access to the database.
7. The subscriber gets a list of other subscriber servers from the publisher. All the subscribers establish a management channel with each other, thus creating a mesh topology.

General Clustering Guidelines

The following guidelines apply to all Unified CM clusters:

Note A cluster may contain a mix of server platforms, but all servers in the cluster must run the same Unified CM software release.

- Under normal circumstances, place all members of the cluster within the same LAN or MAN.
- If the cluster spans an IP WAN, follow the guidelines for clustering over an IP WAN as specified in the section on Clustering Over the IP WAN, page 5-33.
- A Unified CM cluster may contain as many as 20 servers, of which a maximum of eight call processing subscribers (nodes running the Cisco CallManager Service) are allowed. The other server nodes within the cluster may be configured as a dedicated database publisher, dedicated TFTP subscriber, or media resource subscriber.
- When deploying Unified CM on Cisco MCS 7815, MCS 7816, or equivalent servers, there is a maximum limit of two servers in a deployment: one acting as the publisher, TFTP, and backup call processing subscriber node, and the other acting as the primary call processing subscriber. A maximum of 500 phones is supported in this configuration with a Cisco MCS 7816 or equivalent server.
- When deploying a two-server cluster with high-capacity servers, Cisco recommends that you do not exceed 1250 users in the cluster. Above 1250 users, a dedicated publisher and separate servers for primary and backup call processing subscribers is recommended.
- Business Edition 3000 8.5(1) runs on the MCS 7816 server platform, while Business Edition 3000 8.6(1) and later versions run on either the MCS 7816 or the MCS 7890-C1 purpose-built appliance. In either case Business Edition 3000 provides a single instance of Unified CM (a combined publisher and single subscriber instance). A secondary subscriber instance is not configurable.
- Business Edition 5000 runs on a single hardware platform (MCS 7828), and it provides a single instance of Unified CM (a combined publisher and single subscriber instance). A secondary subscriber instance is not configurable.
- Business Edition 6000 runs on a UCS C200 or C220 Rack-Mount Server and provides a single instance of Unified CM (a combined publisher and single subscriber instance). An additional UCS C200 or C220 server may be deployed to provide subscriber redundancy either in an active/standby or load balancing fashion for Cisco Business Edition call processing as well as other co-resident applications. Cisco recommends deploying redundant servers with load balancing so that the load is distributed between the two UCS servers. Alternatively an MCS server can be used to provide Unified CM subscriber redundancy either in active/standby or load balancing fashion.
- When deploying Unified CM on Cisco UCS B-Series or C-Series Servers, just as with a cluster of MCS servers, each Unified CM node instance can be a publisher node, call processing subscriber node, TFTP subscriber node, or media resource subscriber node. As with any Unified CM cluster, only a single publisher node per cluster is supported.

- While the Cisco UCS B-Series Blade Servers and C-Series Rack-Mount Servers do support a local keyboard, video, and mouse (KVM) cable connection that provides a DB9 serial port, a Video Graphics Array (VGA) monitor port, and two Universal Serial Bus (USB) ports, the Unified CM VMware virtual application has no access to these USB and serial ports. Therefore, there is no ability to attach USB devices such as audio cards (MOH-USB-AUDIO=), serial-to-USB connectors (USB-SERIAL-CA=), or flash drives to these servers. The following alternate options are available:
  - For MoH live audio source feed, consider using Cisco IOS-based gateway multicast MoH for live audio source connectivity or deploying one Unified CM subscriber node on an MCS server as part of the Unified CM cluster to allow connectivity of the USB MoH audio card (MOH-USB-AUDIO=).
  - For SMDI serial connections, deploy one Unified CM subscriber node on an MCS server as part of the Unified CM cluster for USB serial connectivity.
  - For saving system install logs, use virtual floppy softmedia.

High Availability for Call Processing

You should deploy the call processing services within a Unified Communications System in a highly available manner so that a failure of a single call processing component will not render all call processing services unavailable.

Hardware Platform High Availability

You should select the call processing platform based not only on the size and scalability of a particular deployment, but also on the redundant nature of the platform hardware.

For example, for highly available deployments you should select platforms with multiple processors and multiple hard disk drives. Not only is this important for higher-scale deployments, but it is also critical for deployments that require high availability so that an individual component failure does not result in loss of features or services

Furthermore, when possible, choose platforms with dual power supplies to ensure that a single power supply failure will not result in the loss of a platform. See Table 8-2 to determine which platforms support dual power supplies. Plug platforms with dual power supplies into two different power sources to avoid the failure of one power circuit causing the entire platform to fail. The use of dual power supplies combined with the use of uninterruptible power supply (UPS) sources will ensure maximum power availability. In deployments where dual power supply platforms are not feasible, Cisco still recommends the use of a UPS in situations where building power does not have the required level of power availability.

Network Connectivity High Availability

Connectivity to the IP network is also a critical consideration for maximum performance and high availability. Connect call processing platforms to the network at the highest possible speed to ensure maximum throughput, typically 1000 Mbps or 100 Mbps full-duplex depending on the platform. If 1000 or 100 Mbps network access is not available on smaller deployments, then use 10 Mbps full-duplex. Whenever possible, ensure that platforms are connected to the network using full-duplex, which can be achieved with 10 Mbps and 100 Mbps by hard-coding the network switch port and the platform interface port. For 1000 Mbps, Cisco recommends using Auto/Auto for speed and duplex configuration on both the platform interface port and the network switch port.

Note A mismatch will occur if either the platform interface port or the network switch port is left in Auto mode and the other port is configured manually. The best practice is to configure both the platform port and the network switch port manually, with the exception of Gigabit Ethernet ports which should be set to Auto/Auto.

In addition to speed and duplex of IP network connectivity, equally important is the resilience of this network connectivity. Unified communications deployments are highly dependent on the underlying network connectivity for true redundancy. For this reason it is critical to deploy and configure the underlying network infrastructure in a highly resilient manner. For details on designing highly available network infrastructures, see the chapter on Network Infrastructure, page 3-1. In all cases, the network should be designed so that, given a switch or router failure within the infrastructure, a majority of users will have access to a majority of the services provided within the deployment.

To maximize call processing availability, locate and connect call processing platforms in separate buildings and/or separate network switches when possible to ensure that the impact to call processing will be minimized if there is a failure of the building or network infrastructure switch. With Unified CM call processing, this means distributing cluster server nodes among multiple buildings or locations within the LAN or MAN deployment whenever possible. And at the very least, it means physically distributing network connections between different physical network switches in the same location.

Furthermore, even though Cisco Unified CME and Cisco Business Edition are standalone call processing entities, providing physical distribution and therefore redundancy for these call processing types still makes sense when deploying multiple call processing entities. Whenever possible in those scenarios, install each instance of Unified CME or Business Edition in a different physical location within the network, or at the very least physically attach them to different network switches.

Besides deploying a highly available network infrastructure and physically distributing call processing platforms across network components and locations, it is also good practice to provide highly available physical connections to the network from each call processing entity. Whenever possible, use dual network attachments to connect the platform to two different ports on two physically separate network switches so that a single upstream hardware port or switch failure will not result in loss of network connectivity for the platform. A Unified CME router platform can have more than one physical network interface and can be dual-attached to a network. Likewise, the MCS server platform for Unified CM and Business Edition 5000 call processing types can also be dual-attached to the network using network interface card (NIC) teaming.

NIC Teaming for Network Fault Tolerance

The NIC teaming feature allows a Cisco MCS (or HP or IBM equivalent server) to be connected to the IP network through two NICs and, therefore, two physical cables. NIC teaming prevents network downtime by transferring the workload from the failed port to the working port. NIC teaming cannot be used for load balancing or increasing the interface speed. NIC teaming is supported on dual-NIC Cisco MCS platforms (or HP or IBM equivalents).

Note The MCS 7815 platform (or HP or IBM equivalent) has only a single network interface port and therefore cannot perform NIC teaming.

UCS Network Fault Tolerance

Cisco UCS B-Series Blade Servers leverage the UCS network attachment infrastructure as well as the underlying network attached storage area network (SAN). This back-end UCS network infrastructure, including redundant parallel switching fabric extenders and interconnects as well as Fibre Channel or gigabit ethernet uplinks, provides highly available network attachment and storage for these servers. For

details on highly available virtual data center deployments of the UCS network and storage infrastructure, refer to the document on *Designing Secure Multi-Tenancy into Virtualized Data Centers*, available at
http://www.cisco.com/en/US/docs/solutions/Enterprise/Data_Center/Virtualization/securecldg.html.

Unified CM High Availability

Because of the underlying Unified CM clustering mechanism, a Unified Communications System has additional high availability considerations above and beyond hardware platform disk and power component redundancy, physical network location, and connectivity redundancy. This section examines call processing subscriber redundancy considerations, call processing load balancing, and redundancy of additional cluster services.

Call Processing Redundancy

Unified CM provides the following call processing redundancy configuration options or schemes:

- Two to one (2:1) — For every two primary call processing subscribers, there is one shared secondary or backup call processing subscriber.
- One to one (1:1) — For every primary call processing subscriber, there is a secondary or backup call processing subscriber.

These redundancy schemes are facilitated by the built-in registration failover mechanism within the Unified CM cluster architecture, which enables endpoints to re-register to a backup call processing subscriber node when the endpoint's primary call processing subscriber node fails. The registration failover mechanism can achieve failover rates for Skinny Client Control Protocol (SCCP) IP phones of approximately 125 registrations per second. The registration failover rate for Session Initiation Protocol (SIP) phones is approximately 40 registrations per second.

The call processing redundancy scheme you select determines not only the fault tolerance of the deployment, but also the fault tolerance of any upgrade.

With 1:1 redundancy, multiple primary call processing subscriber failures can occur without impacting call processing capabilities. With 2:1 redundancy, on the other hand, only one of the primary call processing subscribers out of the two primary call processing subscribers that share a backup call processing subscriber can fail without impacting call processing. If the total number of endpoints registered across both primary subscribers and the traffic to those two primary subscribers are within the capacity limits of a single subscriber, then the backup subscriber is able to handle the failure of both primary subscribers.

Note Do not deploy 2:1 redundancy if the total capacity utilization across the two primary subscribers would exceed the capacity of the backup subscriber. For example, if the call processing capacity or endpoints capacity utilization exceeds 50% on both primary subscribers, the backup subscriber would not be able to handle call processing services properly if both primary subscribers fail. In these scenarios, for example, some endpoints might not be able to register, some new calls might not be established, and some services and features might not operate properly because the backup subscriber system capacity has been exceeded.

Likewise, with the 1:1 redundancy scheme, upgrades to the cluster can be performed with only a single set of endpoint registration failover periods impacting the call processing services. Whereas with the 2:1 redundancy scheme, upgrades to the cluster can require multiple registration failover periods.

A Unified CM cluster can be upgraded with minimal impact to the services. Two different versions (releases) of Unified CM may be on the same server, one in the active partition and the other in the inactive partition. All services and devices use the Unified CM version in the active partition for all Unified CM functionality. During the upgrade process, the cluster operations continue using its current release of Unified CM in the active partition, while the upgrade version gets installed in the inactive partition. Once the upgrade process is complete, the servers can be rebooted to switch the inactive partition to the active partition, thus running the new version of Unified CM.

With the 1:1 redundancy scheme, the following steps enable you to upgrade the cluster while minimizing downtime:

Step 1 Install the new version of Unified CM in the inactive partition, first on the publisher and then on all subscribers (call processing, TFTP, and media resource subscribers). Do not reboot.

Step 2 Reboot the publisher and switch to the new version.

Step 3 Reboot the TFTP subscriber node(s) one at a time and switch to the new version.

Step 4 Reboot any dedicated media resource subscriber nodes one at a time and switch to the new version.

Step 5 Reboot the backup call processing subscribers one at a time and switch to the new version.

Step 6 Reboot the primary call processing subscribers one at a time and switch to the new version. Device registrations will fail-over to the previously upgraded and rebooted backup call processing subscribers. After each primary call processing subscriber is rebooted, devices will begin to re-register to the primary call processing subscriber.

With this upgrade method, there is no period (except for the registration failover period) when devices are registered to subscriber servers that are running different versions of the Unified CM software.

While the 2:1 redundancy scheme allows for fewer servers in a cluster, registration failover occurs more frequently during upgrades, increasing the overall duration of the upgrade as well as the amount of time call processing services for a particular endpoint will be unavailable. Because there is only a single backup call processing subscriber per pair of primary call processing subscribers, it might be possible to reboot to the new version on only one of the primary call processing subscribers in a pair at a time in order to prevent oversubscribing the single backup call processing subscriber. As a result, there may be a period of time after the first primary call processing subscriber in each pair is switched to the new version, in which endpoint registrations will have to be moved from the backup subscriber to the newly upgraded primary subscriber before the endpoint registrations on the second primary subscriber can be moved to the backup subscriber to allow a reboot to the new version. During this time, not only will endpoints on the second primary call processing subscriber be unavailable while they re-register to the backup subscriber, but until they re-register to a node running the new version, they will also be unable to reach endpoints on other subscriber nodes that have already been upgraded.

Note Before you do an upgrade, Cisco recommends that you back up the Unified CM and Call Detail Record (CDR) database to an external network directory using the Disaster Recovery Framework. This practice will prevent any loss of data if the upgrade fails.

Note Because an upgrade of a Unified CM cluster results in a period of time in which some or most devices lose registration and call processing services temporarily, you should plan upgrades in advance and implement them during a scheduled maintenance window. While downtime and loss of services to devices can be minimized by selecting the 1:1 redundancy scheme, there will still be some period of time in which call processing services are not available to some or all users.

For more information on upgrading Unified CM, refer to the install and upgrade guides available at

http://www.cisco.com/en/US/products/sw/voicesw/ps556/prod_installation_guides_list.html

Unified CM Redundancy with Survivable Remote Site Telephony (SRST)

Cisco IOS SRST provides highly available call processing services for endpoints in locations remote from the Unified CM cluster. Unified CM clustering redundancy schemes certainly provide a high level of redundancy for call processing and other application services within a LAN or MAN environment. However, for remote locations separated from the central Unified CM cluster by a WAN or other low-speed links, SRST can be used as a redundancy method to provide basic call processing services to these remote locations in the event of loss of network connectivity between the remote and central sites. Cisco recommends deploying SRST-capable Cisco IOS routers at each remote site where call processing services are considered critical and need to be maintained in the event that connectivity to the Unified CM cluster is lost. Endpoints at these remote locations must be configured with an appropriate SRST reference within Unified CM so that the endpoint knows what address to use to connect to the SRST router for call processing services when connectivity to Unified CM subscribers is unavailable.

Unified CME on a Cisco IOS router can also be used at a remote site to provide enhanced SRST functionality in the event that connectivity to the central Unified CM cluster is lost. Unified CME provides more backup call processing features for the IP phones than are available with the regular SRST feature on a router. However, the endpoint capacities for Unified CME acting as SRST are typically less than for basic SRST.

Call Processing Subscriber Redundancy

Depending on the redundancy scheme chosen (see Call Processing Redundancy, page 8-16), the call processing subscriber will be either a primary (active) subscriber or a backup (standby) subscriber. In the load-balancing option, the subscriber can be both a primary and backup subscriber. When planning the design of a cluster, you should generally dedicate the call processing subscribers to this function. In larger-scale or higher-performance clusters, the call processing service should not be enabled on the publisher and TFTP subscriber nodes. 1:1 redundancy uses dedicated pairs of primary and backup subscribers, while 2:1 redundancy uses a pair of primary subscribers that share one backup subscriber.

The following figures illustrate typical cluster configurations to provide call processing redundancy with Unified CM.

Figure 8-5 Basic Redundancy Schemes

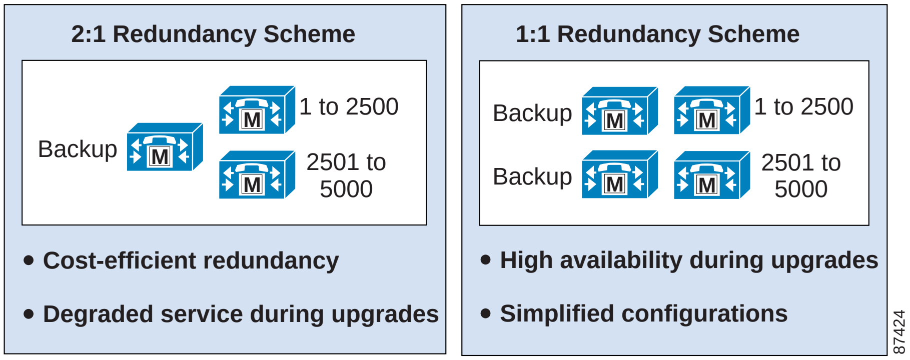

2:1 Redundancy Scheme
Backup
M
1 to 2500
2501 to 5000
• Cost-efficient redundancy
• Degraded service during upgrades
1:1 Redundancy Scheme
Backup
1 to 2500
Backup
2501 to 5000
• High availability during upgrades
• Simplified configurations
87424

Figure 8-5 illustrates the two basic redundancy schemes available. In each case the backup server must be capable of handling the capacity of at least a single primary call processing server failure. In the 2:1 redundancy scheme, the backup might have to be capable of handling the failure of a single call processing server or potentially both primary call processing servers, depending on the requirements of a particular deployment. For information on sizing the capacity of the servers and choosing the hardware platforms, see the section on Capacity Planning for Call Processing, page 8-24.

Note 2:1 redundancy is not supported when using the Cisco MCS 7845-I3 server or a virtual machine deployed with the 10K-User Open Virtualization Archive (OVA) template due to potential overload on the backup subscriber.

Figure 8-6 1:1 Redundancy Configuration Options

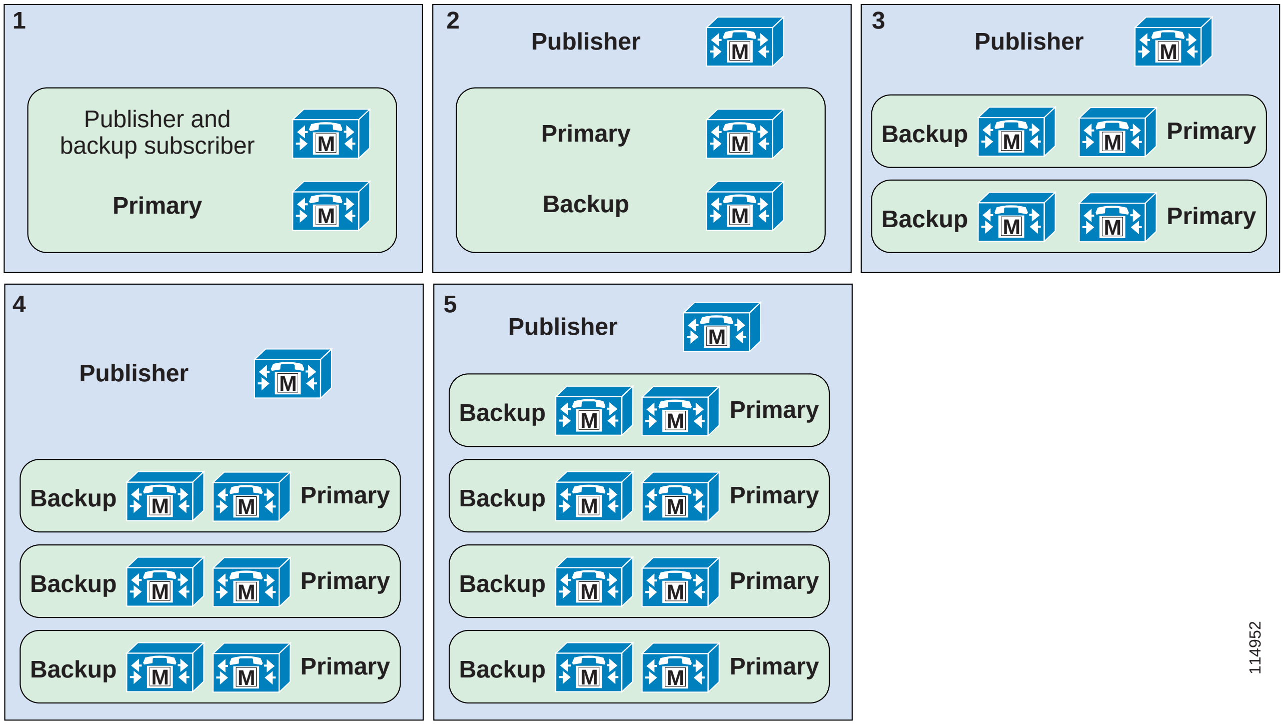
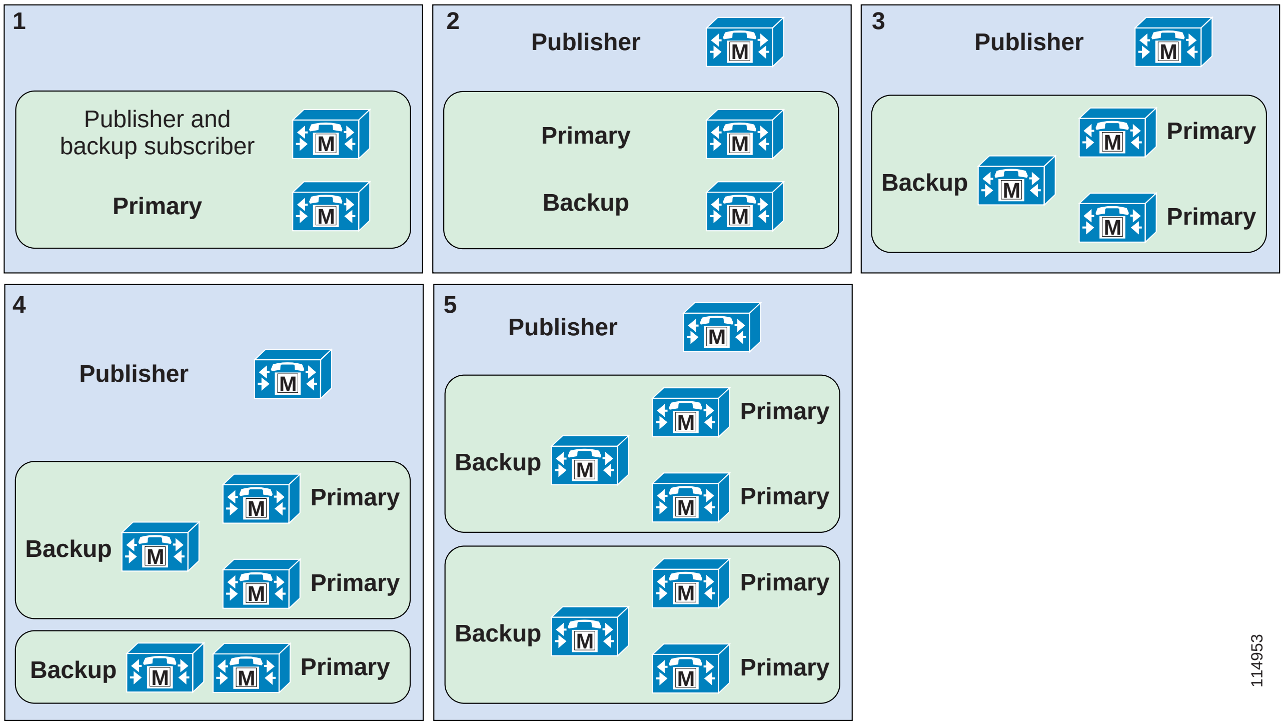


Figure 8-7 2:1 Redundancy Configuration Options



In Figure 8-6, the five options shown all indicate 1:1 redundancy. In Figure 8-7, the five options shown all indicate 2:1 redundancy. In both cases, Option 1 is used for clusters supporting less than 1250 users. Options 2 through 5 illustrate increasingly scalable clusters for each redundancy scheme. The exact scale depends on the hardware platforms chosen or required.

These illustrations show only publisher and call processing subscribers. They do not account for other subscriber nodes such as TFTP and media resources.

Note It is possible to define up to three call processing subscribers per Unified CM group. Adding a tertiary subscriber for additional backup extends the above redundancy schemes to 2:1:1 or 1:1:1 redundancy. However, with the exception of using tertiary subscriber servers in deployments with clustering over the WAN (see Remote Failover Deployment Model, page 5-43), tertiary subscriber redundancy is not recommended for endpoint devices located in remote sites because failover to SRST will be further delayed if the endpoint must check for connectivity to a tertiary subscriber. The tertiary subscribers also count against the maximum number of call processing subscribers in a cluster (8 call processing subscriber nodes).

Although not shown in the Figure 8-6 or Figure 8-7, it is also possible to deploy a single-server cluster with an MCS 7825 or larger server. With an MCS 7825 or equivalent server, the endpoint configuration and registration limit is 500 for a single-server cluster. With a higher-availability server, the single-server cluster should not exceed 1000 endpoint configuration and registrations. Note that in a single-server configuration, there is no backup call processing subscriber and therefore no cluster redundancy mechanism. Survivable Remote Site Telephony (SRST) can be used as a redundancy mechanism in these types of deployments to provide minimal call processing services during periods when Unified CM is not available. However, Cisco does not recommend a single-server deployment for production environments.

Load Balancing

In Unified CM clusters with the 1:1 redundancy scheme, device registration and call processing services can be load-balanced across the primary and backup call processing subscriber.

Normally a backup server has no devices registered to it unless its primary is unavailable. This makes it easier to troubleshoot a deployment because there is a maximum of four primary call processing subscriber nodes that will be handling the call processing load at a given time. Further, this potentially simplifies configuration by reducing the number of Unified CM redundancy groups and device pools.

In a load-balanced deployment, up to half of the device registration and call processing load can be moved from the primary to the secondary subscriber by using the Unified CM redundancy groups and device pool settings. In this way each primary and backup call processing subscriber pair provides device registration and call processing services to as many as half of the total devices serviced by this pair of call processing subscribers. This is referred to as 50/50 load balancing. The 50/50 load balancing model provides the following benefits:

- Load sharing — The registration and call processing load is distributed on multiple servers, which can provide faster response time.
- Faster failover and failback — Because all devices (such as IP phones, CTI ports, gateways, trunks, voicemail ports, and so forth) are distributed across all active subscribers, only some of the devices fail-over to the secondary subscriber if the primary subscriber fails. In this way, you can reduce by 50% the impact of any server becoming unavailable.

To plan for 50/50 load balancing, calculate the capacity of a cluster without load balancing, and then distribute the load across the primary and backup subscribers based on devices and call volume. To allow for failure of the primary or the backup server, do not let the total load on the primary and secondary subscribers exceed that of a single subscriber server.

Note During upgrades of a Unified CM cluster with 50/50 load balancing, upgrades to the backup call processing subscriber will result in devices registered to that subscriber (up to half of the total devices serviced by the primary and backup subscriber pair) failing over to the primary call processing subscriber.

TFTP Redundancy

Cisco recommends deploying more than one dedicated TFTP subscriber node for a large Unified CM cluster, thus providing redundancy for TFTP services. While two TFTP subscribers are typically sufficient, more than two TFTP servers can be deployed in a cluster.

In addition to providing one or more redundant TFTP subscribers, you must configure endpoints to take advantage of these redundant TFTP nodes. When configuring the TFTP options using DHCP or statically, define a TFTP subscriber node IP address array containing the IP addresses of both TFTP subscriber nodes within the cluster. In this way, by creating two DHCP scopes with two different IP address arrays (or by manually configuring endpoints with two different TFTP subscriber node IP addresses), you can assign half of the endpoint devices to use TFTP subscriber A as the primary and TFTP subscriber B as the backup, and the other half to use TFTP subscriber B as the primary and TFTP subscriber A as the backup. In addition to providing redundancy during a failure of one TFTP subscriber, this method of distributing endpoints across multiple TFTP subscribers provides load balancing so that one TFTP subscriber is not handling all the TFTP service load.

Note When adding a specific binary or firmware load for a phone or gateway, you must add the file(s) to each TFTP subscriber node in the cluster.

CTI Manager Redundancy

All CTI integrated applications communicate with a call processing subscriber node running the CTI Manager service. Further, most CTI applications have the ability to specify redundant CTI Manager service nodes. For this reason, Cisco recommends activating the CTI Manager service on at least two call processing subscribers within the cluster. With both a primary and backup CTI Manager configured, in the event of a failure the application will switch to a backup CTI Manager to receive CTI services.

As stated previously, the CTI Manager service can be enabled only on call processing subscribers, therefore there is a maximum of eight CTI Managers per cluster. Cisco recommends that you load-balance CTI applications across the enabled CTI Managers in the cluster to provide maximum resilience, performance, and redundancy.

Generally, it is good practice to associate devices that will be controlled or monitored by a CTI application with the same server pair used for the CTI Manager service. For example, an interactive voice response (IVR) application requires four CTI ports. They would be provisioned as follows, assuming the use of 1:1 redundancy and 50/50 load balancing:

- Two CTI ports would have a Unified CM redundancy group of server A as the primary call processing subscriber and server B as the backup subscriber. The other two ports would have a Unified CM redundancy group of server B as the primary subscriber and server A as the backup subscriber.
- The IVR application would be configured to use the CTI Manager on subscriber A as the primary and subscriber B as the backup.

The above example allows for redundancy in case of failure of the CTI Manager on subscriber A and also allows for the IVR call load to be spread across two servers. This approach also minimizes the impact of a Unified CM subscriber node failure.

For more details on CTI and CTI Manager, see Computer Telephony Integration (CTI), page 8-30.

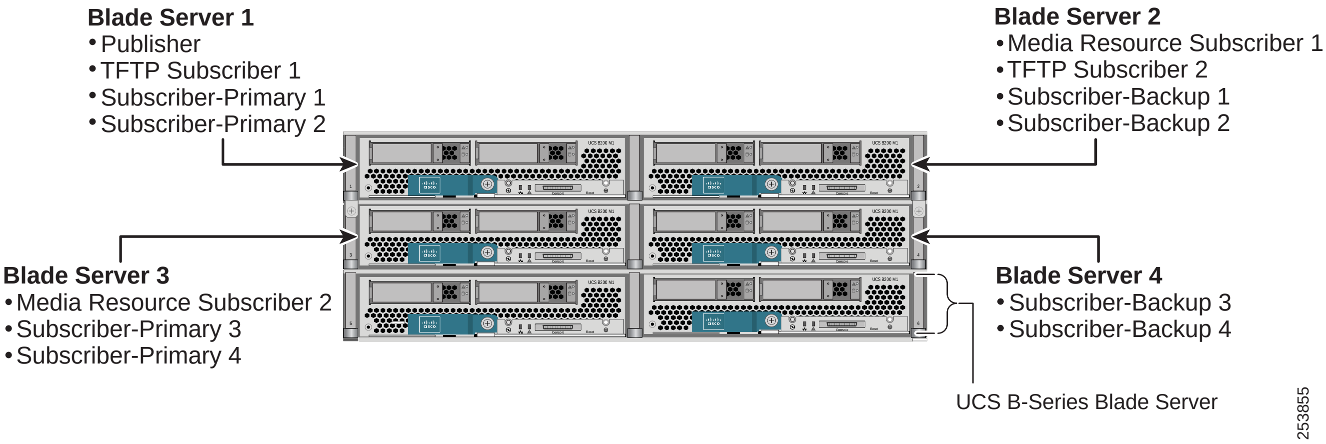
UCS Call Processing Redundancy with Virtualized Platforms

For deployments of Unified CM as a virtualized application running on Cisco UCS B-Series Blade Servers, C-Series Rack-Mount Servers, or third-party servers, all previous call processing, TFTP, and CTI Manager redundancy schemes still apply.

As illustrated in Figure 8-8, observe the following guidelines when deploying Unified CM as a virtualized application to ensure the highest level of call processing redundancy:

- Each primary call processing subscriber node instance should reside on a different physical UCS B-Series or C-Series server than its backup call processing subscriber node instance. This ensures that the failure of a server containing the primary call processing node instance does not impact the system's ability to provide endpoints with access to their backup call processing subscriber node.
- When deploying multiple TFTP or media resource subscriber nodes instances for redundancy of those services, always distribute redundant subscriber nodes across more than one UCS B-Series or C-Series server to ensure that a failure of a single server does not eliminate those services. This ensures that, given the failure of a server containing a TFTP or media resource subscriber, endpoints will still be able to access TFTP and media resource services on a subscriber node residing on another server. Endpoints can also be distributed among redundant TFTP and media resource subscriber node instances to balance system load in non-failure scenarios.
- When deploying CTI applications, always make sure that call processing subscriber node instances running the CTI Manager service are distributed across more than one UCS B-Series or C-Series server to ensure that a failure of a single server does not eliminate CTI services. Further, CTI applications should be configured to use the CTI Manager service running on the subscriber node instance on one server as the primary CTI Manager and the CTI Manager service running on the subscriber node on another server as the backup CTI Manager.

Figure 8-8 Unified CM Server Node Distribution on UCS



In addition to distributing subscriber node instances across multiple blades, when using blade servers you may distribute subscriber node instances across multiple blade chassis for additional redundancy and scalability.

For more information about redundancy and provisioning of host resources for virtual machines, refer to the documentation at http://www.cisco.com/go/uc-virtualized.

Cisco Business Edition High Availability

The main considerations for high availability of Cisco Business Edition are network connectivity, power, and redundancy for call processing and registration.

As shown in Table 8-2, both the MCS 7816 platform used for Business Edition 3000 and the MCS 7828 platform used for Business Edition 5000 have dual IP interfaces or NICs for redundant network attachment. However, only Business Edition 5000 supports NIC Teaming for network connectivity redundancy. Business Edition 3000 installed on an MCS 7816 server does not support NIC Teaming. The MCS 7980-C1 purpose-built appliance for Business Edition 3000 has only a single IP interface and therefore does not provide support for redundant network attachment or NIC Teaming.

Business Edition 3000 and Business Edition 5000 each reside on their own single standalone platforms (a combined publisher and single subscriber instance with no ability to configure a secondary subscriber instance). They do not support node clustering and therefore cannot leverage the call processing redundancy schemes available with Unified CM. For this reason, the only way to provide call processing and registration redundancy for endpoints in these types of deployments is by using SRST or Unified CME acting as SRST. However, only Business Edition 5000 supports SRST. There is not ability to provide highly available call processing and registration with Business Edition 3000.

On the other hand, Business Edition 6000 does provide redundancy for call processing and registration services by clustering additional Cisco Unified CM nodes. A second Business Edition 6000 server (UCS C200 or C220 Rack-Mount Server or MCS server) can be deployed to provide high availability for call processing as well as other applications and services.

Note More than two UCS C200 or C220 Rack-Mount Servers may be clustered for a Business Edition 6000 deployment to provide additional redundancy and/or geographic distribution as with a clustering over the WAN deployment. However, the total number of users across the cluster may not exceed 1,000 and the total number of configured devices across the cluster may not exceed 1,200. A deployment of UCS C200 or C220 Rack-Mount Servers in a cluster exceeding 1,000 users and 1,200 configured devices is considered a regular Unified CM cluster, and as such the deployment must follow high availability design guidance for regular Unified CM. (See Unified CM High Availability, page 8-16.)

Capacity Planning for Call Processing

Call processing capacity planning is critical for successful unified communications deployments. Given the many features and functions provided by call processing services as well as the many types of devices for which call processing entities can provide registration and transaction services, it important to size the call processing infrastructure and its individual components to ensure they meet the capacity needs of a particular deployment.

IP phones, software clients, voicemail ports, CTI (TAPI or JTAPI) devices, gateways, and DSP resources for media services such as transcoding and conferencing, all register to a call processing entity. Each of these devices requires resources from the call processing platform with which it is registered. The required resources can include memory, processor usage, and disk I/O.

Besides adding registration load to call processing platforms, after registration each device then consumes additional platform resources during transactions, which are normally in the form of calls. For example, a device that makes only 6 calls per hour consumes fewer resources than a device making 12 calls per hour.

For more information about call processing sizing and for a complete discussion of system sizing, capacity planning, and deployment considerations, see the chapter on Unified Communications Design and Deployment Sizing Considerations, page 29-1.

Unified CME Capacity Planning

When deploying Unified CME, it is critical to select a Cisco IOS router platform that provides the desired capacity in terms of number of supported endpoints required. In addition, platform memory capacity should also be considered if the Unified CME router is providing additional services above and beyond call processing, such as IP routing, DNS lookup, dynamic host configuration protocol (DHCP) address services, or VXML scripting.

Unified CME can support a maximum of 450 endpoints on a single Cisco IOS platform; however, each router platform has a different endpoint capacity based on the size of the system. Because Unified CME is not supported within the Cisco Unified Communications Sizing Tool, it is imperative to follow capacity information provided in the product data sheets available at

http://www.cisco.com/en/US/products/sw/voicesw/ps4625/products_data_sheets_list.html

Unified CM Capacity Planning

This section examines capacity planning for Unified CM. The recommendations provided in this section are based on calculations made using the Unified Communications Sizing Tool, with default trace levels and call detail records (CDRs) enabled. In some cases higher levels of performance and capacity can be achieved by disabling, reducing, or reconfiguring other functions that are not directly related to processing calls. Enabling and increasing utilization of these functions can also have an impact on the call processing capabilities of the system and in some cases can reduce the overall capacity. These functions include tracing, call detail recording, highly complex dial plans, and other services that are co-resident on the Unified CM platform. Highly complex dial plans can include multiple line appearances as well as large numbers of partitions, calling search spaces, route patterns, translations, route groups, hunt groups, pickup groups, route lists, call forwarding, co-resident services, and other co-resident applications. All of these functions can consume additional resources within the Unified CM system.

You can use the following techniques to improve system performance:

- Install additional certified memory in the server, up to the maximum supported for the particular platform. Cisco recommends doubling the RAM in MCS 7825 and MCS 7835 or equivalent servers with large configurations for that server class. Verification using the Cisco Real Time Monitoring Tool (RTMT) will indicate if this memory upgrade is required. As the server approaches maximum utilization of physical memory, the operating system will start to swap to disk. This swapping is a sign that additional physical memory should be installed.
- A Unified CM cluster with a very large dial plan containing many gateways, route patterns, translation patterns, and partitions, can take an extended amount of time to initialize when the Cisco CallManager Service is first started. If the system does not initialize within the default time, you can modify the system initialization timer (a Unified CM service parameter) to allow additional time for the configuration to initialize. For details on the system initialization time, refer to the online help for Service Parameters in Unified CM Administration.

Unified CM Capacity Planning with Virtualized Platforms

In a virtualized deployment, most Unified Communications applications such as Unified CM must be installed using a predefined template that specifies the configuration of the virtual machine's virtual hardware. These templates are distributed through Open Virtualization Archives (OVA), an open standards-based method for packaging and distributing virtual machine templates.

These OVA templates define the number of virtual CPU, the amount of virtual memory, the number and size of hard drives, and so forth, and they determine the capacity of the application. For Unified CM, there are multiple OVA templates available, one for almost each server class (although there is no template corresponding to the MCS 7815 or MCS 7816). A Unified CM virtual machine instance running on a VMware or Cisco UCS server typically has the same capacity as a Unified CM node running directly on a Cisco MCS server when using the corresponding OVA template. For example, the OVA template for Unified CM supporting 7,500 users and/or devices has the same capacity as the MCS 7845-H2/I2 server.

Unified CM Capacity Planning Guidelines and Endpoint Limits

The following capacity guidelines apply to Cisco Unified CM:

- Within a cluster, a maximum of 8 call processing subscriber nodes can be enabled with the Cisco CallManager Service. Other servers may be used for more dedicated functions such as publisher, TFTP subscribers, and media resources subscribers.
- Each cluster can support configuration and registration for a maximum of:
  - 40,000 secured or unsecured SCCP or SIP endpoints with Unified CM 8.6(1) and later releases
  - 30,000 secured or unsecured SCCP or SIP endpoints with Unified CM 8.5 and earlier releases.
- A cluster consisting of server node instances running on VMware can support different capacities depending on the OVA template that is chosen. For most Cisco MCS server classes, there is a corresponding OVA template that provides a Unified CM instance with the same capacities (number of phones, gateways, locations, regions, CTI connections, and so forth) as the MCS server class. Because multiple virtual machine instances can run on the same blade or server, the total capacity on a blade or server can therefore be higher than on an MCS server.
- The maximum recommended trace setting for Unified CM is 2,000 files of 2 MB for both System Diagnostic Interface (SDI) and Signaling Distribution Layer (SDL) traces, for a total of 4,000 files. Each process has a setting for maximum number of files, and each process is allowed 2,000 files for SDL and 2,000 files for SDI. Trace settings for all other components must be configured within the limit of 126 MB (for example, 63 files of 2 MB each). These are suggested upper limits. Unless specific troubleshooting under high call rates requires increasing the maximum file setting, the default settings are sufficient for collecting sufficient traces in most circumstances.

For more information about Unified CM capacity planning considerations, including sizing limits as well as a complete discussion of system sizing, capacity planning, and deployment considerations, see the chapter on Unified Communications Design and Deployment Sizing Considerations, page 29-1.

Megacluster

The term *megacluster* defines and identifies certain Unified CM deployments that allow for further increases in scalability. A megacluster provides more device capacity through the support of additional Unified CM subscriber nodes, with a maximum of eight Unified CM subscriber pairs (1:1 redundancy) per megacluster, thus allowing for a maximum of 80,000 devices with Cisco Unified CM 8.6 and later releases.

A megacluster can also be deployed where customers simply require non-locally redundant call processing functionality, rather than using Survivable Remote Site Telephony (SRST), to scale beyond the maximum eight sites allowed in a standard cluster deployment and up to 16 Unified CM subscriber nodes per megacluster. For example, consider a large hospital that has twelve locations and each location has only 1,000 devices. This total of 12,000 devices could be accommodated within a standard cluster, which has a maximum device capacity of 40,000 devices. However, in this case it is the need for additional Unified CM subscribers, rather than additional device capacity, that requires a megacluster deployment. In this example, a Unified CM subscriber node could be deployed in each location, and each Unified CM subscriber could serve as the primary subscriber for the local endpoints and as a backup subscriber for endpoints from another location.

When considering a megacluster deployment, the primary areas impacting capacity are as follows:

- The megacluster may contain a total of 21 servers consisting of 16 subscribers, 2 TFTP servers, 2 music on hold (MoH) servers, and 1 publisher
- Server type must be either Cisco MCS 7845-I3/H3 class or Cisco Unified Computing System (UCS) C-Series or B-Series using the 10K Open Virtualization Archive (OVA) template.
- Redundancy model must be 1:1.

All other capacities relating to a standard cluster also apply to a megacluster. Note that support for a megacluster deployment is granted only following the successful review of a detailed design, including the submission of the results from the Cisco Unified Communications Sizing Tool. For more information about the Cisco Unified Communications Sizing Tool and the sizing of Unified CM standard clusters and megaclusters, see the chapter on Unified Communications Design and Deployment Sizing Considerations, page 29-1.

Due to the many potential complexities surrounding megacluster deployments, customers who wish to pursue such a deployment must engage either their Cisco Account Team or their certified Cisco Unified Communications Partner.

Note Unless otherwise specified, all information contained within this SRND that relates to call processing deployments (including capacity, high availability, and general design considerations) applies only to a standard cluster.

Cisco Business Edition Capacity Planning

Just as with Unified CM, many types of devices can register with Cisco Business Edition, and each of these devices requires registration and transaction resources from the platform with which it is registered. Likewise the users and their busy hour call attempts (BHCA) consume additional system resources. Each Cisco Business Edition system has specific user, endpoint, and BHCA capacity thresholds based on the available system resources of the platform. The maximum number of users and endpoints supported by Cisco Business Edition are 1,000 and 1,200 respectively. The maximum BHCA supported by Cisco Business Edition is 5,000.

For more information about Cisco Business Edition capacity planning considerations, including sizing examples and per-platform sizing limits as well as a complete discussion of system sizing, capacity planning, and deployment considerations, see the chapter on Unified Communications Design and Deployment Sizing Considerations, page 29-1.

For additional information on Cisco Business Edition capacities as well as all other Cisco Business Edition product information, refer to the following product documentation:

- Cisco Business Edition 3000

  http://www.cisco.com/en/US/products/ps11370/tsd_products_support_series_home.html
- Cisco Business Edition 5000

  http://www.cisco.com/en/US/products/ps7273/tsd_products_support_series_home.html
- Cisco Business Edition 6000

  http://docwiki.cisco.com/wiki/Cisco_Unified_Communications_Manager_Business_Edition_6000

Design Considerations for Call Processing

Observe the following design recommendations and guidelines when deploying Cisco call processing:

Cisco Unified CME

- Unified CME supports a maximum of 450 endpoints. However, depending on the Cisco IOS router model, endpoint capacity could be significantly lower. For additional information about Unified CME platforms and capacities, refer to the Cisco Unified Communications Manager Express compatibility information available at http://www.cisco.com/en/US/products/sw/voicesw/ps4625/products_device_support_tables_list.html.
- When possible, dual-attach the Unified CME router to the network using multiple IP interfaces to provide maximum network availability. Likewise, if multiple instances of Unified CME are required in the same deployment, distribute them across multiple physical switches or locations.
- When possible, deploy the Unified CME router with dual power supplies and/or an uninterruptible power supply (UPS) in order to provide maximum availability of the platform.

Cisco Business Edition

- Business Edition 3000 runs on either the MCS 7816 or the MCS 7890-C1 (with version 8.6(1) and later) acting as a combined publisher and single subscriber instance. A secondary subscriber instance is not configurable.
- Business Edition 5000 runs on a single hardware platform (MCS 7828) acting as a combined publisher and single subscriber instance. A secondary subscriber instance is not configurable.
- Business Edition 6000 runs on a UCS C200 or C220 Rack-Mount Server acting as a combined publisher and single subscriber instance. A second UCS C200 or C220 server can be deployed to provide call processing redundancy by means of a secondary subscriber. Alternatively an MCS server can be used to provide redundancy.

Note More than two UCS C200 or C220 Rack-Mount Servers may be clustered for a Business Edition 6000 deployment to provide additional redundancy and/or geographic distribution. However, the total number of users across the cluster may not exceed 1,000 and the total number of configured devices across the cluster may not exceed 1,200.

- Business Edition 6000 supports a maximum of 1,200 endpoints. However, actual endpoint capacity depends on total system BHCA, which cannot exceed a maximum of 5,000. For additional information about Cisco Business Edition capacity, including sizing examples and per-platform sizing limits, see the chapter on Unified Communications Design and Deployment Sizing Considerations, page 29-1.
- Dual-attach the MCS 7828 server for Business Edition 5000 to the network using NIC teaming to provide maximum high availability. Business Edition 3000 does not support NIC teaming.
- If multiple instances of Business Edition 5000 or Business Edition 6000 are required in the same deployment, distribute them across multiple physical switches.
- Because some of the Cisco Business Edition platforms (MCS 7816, MCS 7828, MCS 7890-C1, and UCS C200) do not have or support dual power supplies, use an uninterruptible power supply (UPS) to provide maximum availability of those platforms.
- When deploying Business Edition 6000 with two servers for high availability (two UCS C200/C220 Rack-Mount Servers, or one UCS C200/C220 Rack-Mount Server and one MCS server), device registration should be load-balanced between the two servers in order to distribute system load. This is preferable to using the second server for standby redundancy.
- Business Edition 3000 provides support only for very specific types of endpoints and gateways:
  - Business Edition 3000 supports a limited set of endpoints. For a list of supported endpoints, refer to the *Administration Guide for Cisco Business Edition 3000*, available at

    http://www.cisco.com/en/US/products/ps11370/prod_maintenance_guides_list.html
  - Business Edition 3000 PSTN connectivity is supported only through the Cisco 2901 Integrated Services Router (ISR) and only with MGCP backhauled T1/E1 PRI trunks.
  - Business Edition 3000 does not support intercluster trunking and therefore does not support distributed call processing deployments.

Cisco Unified CM

- You can enable a maximum of 8 call processing subscriber nodes (nodes running the Cisco CallManager Service) within a Cisco Unified CM cluster. Additional servers may be dedicated and used for publisher, TFTP, and media resources services. An approved megacluster deployment supports a maximum of 16 call processing subscriber nodes.
- Each Unified CM cluster can support configuration and registration for a maximum of 40,000 secured or unsecured endpoints with Unified CM 8.6(1) and later releases. For Unified CM 8.5 and earlier releases, a maximum of 30,000 secured or unsecured endpoints is supported. For additional information about Unified CM capacity planning, including per-platform sizing limits, see the chapter on Unified Communications Design and Deployment Sizing Considerations, page 29-1.
- When deploying a two-server cluster with high-capacity servers, Cisco recommends that you do not exceed 1250 users in the cluster. Above 1250 users, Cisco recommends a dedicated publisher and separate servers for primary and backup call processing subscribers.
- Cisco recommends using the same server model for all servers in a cluster. However, mixing server models and even different server vendor models within a cluster is supported, provided that all of the individual hardware versions are supported and that all servers are running the same version of Unified CM.
- 2:1 redundancy is not supported when using the Cisco MCS 7845-I3 or the 10K-User Open Virtualization Archive (OVA) template due to potential overload on the backup subscriber.
- Dual-attach MCS servers to the network using NIC teaming to provide maximum high availability. The MCS 7815 has only a single network interface port and therefore cannot perform NIC teaming.

- Whenever possible, distribute the Unified CM servers across multiple physical switches within the network and across multiple physical locations within the same network to minimize the impact of a switch failure or the loss of a particular network location.
- Deploy SRST or Unified CME acting as SRST on Cisco IOS routers at remote locations to provide fallback call processing services in the event that these locations lose connectivity to the Unified CM cluster.
- Cisco recommends leaving voice activity detection (VAD) disabled in the Unified CM cluster. You should also disable VAD on Cisco IOS H.323 and SIP dial peers by using the **no vad** command.
- When deploying Unified CM as a virtualized application, ensure that server node instances are distributed across rack-mount servers or blades servers within the UCS chassis so that backup or redundant subscriber nodes are on different physical servers than primary subscriber nodes.
- Both UCS B-Series Blade Servers and high-end C-Series Rack-Mount Servers (for example, C210, C240, and C260) can be configured with multiple Open Virtualization Archive (OVA) templates. The largest OVA template supports 10,000 devices and provides the same capacities (number of endpoints, gateways, locations, regions, and so forth) as an MCS 7845-I3 server. For information on proper OVA sizing as well as the use of the Cisco Unified Communications Sizing Tool, see the chapter on Unified Communications Design and Deployment Sizing Considerations, page 29-1.
- While the UCS B-Series Blade Servers and C-Series Rack-Mount Servers do support USB and serial ports through a KVM cable, the Unified CM VMware virtual application has no access to those ports. Therefore, if you deploy Unified CM on UCS, it will not be possible to attach fixed live audio sources for MoH, to make a serial SMDI connection to a legacy voicemail system, or to attach a USB flash drive for writing log files. The following alternate options are available:
  - For MoH live audio source feed, consider using Cisco IOS-based gateway multicast MoH for live audio source connectivity or deploying one Unified CM subscriber node on an MCS server as part of the Unified CM cluster to allow connectivity of the USB MoH audio card (MOH-USB-AUDIO=).
  - For SMDI serial connections, deploy one Unified CM subscriber node on an MCS server as part of the Unified CM cluster for USB serial connectivity.
  - For saving system installation logs, use virtual floppy softmedia.
- Cisco supports Unified CM clusters running some subscriber server node instances on UCS B-Series Blade Servers, some on C-Series Rack-Mount Servers, and other subscriber server node instances on MCS server platforms.

Computer Telephony Integration (CTI)

Cisco Computer Telephony Integration (CTI) extends the rich feature set available on Cisco Unified CM to third-party applications. These Cisco CTI-enabled applications improve user productivity, enhance the communication experience, and deliver superior customer service. At the desktop, Cisco CTI enables third-party applications to make calls from within Microsoft Outlook, open windows or start applications based on incoming caller ID, and remotely track calls and contacts for billing purposes. Cisco CTI-enabled server applications can intelligently route contacts through an enterprise network, provide automated caller services such as auto-attendant and interactive voice response (IVR), as well as capture media for contact recording and analysis.

CTI applications generally fall into one of two major categories:

- First-party applications — Monitor, control, and media termination

  First-party CTI applications are designed to register devices such as CTI ports and route points for call setup, tear-down, and media termination. Because these applications are directly in the media path, they can respond to media-layer events such as in-band DTMF. Interactive voice response and Cisco Attendant Console are examples of first-party CTI applications that monitor and control calls while also interacting with call media.

- Third-party application — Monitor and control

  Third-party CTI applications can also monitor and control calls, but they do not directly control media termination.

  – Monitoring applications

    A CTI application that monitors the state of a Cisco IP device is called a monitoring application. A busy-lamp-field application that displays on-hook/off-hook status or uses that information to indicate a user's availability in the form of Presence are both examples of third-party CTI monitoring applications.

  – Call control applications

    Any application that uses Cisco CTI to remotely control a Cisco IP device using out-of-band signaling is a call control application. Cisco Jabber, when configured to remotely control a Cisco IP device, is a good example of a call control application.

  – Monitor + call control applications

    These are any CTI applications that monitor and control a Cisco IP device. Cisco Unified Contact Center Enterprise is a good example of a combined monitor and control application because it monitors the status of agents and controls agent phones through the agent desktop.

Note While the distinction between a monitor, call control, and monitor + control application is called out here, this granularity is not exposed to the application developer. All CTI applications using Cisco CTI are enabled for both monitoring and control.

The following devices can be monitored or controlled through CTI:

- CTI Route Point
- CTI Port
- Cisco Unified IP Phones supporting CTI
- CTI Remote Device

CTI Remote Device is a new phone type introduced in Cisco Unified CM 9.0. It provides the ability for a CTI application to have monitoring and limited call control capabilities over phones that do not support CTI, such as traditional PSTN phones, mobile phones, third-party phones, or phones attached to a third-party PBX.

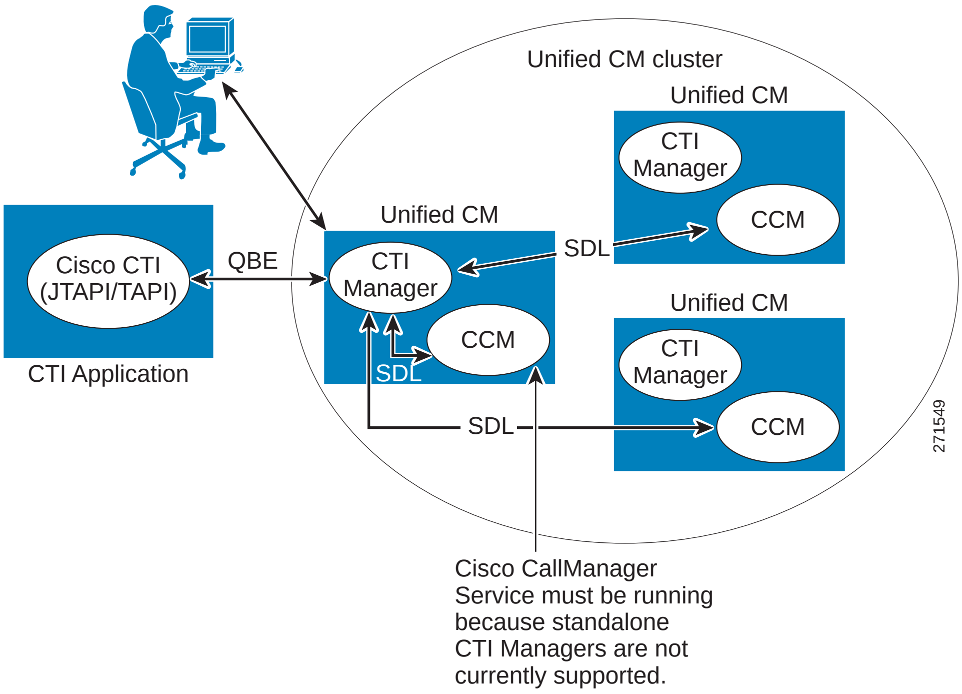
CTI Architecture

Cisco CTI consists of the following components (see Figure 8-9), which interact to enable applications to take advantage of the telephony feature set available in Cisco Unified CM:

- CTI-enabled application — Cisco or third-party application written to provide specific telephony features and/or functionality.

- JTAPI and TAPI — Two standard interfaces supported by Cisco CTI. Developers can choose to write applications using their preferred method library.
- Unified JTAPI and Unified TSP Client — Converts external messages to internal Quick Buffer Encoding (QBE) messages used by Cisco Unified CM.
- Quick Buffer Encoding (QBE) — Unified CM internal communication messages.
- Provider — A logical representation of a connection between the application and CTI Manager, used to facilitate communication. The provider sends device and call events to the application while accepting control instructions that allow the application to control the device remotely.
- Signaling Distribution Layer (SDL) — Unified CM internal communication messages.
- Publisher and subscriber — Cisco Unified Communications Manager (Unified CM) servers.
- CCM — The Cisco CallManager Service (ccm.exe), the telephony processing engine.
- CTI Manager (CTIM) — A service that runs on one or more Unified CM subscribers operating in primary/secondary mode and that authenticates and authorizes telephony applications to control and/or monitor Cisco IP devices.

Figure 8-9 Cisco CTI Architecture



Once an application is authenticated and authorized, the CTIM acts as the broker between the telephony application and the Cisco CallManager Service. (This service is the call control agent and should not be confused with the overall product name Cisco Unified Communications Manager.) The CTIM responds to requests from telephony applications and converts them to Signaling Distribution Layer (SDL) messages used internally in the Unified CM system. Messages from the Cisco CallManager Service are also received by the CTIM and directed to the appropriate telephony application for processing.

The CTIM may be activated on any of the Unified CM subscriber servers in a cluster that have the Cisco CallManager Service active. This allows up to eight CTIMs to be active within a Unified CM cluster. Standalone CTIMs are currently not supported.

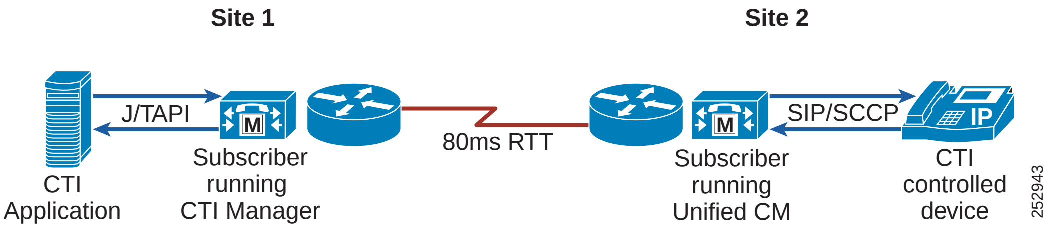
CTI Applications and Clustering Over the WAN

Deployments that employ clustering over the WAN are supported in the following two scenarios:

- CTI Manager over the WAN (see Figure 8-10)

  In this scenario, the CTI application and its associated CTI Manager are on one side of the WAN (Site 1), and the monitored or controlled devices are on the other side, registered to a Unified CM subscriber (Site 2). The round-trip time (RTT) must not exceed the currently supported limit of 80 ms for clustering over the WAN. To calculate the necessary bandwidth for CTI traffic, use the formula in the section on Local Failover Deployment Model, page 5-37. Note that this bandwidth is in addition to the Intra-Cluster Communication Signaling (ICCS) bandwidth calculated as described in the section on Local Failover Deployment Model, page 5-37, as well as any bandwidth required for audio (RTP traffic).

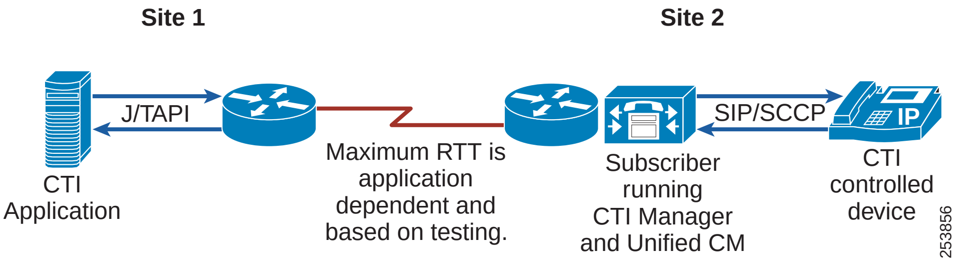
Figure 8-10 CTI Over the WAN



- TAPI and JTAPI applications over the WAN (CTI application over the WAN; see Figure 8-11)

  In this scenario, the CTI application is on one side of the WAN (Site 1), and its associated CTI Manager is on the other side (Site 2). In this scenario, it is up to the CTI application developer or provider to ascertain whether or not their application can accommodate the RTT as implemented. In some cases failover and failback times might be higher than if the application is co-located with its CTI Manager. In those cases, the application developer or provider should provide guidance as to the behavior of their application under these conditions.

Figure 8-11 JTAPI Over the WAN



Note Support for TAPI and JTAPI over the WAN is application dependent. Both customers and application developers or providers should ensure that their applications are compatible with any such deployment involving clustering over the WAN.

Capacity Planning for CTI

The maximum number of supported CTI-controlled devices is 40,000 per cluster. For more information on CTI capacity planning, including per-platform node and cluster CTI capacities as well as CTI resource calculation formulas and examples, see the chapter on Unified Communications Design and Deployment Sizing Considerations, page 29-1.

High Availability for CTI

This section provides some guidelines for provisioning CTI for high availability.

CTI Manager

CTI Manager must be enabled on at least one and possibly all call processing subscribers within the Unified CM cluster. The client-side interfaces (TAPI TSP or JTAPI client) allow for two IP addresses each, which then point to Unified CM servers running the CTIM service. For CTI application redundancy, Cisco recommends having the CTIM service activated on at least two Unified CM servers in a cluster, as shown in Figure 8-12.

Redundancy, Failover, and Load Balancing

For CTI applications that require redundancy, the TAPI TSP or JTAPI client can be configured with two IP addresses, thereby allowing an alternate CTI Manager to be used in the event of a failure. It should be noted that this redundancy is not stateful in that no information is shared and/or made available between the two CTI Managers, and therefore the CTI application will have some degree of re-initialization to go through, depending on the exact nature of the failover.

When a CTI Manager fails-over, just the CTI application login process is repeated on the now-active CTI Manager. Whereas, if the Unified CM server itself fails, then the re-initialization process is longer due to the re-registration of all the devices from the failed Unified CM to the now-active Unified CM, followed by the CTI application login process.

For CTI applications that require load balancing or that could benefit from this configuration, the CTI application can simply connect to two CTI Managers simultaneously, as shown in Figure 8-12.

Figure 8-12 Redundancy and Load Balancing

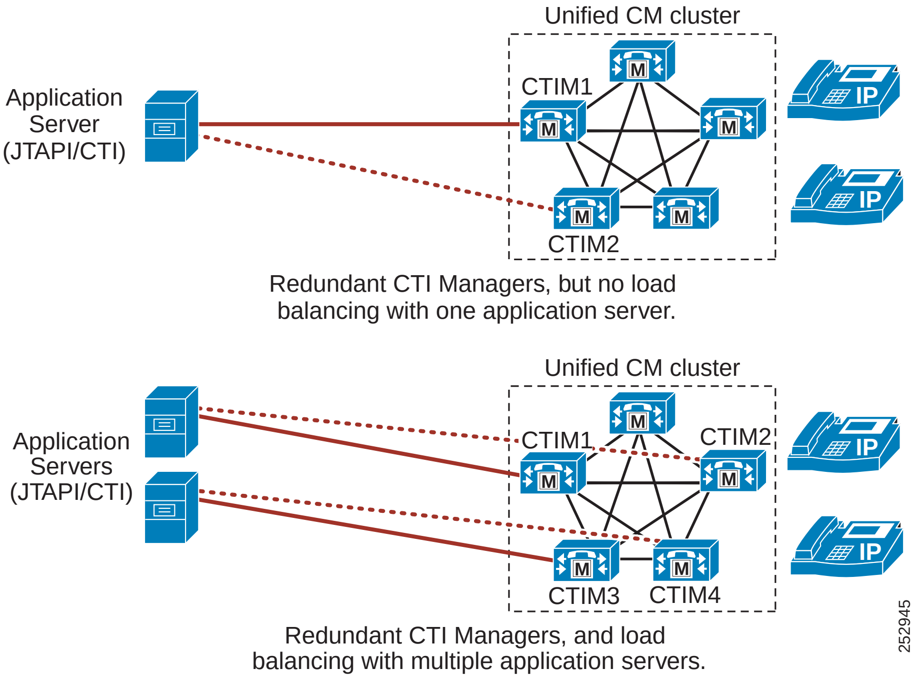

Unified CM cluster
Application Server (JTAPI/CTI)
CTIM1
CTIM2
M
IP
Redundant CTI Managers, but no load balancing with one application server.
Unified CM cluster
Application Servers (JTAPI/CTI)
CTIM1
CTIM2
CTIM3
CTIM4
M
IP
Redundant CTI Managers, and load balancing with multiple application servers.
252945

Figure 8-13 shows an example of this type of configuration for Cisco Unified Contact Center Enterprise (Unified CCE). This type of configuration has the following characteristics:

- Unified CCE uses two Peripheral Gateways (PGs) for redundancy.
- Each PG logs into a different CTI Manager.
- Only one PG is active at any one time.

Figure 8-13 CTI Redundancy with Cisco Unified Contact Center Enterprise

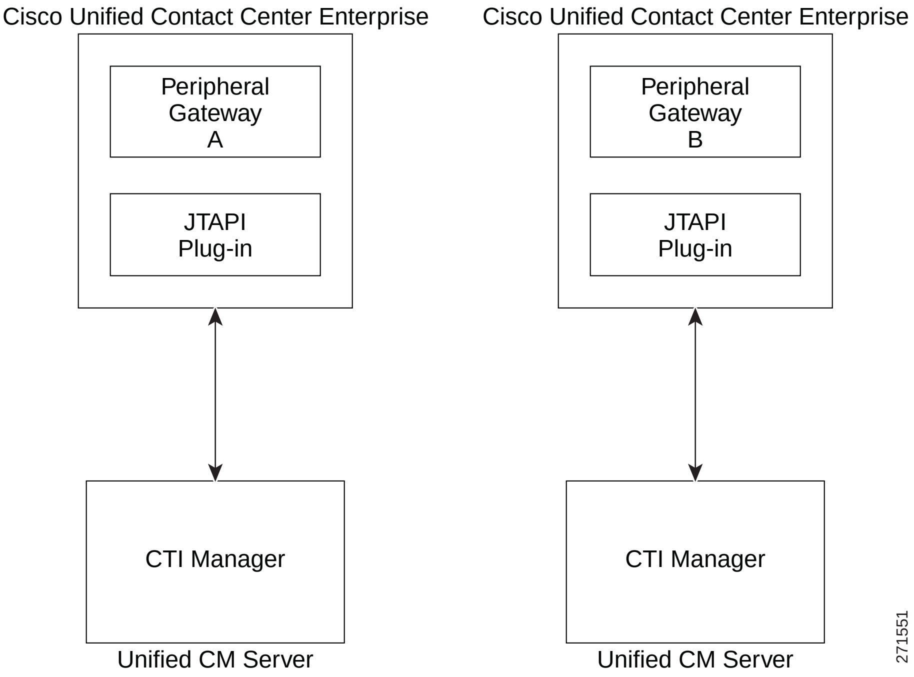
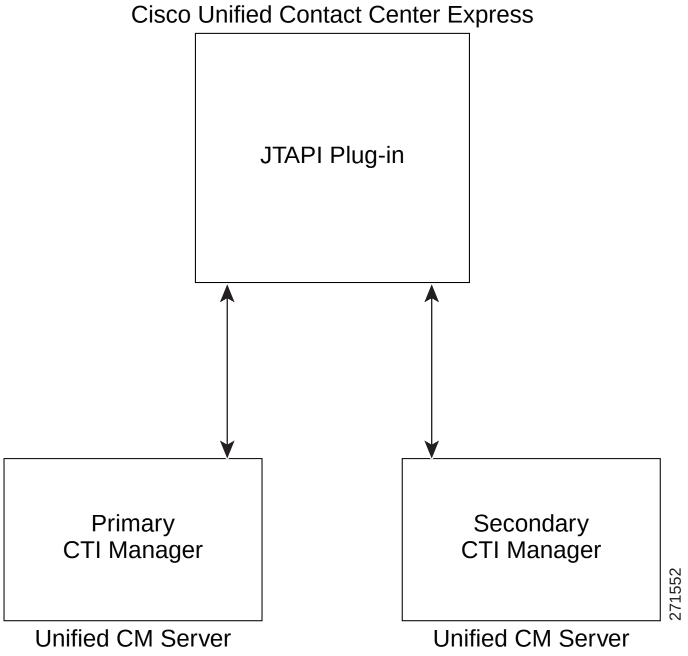


Figure 8-14 shows an example of this type of configuration for Cisco Unified Contact Center Express (Unified CCX). This type of configuration has the following characteristics:

- Unified CCX has two IP addresses configured, one for each CTI Manager.
- If connection to the primary CTI Manager is lost, Unified CCX fails-over to its secondary CTI Manager.

Figure 8-14 CTI Redundancy with Cisco Unified Contact Center Express



Implementation

For guidance and support on writing applications, application developers should consult the Cisco Developer Connection, located at

http://developer.cisco.com/web/cdc/community

Gatekeeper Design Considerations

A single Cisco IOS gatekeeper can provide call routing and call admission control for up to 100 Unified CM clusters in a distributed call processing environment. Multiple gatekeepers can be configured to support thousands of Unified CM clusters. You can also implement a hybrid Unified CM and toll-bypass network by using Cisco IOS gatekeepers to provide communication and call admission control between the H.323 gateways and Unified CM.

Gatekeeper call admission control is a policy-based scheme requiring static configuration of available resources. The gatekeeper is not aware of the network topology, so it is limited to hub-and-spoke topologies.

Most Cisco IOS routers support the gatekeeper feature. For specific platform support for gatekeeper functionality, refer to the *Cisco IOS H323 Gatekeeper Data Sheet*, available at

http://www.cisco.com/en/US/prod/collateral/voicesw/ps6788/vcallcon/ps4139/data_sheet_c78_561921.html

You can configure Cisco IOS gatekeepers in a number of different ways for redundancy, load balancing, and hierarchical call routing. This section considers the design requirements for building a gatekeeper network, but it does not deal with the call admission control or dial plan resolution aspects, which are covered in the chapters on Call Admission Control, page 11-1, and Dial Plan, page 9-1, respectively.

For additional information regarding gatekeepers, refer to the *Cisco IOS H.323 Configuration Guide*, available at

http://www.cisco.com/en/US/products/ps10591/products_installation_and_configuration_guides_list.html

Hardware Platform Selection

The choice of gatekeeper platform is based on the number of calls per second and the number of concurrent calls. A higher number of calls per second requires a more powerful CPU. A higher number of concurrent calls requires more memory. Select Cisco IOS routers with large memory capacity and higher performance CPUs when design requirements include high call volumes and large numbers of simultaneous calls.

For more information about gatekeeper platforms, refer to the *Cisco IOS H323 Gatekeeper Data Sheet*, available at

http://www.cisco.com/en/US/prod/collateral/voicesw/ps6788/vcallcon/ps4139/data_sheet_c78_561921.html

Gatekeeper Redundancy

With gatekeepers providing all call routing and admission control for intercluster communications, redundancy is required. There are two methods for providing gatekeeper redundancy: gatekeeper clustering and directory gatekeeper.

Note Cisco recommends that you use gatekeeper clustering to provide gatekeeper redundancy whenever possible. Do not use Hot Standby Router Protocol (HSRP) for gatekeeper redundancy unless gatekeeper clustering is not available in your software feature set.

Gatekeeper Clustering (Alternate Gatekeeper)

Gatekeeper clustering (alternate gatekeeper) enables the configuration of a "local" gatekeeper cluster, with each gatekeeper acting as primary for some Unified CM trunks and an alternate for others. Gatekeeper Update Protocol (GUP) is used to exchange state information between gatekeepers in a local cluster. GUP tracks and reports CPU utilization, memory usage, active calls, and number of registered endpoints for each gatekeeper in the cluster. Load balancing is supported by setting thresholds for any of the following parameters in the GUP messaging:

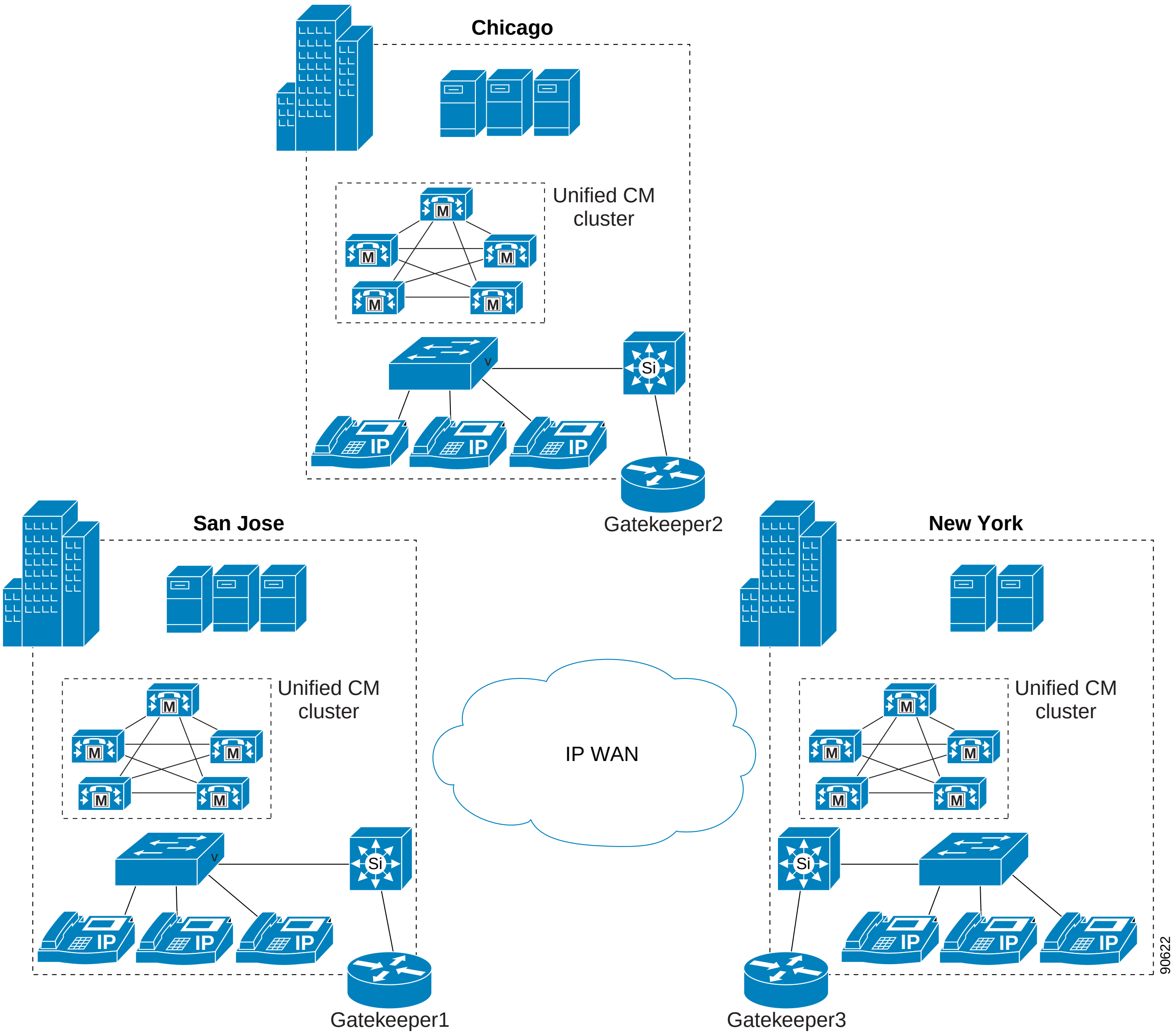
- CPU utilization
- Memory utilization
- Number of active calls
- Number of registered endpoints

With the support of gatekeeper clustering (alternate gatekeeper), stateful redundancy and load balancing is available. Gatekeeper clustering provides the following features:

- Local and remote clusters
- Up to five gatekeepers in a local cluster
- Gatekeepers in local clusters can be located in different subnets or locations
- No failover delay (Because the alternate gatekeeper is already aware of the endpoint, it does not have to go through the full registration process.)
- Gatekeepers in a cluster pass state information and provide load balancing

Figure 8-15 shows three sites with Unified CM distributed call processing and three distributed gatekeepers configured in a local cluster.

Figure 8-15 *Gatekeeper Clustering*



In Figure 8-15, each site's Unified CM cluster registers to the local gatekeeper. The local gatekeeper service is made redundant using gatekeeper clustering such that each local gatekeeper is backed up by a gatekeeper at another site.

Consider the following guidelines when deploying gatekeeper clustering:

- Each Unified CM cluster should have a local zone configured to support Unified CM trunk registrations. This local zone is configured within Unified CM and on the gatekeeper located with the Unified CM cluster. In the example shown in Figure 8-15, the Unified CM cluster located in the San Jose site will have a gatekeeper controlled trunk with a zone name matching the local zone name configured on the San Jose gatekeeper (Gatekeeper 1). Likewise, the Chicago and New York Unified CM clusters will have zone names matching the local zone name on the gatekeepers located in their respective locations (Gatekeeper 2 in Chicago and Gatekeeper 3 in New York).

- A gatekeeper cluster is defined for each local zone, with backup zones on the other gatekeepers configured using the **element** command. In the example shown in Figure 8-15, the San Jose gatekeeper (Gatekeeper 1) has a local zone with elements for both the Chicago gatekeeper (Gatekeeper 2) and New York gatekeeper (Gatekeeper 3). Likewise, the Chicago and New York gatekeepers have local zones with elements for both the San Jose gatekeeper and each other (respectively).
- Use the **gw-type-prefix** command to allow all locally unresolved calls to be forwarded to a device registered with the configured technology prefix in the local zone. In the example shown in Figure 8-15, each Unified CM gatekeeper controlled trunk is configured with a technology prefix of 1#* and the gatekeeper at each site is configured with a default-technology gw-type-prefix of 1#*.
- Load balancing between clustered gatekeepers is configured using the **load-balance** command. Given the example shown in Figure 8-15, each site's gatekeeper can be configured to load balance or move endpoint/gateway registration from the local gatekeeper to the alternate gatekeeper within the cluster based on thresholds for CPU utilization, memory utilization, number of endpoints, and/or number of calls. For example, the San Jose gatekeeper (Gatekeeper 1) might be configured to move endpoint or gateway registrations to the Chicago gatekeeper based on a high-water CPU and memory threshold of 80%. In that case, if the San Jose gatekeeper's memory or CPU utilization reaches 80%, the gatekeeper will begin sending Chicago gatekeeper information in the H.323 Registration, Admission, and Status (RAS) messages it sends to the San Jose Unified CM cluster to maintain trunk registration state. Likewise, the other gatekeepers in Chicago and New York could be similarly configured to load-balance local Unified CM trunk registration loads at those sites to gatekeepers located in other sites.
- When routing calls between the three Unified CM clusters in Figure 8-15, the gatekeeper at each site should be configured to check that appropriate bandwidth is available on the network between that location and the location to which the call is being routed. If there is not sufficient bandwidth, the call should not be routed. The **bandwidth interzone** command is recommended for specifying interzone call bandwidth between the distribute Unified CM locations.
- Use the **arq reject-unknown-prefix** command to guard against potential call routing loops across redundant Unified CM trunks within a cluster. This command prevents the gatekeeper from forwarding call routing requests back to the local gateway or Unified CM trunk when the dialed prefix does not match a defined prefix.

For additional information regarding gatekeeper deployment and configuration, refer to the *Cisco IOS H.323 Configuration Guide*, available at

http://www.cisco.com/en/US/products/ps10591/products_installation_and_configuration_guides_list.html

Directory Gatekeeper Redundancy

You can implement directory gatekeeper redundancy by using HSRP or by configuring multiple identical directory gatekeepers. When a gatekeeper is configured with multiple remote zones using the same zone prefix, the gatekeeper can use either of the following methods:

- Sequential LRQs (default)

  Redundant remote zones (matching zone prefixes) are assigned a cost, and LRQs are sent to the matching zones in order based on the cost values. Using sequential LRQs saves WAN bandwidth by not blasting LRQs to all matching gatekeepers.

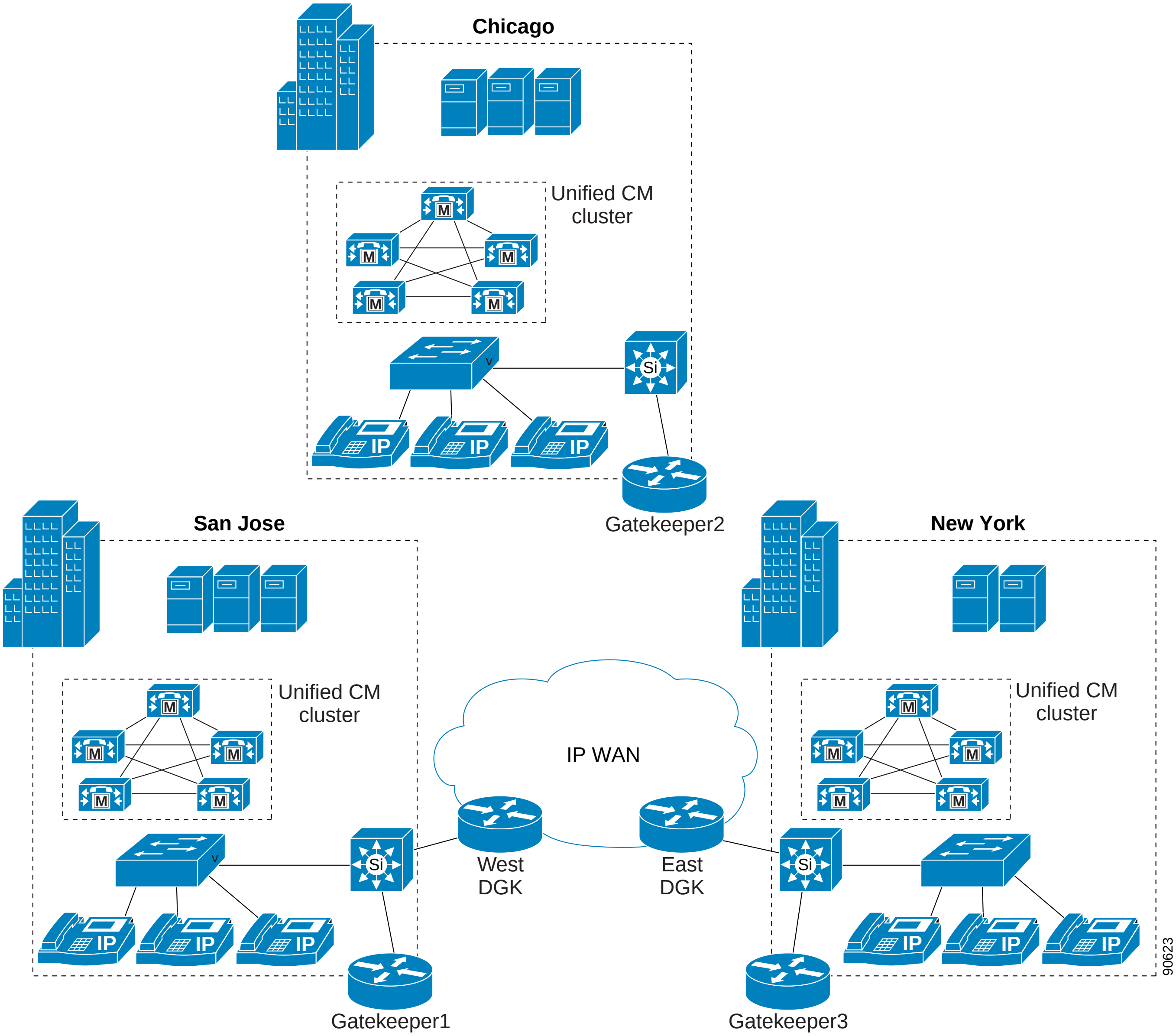
- LRQ Blast

  LRQs are sent to redundant zones (matching zone prefixes) simultaneously. The first gatekeeper to respond with an Location Confirm (LCF) is the one that is used.

Cisco recommends that you use multiple active directory gatekeepers with sequential LRQs, thus allowing directory gatekeepers to be placed in different locations. Using HSRP requires both directory gatekeepers to be located in the same subnet, and only one gatekeeper can be active at any time.

Figure 8-16 shows the same three-site Unified CM distributed call processing deployment with three distributed local gatekeepers as shown in Figure 8-15. However, unlike the deployment illustrated in Figure 8-15, the deployment in Figure 8-16 depicts the three distributed local gatekeepers relying on two active directory gatekeepers for redundant inter-site call routing (rather than relying on alternate or clustered gatekeepers).

Figure 8-16 Redundant Directory Gatekeepers



Consider the following guidelines when deploying redundant directory gatekeepers:

- When configuring redundancy for directory gatekeepers, configure each directory gatekeeper with a local zone. In the example shown in Figure 8-16, the directory gatekeeper located in San Jose (West DGK) is configured with one local zone name and IP address, while the directory gatekeeper located in New York (East DGK) is configured with another local zone name and IP address.
- Directory gatekeepers should be configured with remote zones corresponding to each gatekeeper in the network. In the example shown in Figure 8-16, both the directory gatekeeper in San Jose (West DGK) and the one in New York (East DGK) are configured with a remote zone corresponding to the gatekeeper at the San Jose site (Gatekeeper 1), a remote zone corresponding to the gatekeeper at the Chicago site (Gatekeeper 2), and a remote zone corresponding to the gatekeeper at the New York site (Gatekeeper 3). The configuration for these remote sites is the same on both directory gatekeepers.
- Each directory gatekeeper is configured with dialed number prefixes corresponding to each remote zone for inter-zone call routing. In the example shown in Figure 8-16, both directory gatekeepers are configured with prefixes corresponding to the local area code serviced by each site's gatekeeper. For example, the prefix 408 is configured for the San Jose gatekeeper (Gatekeeper 1) remote zone, the prefix 720 is configured for the Chicago gatekeeper (Gatekeeper 2) remote zone, and the prefix 212 is configured for the New York gatekeeper (Gatekeeper 3) remote zone. Additional prefixes can be configured for each remote zone as needed to accommodate other dialed number prefixes. Because calls are never routed to the local directory gatekeeper zone, a prefix is not required for those zones. In addition to configuring specific prefixes, the wildcard notation * can be used to match all prefixes not explicitly defined.
- Configure the **lrq forward-queries** command on each directory gatekeeper to ensure that call setup location requests (LRQ) received from one gatekeeper are forwarded to one of the other gatekeepers as appropriate for service based on dialed prefixes. Given the example shown in Figure 8-16, both the directory gatekeeper in San Jose (West DGK) and the one in New York (East DGK) should be configured to forward LRQ queries.

Note Directory gatekeepers do not contain any active endpoint registrations and do not supply any bandwidth management.

- Just as with the previous example (Figure 8-15), the local site gatekeepers shown at each site in Figure 8-16 provide gatekeeper services and registration for Unified CM cluster trunks at each site.
- Each local site gatekeepers is configured with a remote zone for each directory gatekeeper. Given the example depicted in Figure 8-16, the local gatekeeper in the San Jose site (Gatekeeper 1) has remote zones configured for both the directory gatekeeper in San Jose (West DGK) and the directory gatekeeper in New York (East DGK). The local gatekeepers at the Chicago (Gatekeeper 2) and New York (Gatekeeper 3) sites are configured identically.
- Each local site gatekeeper should be configured to limit bandwidth between the local gatekeeper zone and any remote zones configured. In the example shown in Figure 8-16, each local gatekeeper is configured with the **bandwidth remote** command determining the amount of bandwidth available for routing calls to the remote zone directory gatekeepers. For example, the **bandwidth remote** command is configured on the San Jose gatekeeper (Gatekeeper 1) to limit the available bandwidth for routing calls to the remote zone defined for the directory gatekeeper in San Jose (West DGK) and the remote zone defined for the directory gatekeeper in New York (East DGK). This in turn limits the bandwidth available for routing calls between the San Jose site gatekeeper (Gatekeeper 1) and either of the other site gatekeepers (Gatekeeper 2 or Gatekeeper 3). This same configuration would be replicated on the other local site gatekeepers.

- Each local site gatekeeper is configured with zone prefixes for the local zone corresponding to the local gatekeeper and for both remote zones corresponding to the two directory gatekeepers. The former local zone prefix handles call routing to the local Unified CM cluster, while the latter remote zone prefixes handle inter-zone call routing to the other gatekeeper sites. Given the example depicted in Figure 8-16, the local gatekeeper in the San Jose site (Gatekeeper 1) is configured with a local zone prefix of 408 and remote zone prefixes of ten dots (.) corresponding to the two directory gatekeepers (East DGK and West DGK). These ten dot (.) prefixes match all normalized ten-digit E.164 dialed numbers that do not begin with the local zone prefix of 408. Thus all calls routed by Gatekeeper 1 that do not begin with 408 will be routed to one of the other gatekeeper sites through one of the directory gatekeepers. The local gatekeepers at the Chicago (Gatekeeper 2) and New York (Gatekeeper 3) sites are configured with local zone prefixes 720 and 212 respectively, along with the same general remote zone ten-dot prefixes.
- Sequential location requests (LRQs) are used by default when matching zone prefixes are configured. In the example shown in Figure 8-16, for all calls routed to dialed numbers that do not start with 408, the local gatekeeper at the San Jose site (Gatekeeper 1) will first send an LRQ to the directory gatekeeper located in San Jose (West DGK) based on the generic ten-dot (.) prefix configured for the remote zone corresponding to the West DGK. If a response is not received from the West DGK, then Gatekeeper 1 will send an LRQ to the directory gatekeeper located in New York (East DGK) based on the generic ten-dot (.) prefix configured for the remote zone corresponding to the East DGK. Similarly, local gatekeepers at the Chicago (Gatekeeper 2) and the New York (Gatekeeper 3) sites will first send an LRQ to one of the directory gatekeepers based on remote zone and ten-dot (.) prefix configuration and to the second directory gatekeeper if a response is not received from the first.
- Just as with the previous gatekeeper clustering example (Figure 8-15), the **gw-type-prefix** command is used to ensure all locally unresolved calls are forwarded to a device registered with the configured technology prefix in the local zone. Likewise, the **arq reject-unknown-prefix** command guards against potential call routing loops across redundant Unified CM trunks within a cluster.

For additional information regarding directory gatekeeper deployment and configuration, refer to the *Cisco IOS H.323 Configuration Guide*, available at

http://www.cisco.com/en/US/products/ps10591/products_installation_and_configuration_guides_list.html

Interoperability of Unified CM and Unified CM Express

This section explains the requirements for interoperability and internetworking of Cisco Unified CM with Cisco Unified Communications Manager Express (Unified CME) using H.323 or SIP trunking protocol in a multisite IP telephony deployment. This section highlights the recommended deployments between phones controlled by Unified CM and phones controlled by Unified CME.

This section covers the following topics:

Overview of Interoperability Between Unified CM and Unified CME

Either H.323 or SIP can be used as a trunking protocol to interconnect Unified CM and Unified CME. When deploying Unified CM at the headquarters or central site in conjunction with one or more Unified CME systems for branch offices, network administrators must choose either the SIP or H.323 protocol after careful consideration of protocol specifics and supported features across the WAN trunk. Using H.323 trunks to connect Unified CM and Unified CME has been the predominant method in past years, until more enhanced capabilities for SIP phones and SIP trunks were added in Unified CM and Unified CME. This section first describes some of the features and capabilities that are independent of the trunking protocol for Unified CM and Unified CME interoperability, then it explains some of the most common design scenarios and best practices for using SIP trunks and H.323 trunks.

Call Types and Call Flows

In general, Unified CM and Unified CME interworking allows all combination of calls from SCCP IP phones to SIP IP phones, or vice versa, across a SIP trunk or H.323 trunk. Calls can be transferred (blind or consultative) or forwarded back and forth between the Unified CM and Unified CME SIP and/or SCCP IP phones.

When connected to Unified CM via H.323 trunks, Unified CME can auto-detect Unified CM calls. When a call terminating on Unified CME is transferred or forwarded, Unified CME regenerates the call and routes the call appropriately to another Unified CME or Unified CM by hairpinning the call. Unified CME hairpins the call legs from Unified CM for the VoIP calls across SIP or H.323 trunks when needed. For more information on allowing auto-detection on a non-H.450 supported Unified CM network and for enabling or disabling supplementary services for H450.2, H450.3, or SIP, refer to the Unified CME product documentation available at

http://www.cisco.com/en/US/products/sw/voicesw/ps4625/tsd_products_support_series_home.html

When connected to Unified CM via SIP trunks, Unified CME does not auto-detect Unified CM calls. By default, Unified CME always tries to redirect calls using either a SIP Refer message for call transfer or a SIP 302 Moved Temporarily message for call forward; if that fails, Unified CME will then try to hairpin the call.

Music on Hold

While Unified CM can be enabled to stream MoH in both G.711 and G.729 formats, Unified CME streams MoH only in G.711 format. Therefore, when Unified CME controls the MoH audio on a call placed on hold, it requires a transcoder to transcode between a G.711 MoH stream and a G.729 call leg.

Ad Hoc and Meet Me Hardware Conferencing

Hardware DSP resources are required for both Ad Hoc and Meet Me conferences. Whether connected via SIP, H.323, or PSTN, both Unified CM and Unified CME phones can be invited or added to an Ad Hoc conference to become conference participants as along as the phones are reachable from the network. When calls are put on hold during an active conference session, music will not be heard by the conference participants in the conference session.

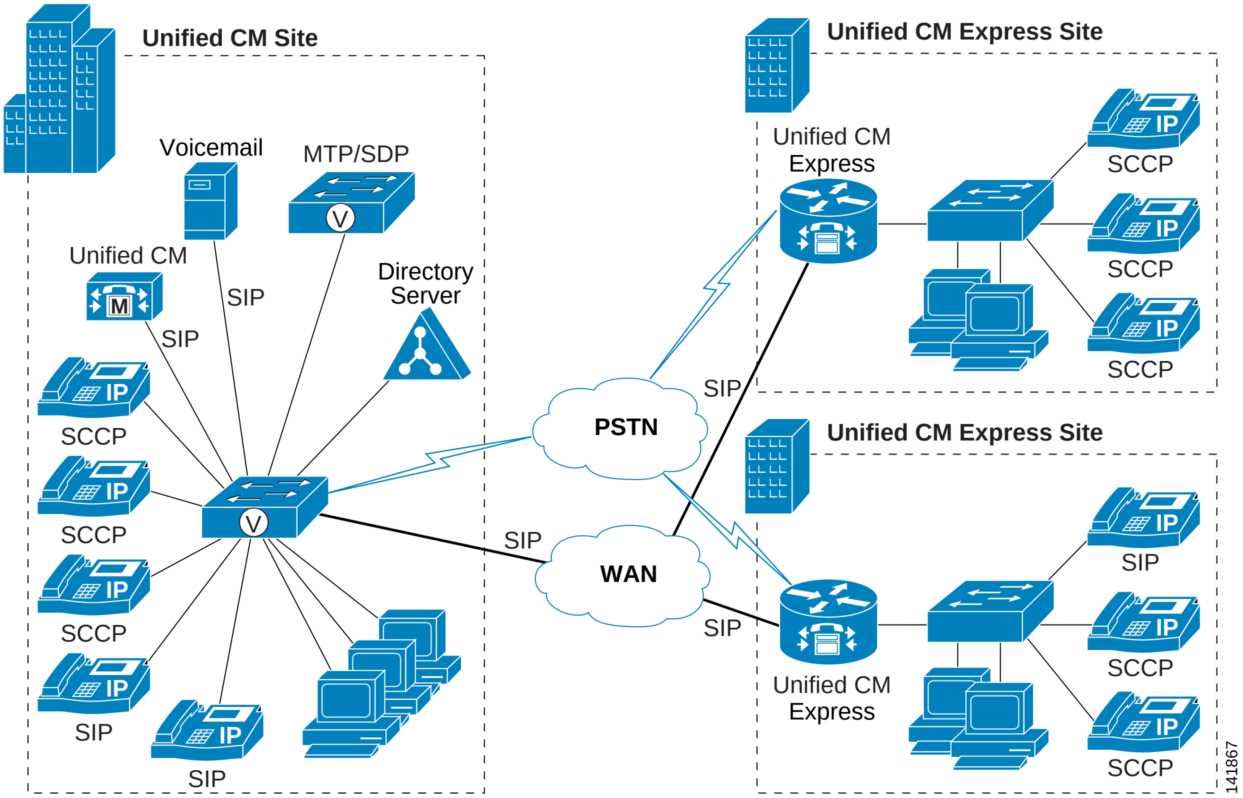
For information on required and supported DSP resources and the maximum number of conference participants allowed for Ad Hoc or Meet Me conferences, refer to the Unified CME product documentation available at

http://www.cisco.com/en/US/products/sw/voicesw/ps4625/tsd_products_support_series_home.html

Unified CM and Unified CME Interoperability via SIP in a Multisite Deployment with Distributed Call Processing

Unified CM can communicate directly with Unified CME using a SIP interface. Figure 8-17 shows a Cisco Unified Communications multisite deployment with Unified CM networked directly with Cisco Unified CME using a SIP trunk.

Figure 8-17 Multisite Deployment with Unified CM and Unified CME Using SIP Trunks



Best Practices

Follow these guidelines and best practices when using the deployment model illustrated in Figure 8-17:

- Configure a SIP Trunk Security Profile with **Accept Replaces Header** selected.
- Configure a SIP trunk on Unified CM using the SIP Trunk Security Profile created, and also specify a ReRouting CSS. The ReRouting CSS is used to determine where a SIP user (transferor) can refer another user (transferee) to a third user (transfer target) and which features a SIP user can invoke using the SIP 302 Redirection Response and INVITE with Replaces.
- For SIP trunks there is no need to enable the use of media termination points (MTPs) when using SCCP endpoints on Unified CME. However, SIP endpoints on Unified CME require the use of media termination points on Unified CM to be able to handle delayed offer/answer exchanges with the SIP protocol (that is, the reception of INVITEs with no Session Description Protocol).

- Route calls to Unified CME via a SIP trunk using the Unified CM dial plan configuration (route patterns, route lists, and route groups).
- Use Unified CM device pools and regions to configure a G.711 codec within the site and the G.729 codec for remote Unified CME sites.
- Configure the **allow-connections sip to sip** command under **voice services voip** on Unified CME to allow SIP-to-SIP call connections.
- For SIP endpoints, configure the **mode cme** command under **voice register global**, and configure **dtmf-relay rtp-nte** under the **voice register pool** commands for each SIP phone on Unified CME.
- For SCCP endpoints, configure the **transfer-system full-consult** command and the **transfer-pattern .T** command under **telephony-service** on Unified CME.
- Configure the SIP WAN interface voip dial-peers to forward or redirect calls, destined for Unified CM, with **session protocol sipv2** and **dtmf-relay** [**sip-notify** | **rtp-nte**] on Unified CME.

Design Considerations

This section first covers some characteristics and design considerations for Unified CM and Unified CME interoperability via SIP in some main areas such as supplementary services for call transfer and forward, presence service for busy lamp field (BLF) notification for speed-dial buttons and directory call lists, and out-of-dialog (OOD-Refer) for integration with partner applications and third-party phone control for click-to-dial between the Unified CM phones and Unified CME phones. The section also covers some general design considerations for Unified CM and Unified CME interoperability via SIP.

Supplementary Services

SIP Refer or SIP 302 Moved Temporarily messages can be used for supplementary services such as call transfer or call forward on Unified CME or Unified CM to instruct the transferee (referee) or phone being forwarded (forwardee) to initiate a new call to the transfer-to (refer-to) target or forward-to target. No hairpinning is needed for call transfer or call forward scenarios when the SIP Refer or SIP 302 Moved Temporarily message is supported.

However, **supplementary-service** must be disabled if there are certain extensions that have no DID mapping or if Unified CM or Unified CME does not have a dial plan to route the call to the DID in the SIP 302 Moved Temporarily message. When **supplementary-service** is disabled, Unified CME hairpins the calls or sends a re-invite SIP message to Unified CM to replace the media path to the new called party ID. Both signaling and media are hairpinned, even when multiple Unified CMEs are involved for further call forwards. The **supplementary-service** can also be disabled for transferred calls. In this case, the SIP Refer message will not be sent to Unified CM, but the transferee (referee) party and transfer-to party (refer-to target) are hairpinned.

Note Supplementary services can be disabled with the command **no supplementary-service sip moved-temporarily** or **no supplementary-service sip refer** under **voice service voip** or **dial-peer voice xxxx voip**.

The following examples illustrate the call flows when supplementary services are disabled:

- Unified CM phone B calls Unified CME phone A, which is set to call-forward (all, busy, or no answer) to phone C (either a Unified CM phone, a Unified CME phone on the same or different Unified CME, or a PSTN phone).

  Unified CME does not send the SIP 302 Moved Temporarily message to Unified CM, but hairpins the call between Unified CM phone B and phone C.
- Unified CM phone B calls Unified CME phone A, which transfer the call to phone C (either a Unified CM phone, a Unified CME phone, or a PSTN phone).

  Unified CME does not send the SIP Refer message to Unified CM, but hairpins the call between Unified CM phone B and phone C.

General Design Considerations for Unified CM and Unified CME Interoperability via SIP

- Disable **supplementary-service** if SIP 302 Moved Temporarily or SIP Refer messages are not supported by Unified CM, otherwise Unified CM cannot route the call to the transfer-to or forward-to target.
- In a SIP-to-SIP call scenario, a Refer message is sent by default from the transferor to the transferee, the transferee sets up a new call to the transfer-to target, and the transferor hears ringback tone by default while waiting for the transfer at connect. If **supplementary-service** is disabled on Unified CME, Unified CME will provide in-band ringback tone right after the call between the transferee and transfer-to target is connected.
- Presence service is supported on Unified CM and Unified CME via SIP trunk only.
- The OOD-Refer feature allows third-party applications to connect two endpoints on Unified CM or Unified CME through the use of the SIP REFER method. Consider the following factors when using OOD-Refer:
  - Both Unified CM and Unified CME must be configured to enable the OOD-Refer feature.
  - Call Hold, Transfer, and Conference are not supported during an OOD-Refer transaction, but they are not blocked by Unified CME.
  - Call transfer is supported only after the OOD-Refer call is in the connected state and not before the call is connected; therefore, call transfer-at-alert is not supported.
- Control signaling in TLS is supported, but SRTP is not supported over the SIP trunk.
- SRTP over a SIP trunk is a gateway feature in Cisco IOS for Unified CM. SRTP support is not available with Unified CM and Unified CME interworking via SIP trunks.

Note When multiple PSTN connections exist (one for Unified CM and one for Unified CME), fully attended transfer between a Unified CM endpoint and a Unified CME endpoint to a PSTN endpoint will fail. The recommendation is to use blind transfer when using multiple PSTN connections, and it is configured under **telephony-service** as **transfer-system full-blind**.

Unified CM and Unified CME Interoperability via H.323 in a Multisite Deployment with Distributed Call Processing

There are two deployment options to achieve interoperability between Unified CM and Unified CME via H.323 connections in a multisite WAN deployment with distributed call processing. The first option is to deploy a Cisco Unified Border Element as a front-end device of Unified CM, which has a peer-to-peer H.323 connection with a remote Unified CME system. The Cisco Unified Border Element performs dial plan resolution between Unified CM and Unified CME, and it also terminates and re-originates call signaling messages between the two. The Cisco Unified Border Element acts as a proxy device for a system that does not support H.450 for its supplementary services, such as Unified CM, which uses Empty Capability Sets (ECS) to invoke supplementary services. The Cisco Unified Border Element can also act as the PSTN gateway for the Unified CM cluster so that a separate PSTN gateway is not needed.

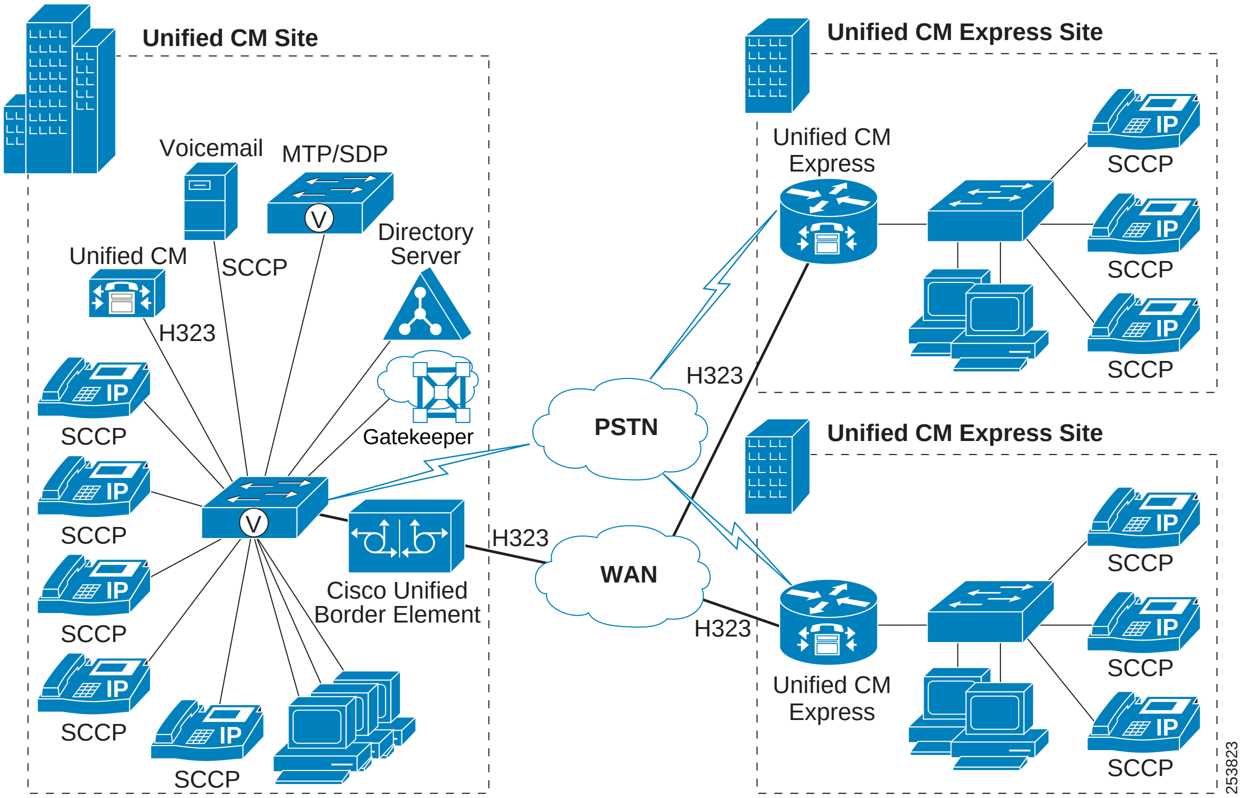
The second option is to deploy a via-zone gatekeeper. Unified CM, Unified CME, and the Cisco Unified Border Element all register with the via-zone gatekeeper as VoIP gateway devices. The via-zone gatekeeper performs dial plan resolution and bandwidth restrictions between Unified CM and Unified CME. The via-zone gatekeeper also inserts a Cisco Unified Border Element in the call path to interwork between ECS and H.450 to invoke the supplementary services. For detailed information on the via-zone gateway and Cisco Unified Border Element, see the chapter on Call Admission Control, page 11-1.

These two deployment options have the following differences:

- With the first option, the Cisco Unified Border Element registers with Unified CM as an H.323 gateway device; with the second option, it registers with via-zone gatekeeper as a VoIP gateway device.
- With the first option, the Cisco Unified Border Element performs dial plan resolution based on the VoIP dial-peer configurations on the Cisco Unified Border Element; with the second option, the via-zone gatekeeper performs dial plan resolution based on the gatekeeper dial plan configuration.
- With the first option, there is no call admission control mechanism that oversees both call legs; with the second option, the via-zone gatekeeper performs gatekeeper zone-based call admission control.
- With the second option, the via-zone gatekeeper can also act as an infrastructure gatekeeper for Unified CM, to manage all dial plan resolution and bandwidth restrictions between Unified CM clusters, between a Unified CM cluster and a network of H.323 VoIP gateways, or between a Unified CM cluster and a service provider's H.323 VoIP transport network.

Figure 8-18 shows H.323 integration between Unified CM and Unified CME using a via-zone gatekeeper and Cisco Unified Border Element.

Figure 8-18 Multisite Deployment with Unified CM and Unified CME Using a Cisco Unified Border Element or Via-Zone Gatekeeper



Best Practices

This section discusses configuration guidelines and best practices when using the deployment model illustrated in Figure 8-18 with the second deployment option (via-zone gatekeeper):

- Configure a gatekeeper-controlled H.225 trunk between Unified CM and the via-zone gatekeeper. Media termination point (MTP) resources are required over the trunk only when Unified CME tries to initiate an outbound H.323 fast-start call.
- The **Wait For Far End H.245 Terminal Capability Set** (TCS) option must be unchecked to prevent stalemate situations from occurring when the H.323 devices at both sides of the trunk are waiting the far end to send TCS first and the H.245 connection times out after a few seconds.
- Configure the Unified CM service parameter **Send H225 user info message** to **H225 info for Call Progress Tone**, which will make Unified CM send the H.225 Info message to Unified CME to play ringback tone or tone-on-hold.
- Use the Unified CM dial plan configuration (route patterns, route lists, and route groups) to send calls destined for Unified CME to the gatekeeper-controlled H.225 trunk.
- Register Unified CME and the Cisco Unified Border Element as H.323 gateways with the via-zone gatekeeper.

- Configure the **allow-connection h323 to h323** command on the Cisco Unified Border Element to allow H.323-to-H.323 call connections. This command is optional to configure on Unified CME. Configure **allow-connection h323 to sip** if Cisco Unity Connection is used on Unified CME.
- Supplementary services such as transfer and call forward will result in calls being media hairpinned when the two endpoints reside in the same Unified CME branch location.

Note The only configuration difference between the two deployment options is that the first option requires configuring the Cisco Unified Border Element as an H.323 gateway device in Unified CM. The rest of the configuration guidelines listed above are the same for both options.

Note When multiple PSTN connections exist (one for Unified CM and one for Unified CME), fully attended transfer between a Unified CM endpoint and a Unified CME endpoint to a PSTN endpoint will fail. The recommendation is to use blind transfer when using multiple PSTN connections, and it is configured under **telephony-service** as **transfer-system full-blind**.

Design Considerations

In an H.323 deployment, Unified CME supports call transfer, call forward with H.450.2, and H.450.3 as part of the H.450 standards. However, Unified CM does not support H.450, and supplementary services such as call transfer, call forward, call hold or resume are done using the Empty Capabilities Set (ECS). Therefore, when calls are transferred or forwarded between Unified CM and Unified CME, they are hairpinned and routed with a Cisco Unified Border Element and with or without a gatekeeper, as described as the two deployment models in the previous section. This section lists some of the design considerations and best practices for Unified CM and Unified CME interoperability via H.323.

Supplementary Services Such as Call Transfer and Call Forward

Unified CME can auto-detect Unified CM, which does not support H.450, by using H.450.12 protocol to automatically discover the H.450.*x* capabilities. Unified CME uses VoIP hairpin routing for calls between Unified CM and Unified CME. When the call is terminated, Unified CME hairpins the call from the Unified CM phone by re-originating and routing the call as appropriate.

Note When Unified CME detects that Unified CM does not support H.450, Unified CME hairpins the calls by hairpinning both signaling and media at Unified CME. This causes double the amount of bandwidth to be consumed when calls are transferred or forwarded across the WAN. (For example, if a Unified CM phone calls a Unified CME phone and the Unified CME phone transfers the call to a second Unified CM phone, Unified CME hairpins both the signaling and media even though the call is between two Unified CM phones.) To avoid this double bandwidth consumption on the WAN, Cisco recommends using the Cisco Unified Border Element to act as an H.450 tandem gateway and to allow for H.450-to-ECS mapping for supplementary services such as call transfer or call forward.

Supported Call Flows

Unified CME is a back-to-back user agent (B2BUA), thus call flows work from SCCP phone to SCCP phone and from SCCP phone to SIP phone. SIP phone calls work over H.323 trunks, but supplementary features are not supported.

Security

Unified CME provides secure signaling with TLS and media encryption with SRTP. Unified CM also supports secure signaling via TLS and secure media via SRTP. However, interworking between secure Unified CM and secure Unified CME is not supported.

Video

Observe the following design considerations when implementing video functionality with Unified CME:

- All endpoints on Unified CM and Unified CME must be configured as video-capable endpoints. The video codec and formats for all the video-capable endpoints must match.
- Unified CM and Unified CME support basic video calls; however, supplementary services such as call transfer and call forward are not supported for video calls between Unified CM and Unified CME. To support supplementary services with Unified CME, H.450 must be enabled on all Unified CMEs and voice gateways. Because Unified CM does not support H.450, video calls will revert to audio-only calls when supplementary services are needed between Unified CM phones and Unified CME phones.
- Conference calls revert to audio only.
- WAN bandwidth must meet the minimum video bit rate of 384 kbps for video traffic to traverse the WAN.

H.320 Video via ISDN

Observe the following design considerations when implementing H.320 video functionality via ISDN:

- When directly connected to an H.320 endpoint via a PRI or BRI interface, Unified CME and Cisco IOS routers currently support only 128 kbps video calls.
- When H.320 is enabled on Unified CME and PSTN gateways to interwork with Unified CM, use a separate dial-peer for video calls to differentiate them from voice-only calls. Configure **bear-cap speech** under the **voice-port** configuration on Unified CME.
- H.320 does not support supplementary services.

General Design Considerations for Unified CM and Unified CME Interoperability via H.323

- Configure Unified CME to auto-detect Unified CM by using H.450.12 to hairpin the calls between Unified CM and Unified CME phones.
- For SCCP-to-SCCP calls or SCCP-to-SIP calls, an H.323 trunk can be deployed between Unified CM and Unified CME.
- While Unified CME supports secure signaling with TLS and secure media with SRTP, conferencing call flows cannot be secured. Further, security interoperability is not supported between Unified CM and Unified CME phones.
- Deploy video only for SCCP phones (with support of basic calls), and not for SIP phones.
- MTP functionality is not compatible with video; for video calls to work, the MTP feature must be disabled (unchecked).
- Make sure that IP connectivity between Unified CM and Unified CME works properly.
- Make sure the local video setup works correctly for each Unified CME local zone and Unified CM location (local SCCP).
- Use the existing voice dial-plan infrastructure.

- Observe the following guidelines for video traffic shaping:
  - Mark the video and audio channels of a video call with CoS 4 to preserve lip-sync and to separate video from audio-only calls.
  - Place voice and video traffic in different queues.
  - Use Priority Queuing (PQ) for voice and video traffic. Two different policies are required for voice-only calls and video (voice stream + video stream) calls based on Classifications. Voice calls are protected from video calls because the voice stream in a video call is marked the same as the video stream in the video call.
- Video should not be deployed in links with less than 768 kbps of bandwidth.
- With link speeds greater than 768 kbps and with proper call admission control to avoid oversubscription, placing video traffic in a PQ does not introduce a noticeable increase in delay to the voice packets.
- There is no need to configure fragmentation for speeds greater than 768 kbps.
- cRTP is not recommended for video packets. (Because video packets are large, cRTP is of no help with video.)
- Voice and video traffic should occupy no more than 33% of the link capacity.
- When calculating video bandwidth, add 20% to the total video data rate of the call to account for overhead.

For more details on integrating Unified CME with Unified CM through H.323, refer to the *Cisco Unified CME Solution Reference Network Design Guide*, available at

http://www.cisco.com/en/US/products/sw/voicesw/ps4625/products_implementation_design_guides_list.html

CHAPTER 9

Dial Plan

Revised: April 30, 2013; OL-27282-05

The dial plan is one of the key elements of a Unified Communications system, and an integral part of all call processing agents. Generally, the dial plan is responsible for instructing the call processing agent on how to route calls. Specifically, the dial plan performs the following main functions:

- Endpoint addressing

  Reachability of internal destinations is provided by assigning directory numbers (DNs) to all endpoints (such as IP phones, fax machines, and analog phones) and applications (such as voicemail systems, auto attendants, and conferencing systems)

- Path selection

  Depending on the calling device, different paths can be selected to reach the same destination. Moreover, a secondary path can be used when the primary path is not available (for example, a call can be transparently rerouted over the PSTN during an IP WAN failure).

- Calling privileges

  Different groups of devices can be assigned to different classes of service, by granting or denying access to certain destinations. For example, lobby phones might be allowed to reach only internal and local PSTN destinations, while executive phones could have unrestricted PSTN access.

- Digit manipulation

  In some cases, it is necessary to manipulate the dialed string before routing the call; for example, when rerouting over the PSTN a call originally dialed using the on-net access code, or when expanding an abbreviated code (such as 0 for the operator) to an extension. Digit manipulation is also used to adapt the local dialing habits of a user to the global routes used to select a path for a call. For example, a French user may dial 0 00 1 212 555 1234 to call a number in New York. That same number is reachable to a caller in Chicago by dialing 9 1 212 555 1234. Both localized user inputs can be translated to a global form of +1 212 555 1234, so that a single route is used to select a path for the call.

- Call coverage

  Special groups of devices can be created to handle incoming calls for a certain service according to different rules (top-down, circular hunt, longest idle, or broadcast). The dial plan information covered in this chapter applies to any Unified Communications deployment model; in particular, when deploying multi-site systems, the system designer should pay close attention to the site-specific dialing habits as well as site-specific routing of calls, such as the use of a gateway co-located with a specific group of users.

This chapter presents information intended to guide the system designer toward a dial plan that accommodates the legacy dialing habits of telephony users, while also taking advantage of new functionality afforded by the increasing integration between computing technology and telephony, such as dialing from contacts, click-to-call actions from computers and smart phones, and adoption of mobility-related features. The chapter is structured to offer information of the following main areas:

- Planning Considerations, page 9-4

  This section analyzes the thought process involved in planning an IP Telephony dial plan, ranging from the number of digits used for internal extensions to the overall architecture of a company's internal dial plan. (Prerequisite: Some familiarity with dial plans in general.)

- Design Considerations, page 9-11

  This section contains design and deployment guidelines related to multisite IP Telephony networks, endpoint addressing methods, approaches to building classes of service, and call coverage functionality. (Prerequisite: A working knowledge of Cisco Unified Communications Manager and Cisco IOS is recommended.)

- Dial Plan Elements, page 9-84

  This section provides detailed background explanations of the elements of a Cisco Unified Communications dial plan. Covered topics include call routing logic, calling privileges, and digit manipulation techniques for various Cisco products. (Prerequisite: A working knowledge of Cisco Unified Communications Manager and Cisco IOS is recommended.) This section does not supersede product-specific documentation, nor does it present all the information available in the Cisco Unified Communications Manager help files. It does highlight some fundamental functionality elements essential to the understanding of design-related concepts presented herein.

For more details, refer to the *Cisco Unified Communications Manager System Guide*, the *Cisco IOS Voice, Video, and Fax Configuration Guide, Release 12.2*, and other product documentation available at

http://www.cisco.com

What's New in This Chapter

Table 9-1 lists the topics that are new in this chapter or that have changed significantly from previous releases of this document.

Table 9-1 New or Changed Information Since the Previous Release of This Document

New or Revised Topic	Described in	Revision Date
A few minor corrections and changes	Various sections	April 30, 2013
Calling party transformations	Calling Party Transformations on IP Phones, page 9-85	October 31, 2012
Case sensitivity of URI dialing	Case Sensitivity, page 9-79	October 31, 2012
SAF Forwarder CLI configuration	SAF Forwarders, page 9-154	September 28, 2012
Minor correction for blocked patterns	Globalized Numbers and Class of Service, page 9-62	August 31, 2012
Calling party number globalization	Phone Calling Party Number Globalization, page 9-17	June 28, 2012
Calling party number in missed and received calls directories on Type-A phones	Phone Calling Party Number Localization, page 9-20	June 28, 2012

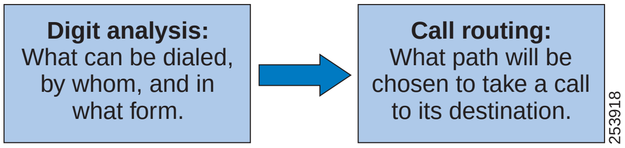
Table 9-1 New or Changed Information Since the Previous Release of This Document (continued)

New or Revised Topic	Described in	Revision Date
Class of service in +E.164 dial plans	Building Classes of Service for Unified CM for +E.164 Dial Plans with the Traditional Approach and Local Route Group, page 9-66	June 28, 2012
URI dialing	Deploying Directory URI Dialing, page 9-78 Directory URIs, page 9-94	June 28, 2012
Routing SIP requests	Routing of SIP Requests in Unified CM, page 9-107	June 28, 2012
Various other updates for Cisco Unified Communications System Release 9.0	Numerous sections throughout this chapter	June 28, 2012

Dial Plan Architecture

Figure 9-1 illustrates the fundamental architecture of dial plans based on Cisco Unified Communications Manager (Unified CM).

Figure 9-1 Basic Dial Plan Architecture


Digit analysis:
What can be dialed, by whom, and in what form.
Call routing:
What path will be chosen to take a call to its destination.
253918

The digit analysis function controls which calls are allowed to a user, to a gateway, or to an application. This function is where call privileges (also known as classes of service) are implemented. The following fundamental constructs are used to implement digit analysis:

- Patterns (such as directory number patterns and translation patterns)

  Patterns are numerical representations of telephone numbers which, when matched, trigger call routing.

- Partitions

  Partitions are used to separate patterns into logical groups. For instance, partitions allow the provisioning of two separate extensions set to 1000.

- Calling search spaces

  Calling search spaces allow control over what groups of patterns a device (such as a phone) can access. For instance, devices at one site can be given access to the local partition containing extension 1000, without having access to a different site's extension 1000.

The call routing function controls the path selection for a call. This function is where IP trunks, PSTN trunks, or even connections to legacy PBXs, are chosen to carry a particular call. The call routing function also allows for the automated failover of calls from, for example, an IP connection as a first choice to a PSTN connection as a backup choice, in case the first choice is not available because no bandwidth is available or because a particular portion of the network is not available.

For both of these functions, Unified CM offers the system designer many tools with which digit manipulation can be effected and with which control over the call processing can be performed for different situations. For example, the system administrator can configure the types of calls allowed from a phone when it is roaming between different sites, how a call is processed when the phone is busy or when it rings with no answer, or which destinations a phone can use when call-forwarding all calls.

The fundamental elements of the individual features of this architecture are presented in multiple documents. The product-specific documentation, along with the help files in Unified CM, offer the most fundamental descriptions of the features. The section on Dial Plan Elements, page 9-84, in this chapter offers the next level of functionality description. The section on Design Considerations, page 9-11 in this document offers the system designer top-down architectural information to be considered when designing a dial plan.

High Availability for Dial Plans

In Cisco Unified CM, dial plan functionality is inherently made available by the clustering capabilities of Unified CM servers. All dial plan configuration is made redundant by the same mechanisms that allow other Unified CM services to be redundant. Specifically, Unified Communications Manager groups are used to control phones, route lists, and gateways, so that a single failure of a Unified CM server does not render any dial plan function unavailable.

For call paths relying on external trunks, an additional level of availability is afforded by the use of alternate routes. For instance, a route list can be used to establish a primary path to a given off-cluster destination, followed by a secondary or even a tertiary choice. If the higher-order choice (such as an IP trunk) is not available, the next order choice will be attempted, until either the call is established successfully or all preconfigured choices have been exhausted.

For calls between on-cluster endpoints, such as calls between two IP phones, if the IP path is not available due to a network failure, the Call Forward Unregistered feature is invoked, allowing the routing of the call through an alternate network such as the PSTN. If the IP path is not available due to lack of bandwidth on the network, the Automated Alternate Routing (AAR) feature is invoked to route the call through an alternate network.

External dial plan resolution subsystems such as H.323 gatekeepers or SIP proxies should also be provisioned with applicable redundancy capabilities, to offer yet another level of high availability.

Capacity Planning for Dial Plans

The configuration of dial plans is typically not onerous on the Unified CM configuration when compared to other capacity-affecting aspects such as the number of gateways, the number of CTI connections, or the rate of call attempts (BHCA). The Cisco Unified Communications Sizing Tool does incorporate dial plan information in its calculations for system provisioning. This tool is available to Cisco employees and partners, with proper login authentication, at http://tools.cisco.com/cucst.

Planning Considerations

The dial plan is the most fundamental attribute of a telephony system. It is at the very core of the user experience because it defines the rules that govern how a user reaches any destination. These rules include:

- Extension dialing — how many digits must be dialed to reach an extension on the system

- Extension addressing — how many digits are used to identify extensions
- Dialing privileges — allowing or not allowing certain types of calls
- Path selection — for example, using the IP network for on-net calls, or using one carrier for local PSTN calls and another for international calls
- Automated selection of alternate paths in case of network congestion — for example, using the local carrier for international calls if the preferred international carrier cannot handle the call
- Blocking of certain numbers— for example, pay-per-minute calls
- Transformation of the called number — for example, retaining only the last five digits of a call dialed as a ten-digit number
- Transformation of the calling number — for example, replacing a caller's extension with the office's main number when calling the PSTN

A dial plan suitable for an IP telephony system is not fundamentally different from a dial plan designed for a traditional TDM telephony system; however, an IP-based system presents the dial plan architect with some new possibilities. For example, because of the flexibility of IP-based technology, telephony users in separate sites who used to be served by different, independent TDM systems can now be included in one, unified IP-based system. These new possibilities afforded by IP-based systems require some rethinking of the way we look at dial plans. This section examines some of the elements that the system planner must consider to properly establish the requirements that drive the design of the dial plan.

Dialed Pattern Recognition

Digit strings dialed by a user on a telephone generally follow patterns. For instance, many enterprises implement a five-digit abbreviated dialing pattern for calls made within the same office location. Also, many enterprises rely on a single-digit access code to represent outside dialing, followed by some quantity of digits to reach a local PSTN number or a long-distance PSTN number (for example, 9 followed by seven digits to reach a local number, or 9 followed by 1 and ten digits to reach a long-distance destination).

The system administrator must plan the system's recognition of these patterns to ensure that the system will act promptly upon detection of a string that corresponds to a predetermined pattern so that users experience no (or minimal) post-dialing delay.

For phones using the Skinny Client Control Protocol (SCCP) and for SIP phones using the Key Press Markup Language (KPML) during dialing, you can implement pattern recognition in Cisco Unified Communications Manager (Unified CM) by configuring route patterns, translation patterns, phone DNs, and so forth. With each digit dialed by the user, the phone sends a signaling message to Unified CM, which performs the incremental work of recognizing a matching pattern. As each key press from the user input is collected, Unified CM's digit analysis provides appropriate user feedback, such as:

- Playing dial tone when the phone first goes off-hook
- Stopping dial tone once a digit has been dialed
- Providing secondary dial tone if an appropriate sequence of digits has been dialed, such as when the off-net access code 9 is dialed

Once digit dialing is completed, Unified CM provides user feedback in the form of call progress tones, such as ringback tone if the destination is in the alerting stage or reorder tone if the destination is invalid.

IP phones running the Session Initiation Protocol (SIP) can be configured with pattern recognition instructions called SIP dial rules. When used, they accomplish the bulk of the task of pattern recognition within the phone. Once a pattern is recognized, the SIP phone sends an invitation to Unified CM to place a call to the number corresponding to the user's input. That action, called a SIP INVITE, is subjected to the Unified CM dial plan in the same way a call from an IP phone running the SCCP protocol would be, except that Unified CM's digit analysis is presented with a completed dial string (that is, all of the digits entered by the user are presented as a block to Unified CM for processing). In this mode of operation, user feedback during the dialing of the digit string is limited to what the phone can provide (see SIP Dial Rules, page 9-90). Once the string has been composed, user feedback can still be provided by Unified CM in the form of call progress tones.

Grouping by Dialing Habits

Most telephony users are used to dialing telephone numbers according to local habits. These are composed of various dialing rules applicable to calls placed to destinations within an office location (intra-site calls), destinations within a company but between different sites (inter-site calls), and destinations outside the company (off-net calls). The form used in dialing these different types of calls varies according to user preferences and local PSTN dialing requirements.

On-Net versus Off-Net Dialing

Calls that originate and terminate on the same telephony network are considered to be on-network (or on-net). By contrast, if a call originates in company A and terminates at company B, it probably has to be routed through different telephony networks: first company A's network, followed by the PSTN, and finally into company B's network. From the caller's perspective, the call was routed off-network (or off-net); from the called party's perspective, the call originated off-net.

In TDM systems, the on-net boundaries of a telephony system are established by the PBX or Centrex system, and they typically do not extend outside of a single site. When they do, they typically do not include sites not immediately on the periphery of a large system hub.

One of the key attributes of IP telephony is its ability to expand the boundaries of calls that can be considered on-net. For instance, telephony users in an enterprise with six dispersed branch offices might be used to reaching one colleague with abbreviated dialing (for example, four-digit dialing) if the called party is located in the same site but dialing a full PSTN number to reach another colleague located in a different site. With an IP-based system where all users are served by the same IP network, it now becomes economically feasible to unite all six branches under a four-digit abbreviated dialing plan, with the IP network as the preferred path and automated overflow to the PSTN as a secondary path if the IP network is congested.

Abbreviated Dialing

Consider an extension with direct inward dial (DID) capability, which can be reached directly from the PSTN. An off-net PSTN caller has to dial the fully qualified PSTN number (for example, 1 415 555 1234) to reach a DID extension. An on-net caller might, however, prefer the ability to reach that same extension by simply dialing the last few digits of the DID number. In a four-digit abbreviated dial plan, the on-net caller would dial only 1234 in this example to reach the same extension.

Dialing can typically be separated into four types:

- Intra-site, on-net abbreviated dialing

  Many systems accommodate four- or five-digit dialing within a site. For example, Cisco employees located in San Jose, California can call the main Cisco reception number using the five-digit string 64000.

- Inter-site, abbreviated on-net dialing

  For example, Cisco employees at any Cisco office can dial the San Jose reception number as 8 526 4000. The digit 8 serves as the inter-site access code, and 52 serves as the site code for San Jose.

  This form is shorter than the alternative of using an off-net form to route the calls on-net (for example, allowing Cisco employees in Canada to reach the Cisco reception in San Jose by dialing 9 1 408 526 4000 while routing the call on-net). Even though the dialing form is similar to that used to reach an off-net destination, the system is configured to keep calls to on-net destinations dialed in the off-net form within the system.

- Inter-site, off-net dialing

  Routing of calls between sites can be handed off to the PSTN. For example, calls made from one site in San Francisco to another site in New York can be dialed either in the on-net or off-net form described above, but routed off-net through the PSTN.

- Off-net dialing

  For calls where the destination is off-net and outside of the company's dial plan, the Unified Communications system must offer a simple, locally significant dialing form to the users.

Avoiding Overlap of Extension Dialing

A telephony system must be configured so that any extension can be reached in an unambiguous manner. To accomplish this goal, the dial plan must satisfy the following requirements:

- All on-net extension dialing must be globally unique. For instance, in a system using an abbreviated four-digit on-net dial plan, there cannot be an extension 1000 in site A and another extension 1000 in site B if the requirement is to reach either of them by dialing only four digits from site C.
- There cannot be any partial overlap between different dial strings.
  - For instance, if 9 is used as an off-net access code in a four-digit abbreviated dial plan (for example, to make PSTN calls), there cannot be any extensions in the 9XXX range. Attempting to do so would create situations where calls are not routed immediately. For example, if a user dialed 9141, the system would have to wait for either more digits (if the user were dialing 9 1 415 555 1234, for example) or the expiration of the interdigit timeout before routing the call to extension 9141. Likewise, if an operator code is used (for example, 0), the entire 0XXX extension range would have to be excluded from a four-digit uniform dial plan.
  - There cannot be overlapping strings of different length. For example, a system with extensions 1000 and 10000 would force users to wait for the interdigit timeout when they dial 1000.

Dialing String Length

The number of dialable extensions determines the quantity of digits needed to dial extensions. For example, a four-digit abbreviated dial plan cannot accommodate more than 10000 extensions (from 0000 to 9999). If 0 and 9 are reserved as operator code and off-net access code, respectively, the number range is further reduced to 8000 (1000 to 8999).

Uniform On-Net Dial Plan

A dial plan can be designed so that all extensions within the system are reached in a uniform way; that is, a fixed quantity of digits is used to reach a given extension from any on-net origination point. Uniform dialing is desirable because of the simplicity it presents to users. A user does not have to remember different ways to dial a number when calling from various on-net locations.

For example, if phone A is reached by dialing 1234 from any on-net location, whether the calling phone is in the same office or at a different site, the enterprise's dial plan is deemed uniform.

When the enterprise consists of few sites, this approach can be used with few complications. The larger the enterprise, in terms of number of extensions and sites, the more of the following challenges it faces in designing a uniform dial plan:

- The number of extensions can exceed the range afforded by the quantity of digits being considered for the dial plan. For instance, if more than 8000 extensions are required (considering the exclusions of the 0XXX and 9XXX ranges), the system may require that an abbreviated dial plan use more than four digits.
- Matching on-net abbreviated extensions to DID numbers means that, when a new DID range is obtained from a local exchange carrier, it cannot conflict with the pre-existing on-net abbreviated dial ranges. For example, if the DID range of 415 555 1XXX exists in a system using a four-digit uniform abbreviated dial plan, and DID range 650 556 1XXX is also being considered, it might be desirable to increase the quantity of digits for on-net dialing to five. In this example, the five-digit on-net ranges 51XXX and 61XXX would not overlap.
- Most systems require the exclusion of certain ranges due to off-net access codes and operator dialing. In such a system where 9 and 0 are reserved codes, no dial plan (uniform or not) could accommodate on-net extension dialing that begins with 9 or 0. This means that DID ranges could not be used if they would force the use of 9 or 0 as the first digit in the dial plan. For instance, in a five-digit abbreviated dial plan, the DID range 415 559 XXXX (or any subset thereof) could not be used. In this example, alternatives include increasing the length of the abbreviated dialing to six or more digits, or avoiding any DID range whose last five digits start with 9, or even not requiring that the DID numbers match the on-net abbreviated extensions.

Once a given quantity of digits has been selected and the requisite ranges have been excluded (for example, ranges beginning with 9 or 0), the remaining dialing space has to be divided between all sites.

Most systems require that two ranges be excluded, thus leaving eight different possibilities for the leading digit of the dial range. Table 9-2 exemplifies the distribution of the dialing space for a typical four-digit uniform dial plan.

Table 9-2 Distribution of Digits in a Typical Four-Digit Uniform Dial Plan

Range	Use	DID Ranges	Non-DID Ranges
0XXX	Excluded; 0 is used as off-net access code		
1XXX	Site A extensions	418 555 1XXX	N/A
2XXX	Site B extensions	919 555 2XXX	N/A
3XXX	Site C extensions	415 555 30XX	3[1-9]XX
4[0-4]XX	Site D extensions	613 555 4[0-4]XX	N/A
4[5-9]XX	Site E extensions	450 555 4[5-9]XX	N/A
5XXX	Site A extensions	418 555 5XXX	N/A

Table 9-2 Distribution of Digits in a Typical Four-Digit Uniform Dial Plan (continued)

Range	Use	DID Ranges	Non-DID Ranges
6XXX	Site F extensions	514 555 6[0-8]XX	69XX
7XXX	Future	XXX XXX 7XXX	7XXX
8XXX	Future	XXX XXX 8XXX	8XXX
9XXX	Excluded; 9 is used as off-net access code		

For the example in Table 9-2, the various sites were assigned numbers in the following ways:

- Site A, the company headquarters, requires more than 1000 extensions, so two entire ranges of numbers have been retained (1XXX and 5XXX). Note that the corresponding DID ranges must also be retained from the site's local exchange carrier.
- Site B has been assigned an entire range (2XXX), allowing for up to 1000 extensions.
- Site C was also assigned an entire range, but it has been split between 100 DID extensions (415 555 30XX) and up to 900 non-DID extensions. If growth requires more extensions for DID, any unassigned numbers from the non-DID range could be used.
- Sites D and E were each assigned 500 numbers from the 4XXX range. Note that their corresponding DID ranges must match each of the site's respective portions of the 4XXX range. Because the DID ranges are for different sites (probably from different PSTN service providers), more coordination effort is required to split ranges between sites. As the quantity of sites assigned within a given range increases, it becomes increasingly difficult (sometimes impossible) to make full use of an entire range.
- Site F's range is split between 900 DID numbers (6[0-8]XX) and 100 non-DID numbers (69XX).
- The ranges 7XXX and 8XXX are reserved for future use.

When implementing a new dial plan, one of the main desires of any planner is to avoid having to change phone numbers. In addition, the extension ranges of any existing phone systems may have overlapped without any problems in the past, but they could be incompatible with a uniform dial plan.

Variable Length On-Net Dial Plan

Systems with many sites or overlapping site extension ranges can benefit from the use of a variable-length dial plan with the following characteristics:

- Within a site, the system retains the use of abbreviated dialing for calls to on-net extensions (for example, four-digit dialing).
- Between sites, users dial an access code followed by a site code and the destination's on-net extension.
- Off-net calls require an access code followed by a PSTN number.

The use of access and site codes (see Table 9-3) enables the on-net dial plan to differentiate between extensions that would overlap if a uniform abbreviated dial plan were implemented.

Table 9-3 Typical Use of Site Codes

Site Code	Range	Use	DID Ranges	Non-DID Ranges
1	1XXX	Site A extensions	418 555 10XX	1[1-9]XX
2	1XXX	Site B extensions	919 555 1XXX	N/A
3	1XXX	Site C extensions	907 555 1XXX	N/A

In Table 9-3, sites A, B, and C are independently assigned the four-digit range 1XXX. For calls from site A to site B under the old telephony system, the calls had to be routed as off-net calls. With the new system, these calls can be dialed as on-net calls.

From site A, users simply dial 1234 to reach extension 1234. But from site B, the dial plan must accommodate the ability to reach extension 1234 at site A without conflicting with site B's own extension 1234. Therefore, each site is assigned a site code.

From site B, merely dialing the combination of site A's code with the desired extension is not feasible: in this case because 11234 would partially overlap with site B's extension 1123, thus causing interdigit timeout issues. If, instead, we assign 8 as an inter-site on-net access code, this would allow site B to dial 81234 to reach site A's extension 1234.

The following factors determine the quantity of digits required to dial an on-net, off-site extension:

- One digit for the inter-site access code
- N digits for the site code, where N is chosen to satisfy the quantity of site codes required (For example, if a system has 13 sites, then a minimum of two digits are required for the site code.)
- The quantity of digits required by the destination site's local dial plan

For example, a system with 75 sites which each use four-digit abbreviated dialing would require a format of 8 + SS + XXXX, where 8 is the on-net access code, SS is a two-digit site code, and XXXX is a four-digit extension number, giving a total of seven digits.

On-Net and Off-Net Access Codes

It is common practice in most enterprise telephony systems to dedicate a digit (for example, 9) as an off-net access code to steer calls to an off-net destination. In the variable-length on-net dial plan, another steering digit (for example, 8) is also required as an on-net access code to dial calls to on-net extensions at other sites. These two access codes, along with the use of an operator access code (for example, 0), implicitly exclude three of the ten possible leading digits of any dialed string. This restriction might not prove convenient for either of the following reasons:

- The users would be required to know the difference between on-net and off-net destinations, and to select the proper access code.
- The exclusion of three entire dialing ranges can become too restrictive or can conflict with some pre-assigned extension ranges. For instance, if a site already uses an abbreviated dialing range beginning with 8, the use of that same digit as an access code would require a change.

In systems where the same off-net access code (for example, 9) is already in use by all sites, it might be desirable to use the same code for both off-net and on-net off-site destinations. This approach has two main implications:

- To avoid partial overlap and interdigit timeout situations, the quantity of digits expected after the access code should be uniform.
- The telephony system must be able to recognize all on-net numbers dialed as off-net numbers and to route them over the IP network. This task can be simple for small systems with only one Unified CM cluster but complex for large systems with multiple Unified CM clusters.

Plan Ahead

Implementing an IP-based system might require changing certain existing user practices. Although it is preferable to plan a new system so that the implementation is as transparent as possible to users, dialing habits might have to be adapted to accommodate the integration of multiple sites that used to be on separate telephony systems. For instance, adapting to a new global, enterprise-wide dial plan might require changing the way a user reaches another user at a different site, the quantity of digits used to make intra-site calls and, in some cases, the extension numbers. To avoid exposing users to multiple generations of dial plan changes, try to anticipate expansion of the enterprise, which could result in the addition of sites in different dialing regions, an increase in the required number of on-net extensions, PSTN renumbering (for example, an area code split), or business expansion into different countries.

Design Considerations

This section presents the following dial plan design considerations for multisite deployments:

- Globalized Design Approach, page 9-12, covers guidelines and best practices that apply to deployments using the globalization dial plan features of Cisco Unified Communications Manager.
- Call Control Discovery, page 9-24, explains how the Service Advertisement Framework (SAF) Call Control Discovery (CCD) service allows clusters to advertise their own hosted DN ranges into the network as well as to subscribe to advertisements generated by other call agents in the network.
- Dial Plan Considerations for the Intercompany Media Engine, page 9-33, describes the Cisco Intercompany Media Engine (IME), which allows participating enterprises to route calls over the Internet between them.
- Design Guidelines for Multisite Deployments, page 9-35, covers guidelines and best practices that apply to all multisite deployment models.
- Choosing a Dial Plan Approach, page 9-38, presents the various approaches to organizing a dial plan for uniform versus variable-length on-net dialing and, for this second option, partitioned versus flat addressing.
- Deploying Dialed Pattern Recognition in SIP Phones, page 9-52, explains how SIP dial rules can be employed to enable SIP phones to recognize certain dialing patterns.
- The following sections analyze in detail two dial plan approaches and provide configuration guidelines for each:
  - Deploying Uniform On-Net Dial Plans, page 9-40
  - Deploying Variable-Length On-Net Dial Plans with Flat Addressing, page 9-43

- The following sections present two alternative ways of configuring classes of service within Unified CM:
  - Building Classes of Service for Unified CM with the Traditional Approach, page 9-54
  - Building Classes of Service for Unified CM with the Line/Device Approach, page 9-57
- Building Classes of Service in Cisco IOS with H.323, page 9-71, explains how to implement classes of service within a Cisco IOS router running the H.323 protocol.
- Deploying Call Coverage, page 9-75, provides guidelines and best practices for implementing call coverage functionality using hunt lists and line groups with Unified CM.

Globalized Design Approach

This section describes dial plan features used to implement simplified call routing based on globalized numbers. The simplification is primarily obtained through the use of a single routing structure for off-net calls, no matter the source of the call. For example, two users in separate countries could use the same route patterns to carry calls to their respective local gateways, instead of requiring site-specific route patterns, each configured to match their respective dialing habits.

The main architectural approach used to attain this globalization can be summarized as follows:

- When a call enters the system, the destination number and the calling number are accepted in their local format but are then immediately globalized by the system.
- Once globalized, the called number is used to route the call to its destination through the use of route patterns expressed in the global form. The global form may be a combination of a global internal, enterprise-specific form such as 81001234 and/or a globalized PSTN representation of a DID number, such as the +E.164 form (for example, +12125551234).
- Once a destination has been identified, the calling and called numbers are localized to the form required by the endpoint, the network, or the system to which the call is to be delivered.

Thus, the guiding principle is:

> Accept localized forms upon call ingress, and globalize them; route the call based on the globalized form; and localize the call to comply to the form required by the destination.

Cisco Unified Communications Manager (Unified CM) offers the following dial plan globalization capabilities:

- Local Route Group, page 9-13
- Support for + Dialing, page 9-13
- Calling Party Number Transformations, page 9-13
- Called Party Number Transformations, page 9-14
- Incoming Calling Party Settings (per Gateway), page 9-14
- Logical Partitioning, page 9-15

Together, these new features enable a Unified CM system to:

- Route calls based on the physical location context of the caller.
- Represent calling and called party numbers in a global form such as that described by the International Telecommunications Union's E.164 recommendation.
- Present calls to users in a format based on local dialing habits.
- Present calls to external networks (for example, the PSTN) in a manner compatible with the local requirements for calling party number, called party number, and their respective numbering types.

- Derive the global form of the calling party number on incoming calls from gateways, based on the calling number digits and the numbering type.
- Control the establishment of calls, as well as the initiation of mid-call features, between endpoints based on policies acting on each endpoint's geolocation, to comply with regulatory requirements in certain countries.

Local Route Group

The Local Route Group offers the ability to create patterns that route off-net calls to a gateway chosen for its proximity to the originating party. For example, a single pattern can be defined to route off-net, intra-country calls for all sites within a given country. Phones at every site can be configured to match this pattern, which then would route the call based on the Local Route Group associated with the calling phone. This allows a phone in site 1 to route calls through the gateway at site 1, while a phone at site 2, still using the same pattern, would route calls through the gateway at site 2. This feature simplifies the configuration of site-specific routing of off-net calls when compared to releases prior to Unified CM 7.0.

Support for + Dialing

Telephone numbers can use the + sign to represent the international dialing access code needed to reach a destination from different countries. For example, +1 408 526 4000 is the international notation for Cisco's main corporate office in the United States. To call this number, an enterprise telephony user from France typically would have to dial 0 00 1 408 526 4000, whereas a caller from the United Kingdom would have to dial 9 00 1 408 526 4000. In each case, + must be replaced with the appropriate off-net access code (as required by the enterprise telephony system) and international access code (as required by the PSTN carrier) relevant for each caller.

The system can route calls directly to destinations defined with +. For example, a user could program a WiFi phone’s speed-dial entry for Cisco's main US office as +1 408 526 4000 and dial it directly when roaming in France, the UK, or anywhere else in the enterprise. In each location, the system would translate the destination number into the locally required digit string to allow the call to be routed properly.

Likewise, phone numbers dialed from a dual-mode phone are routable directly over the mobile carrier network when the phone is in GSM mode, or over the enterprise network when the phone is in WiFi mode, if the called number is represented in the +E.164 form. This allows a user to store a single destination number for a particular contact entry, and dial it no matter to which network the phone is currently attached.

This feature allows users to rely on the system to interpret phone numbers represented in the form described by the ITU E.164 recommendation and to route them properly without requiring the user to edit the number to adapt it manually to the local dialing habits.

Calling Party Number Transformations

The calling party number associated with a call routed through Unified CM might sometimes have to be adapted before it is presented to a phone or to the PSTN. For example, a call from +1 408 526 4000 might have to be presented as coming from 408 526 4000 if the destination phone is in the US or Canada, whereas a call from the same number might have to be presented to a destination phone in France as coming from 00 1 408 526 4000. This is mainly to offer users a presentation of the calling party in the customary form offered by their local PSTN, to maintain user familiarity with identification of the origin of calls ringing in.

Calls offered to gateways might require that the calling party number be manipulated to adapt it to the requirements of the telephony carrier to which the gateway is connected. For example, a call from +1 408 526 4000 offered to a gateway located in France might have to represent the calling number as 1 408 526 4000, with a Calling Party Number Type set to International. Similarly, a call from the same number offered to a gateway located in Canada might have to be represent the calling party number as 408 526 4000, with the Calling Party Number Type set to National.

This feature allows the calling party number to be adapted from the form used to route calls within the Unified CM system, to the form required by phone users or off-cluster networks.

Note Some service providers might not be able to accept calling party numbers representing foreign telephone numbers, due to either technical limitations of their equipment, company policies, or governmental regulations. If calling party numbers cannot be accepted by the provider, the provider will either screen and overwrite the calling party number or reject the call. In some networks two calling party identities can exist for a call: user provided and network provided.

Called Party Number Transformations

The called number associated with a call routed through Unified CM might sometimes have to be adapted before it is presented to the PSTN. For example, a call placed to +1 408 526 4000 requires the called party number be transformed to 1 408 562 4000 with the numbering type set to National if it egresses to the PSTN through a gateway located in Canada. If the same call were re-routed toward a French gateway, the called party number would have to be transformed to 00 1 408 526 4000 with the numbering type set to International.

By manipulating the called party number as well as setting the numbering type for the called number, this feature allows the called party number to be adapted to the form required by off-cluster networks.

Incoming Calling Party Settings (per Gateway)

The calling party number associated with a call as it enters a gateway through a digital interface (for example, ISDN PRI) is also associated with an attribute identifying the calling number’s numbering type as either Unknown, Subscriber, National, or International. When combined, the incoming call's calling number and its associated numbering type allow the system to determine the identity of the caller by stripping and prefixing appropriate digits to the incoming call's calling party number. Incoming Calling Party Settings allow the system to apply separate combinations of stripped and/or prefixed digits to the calling party number for each of the four calling number types.

For example, assume two calls come into a gateway located in Hamburg, Germany. Both feature a calling party number of 691234567. The first call is associated with a numbering type of Subscriber. This means the caller is located in Hamburg, thus the city code of Hamburg (40) is implied, as is the country code of Germany (49). Therefore, a full representation of the incoming call is +49 40 69 1234567, which can be obtained by prefixing +49 40 to the incoming call's calling party number for numbering type Subscriber.

The second call is associated with a numbering type of National. This means the caller is located in Germany, and the number already contains the applicable city code (69 is the city code of Frankfurt), but the country code of Germany (49) is implied. A full representation of the second incoming call is thus +49 69 1234567, which can be obtained by prefixing +49 to the second incoming call's calling party number for numbering type National.

This feature allows the system to globalize incoming calls' calling party numbers based on the incoming party number and numbering type. In previous versions of Unified CM, these settings were implemented through the use of cluster-wide service parameters. Unified CM 7.0 introduced per-gateway settings for this feature, which allow different prefixes for each numbering type to be applied to calls entering different gateways. The settings can be configured on the gateway itself, on the gateway's device pool, or through the cluster-wide service parameters, in order of precedence. A blank entry signifies that no digits will be prefixed; to inherit the settings from the lower-precedence setting, the entry must be set to **Default**.

For all calls within a given numbering type, the prefix and strip-digits operations are applied, with no consideration for the calling party number originally received.

Note Calls coming from SIP trunks or from SIP gateways are all associated with calling party numbering type Unknown.

In particular, the SIP protocol as implemented on SIP gateways and SIP trunks effectively places the incoming calling party number of all calls in the numbering type Unknown. This prevents Unified CM from applying different calling party number modifications for different calling party number categories.

Unified CM 7.1 and later releases allow the use of Incoming Calling Party Settings Calling Search Spaces (CSSs) for each number type. These CSSs are used to apply modifications to the calling party based on Calling Party Transformation Patterns. These patterns use regular expressions to match a subset of cases, followed by separate digit manipulation operations for each subset. This new capability enables Unified CM to apply different calling party number modifications for different calling party number categories. For example, a SIP trunk used to connect to the PSTN could present calls from local, national, and international parties with the numbering type set to Unknown; then each call's calling party number would be used to match a Calling Party Transformation Pattern in the trunk's CSS associated with number type Unknown, thus allowing Unified CM to apply different calling party number modifications for different calling party number categories.

Logical Partitioning

Some countries such as India have Telecom regulations requiring an enterprise's voice infrastructure to use the local PSTN exclusively when connecting calls outside the enterprise. This requires that the voice system be partitioned logically into two systems: one for Closed User Group (CUG) communications within the enterprise, and a second one to access the local PSTN. A call from an enterprise user in location A to another enterprise user in location B could be made within the CUG system; however, a call from an enterprise user in location A to a PSTN destination, no matter the location, must be made through local access to the PSTN in location A.

While existing dial plan tools can be used to prevent a call from completing if it were placed between endpoints outside the CUG, they are not able to prevent new call legs from being established while the call is in progress. For example, assume that an enterprise user in London, England, calls a co-worker in Delhi, India, over the enterprise network. Once the call is established, the user in Delhi conferences in a customer in India, from the same line on which the call from London was received. This mid-call addition (on the same line) of a destination outside the closed user group is not preventable solely by using the existing dial plan tools in Unified CM (such as Calling Search Spaces and Partitions). Unified CM 7.1 and later releases offer logical partitioning functionality, which allows the establishment and enforcement of policies that apply not only to the initial onset of calls, but also to mid-call features such as conference and transfer.

The combination of globalization features available in Unified CM allows the system to accept calls in the local format preferred by the originating users and carriers, to route the calls on-net using global representations of the called and calling numbers, and to deliver the calls to phones or gateways in the local format required by the destination user or network. These three aspects of the dial plan design approach can be summarized as:

- Localized Call Ingress, page 9-16
- Globalized Call Routing, page 9-20
- Localized Call Egress, page 9-20

Localized Call Ingress

Unified Communications systems with multiple sites located in different regions or countries must satisfy different dialing habits from users and different signaling requirements from the service providers to which gateways are connected. Each local case can be different; this requires that the system be able to "translate" the local dialing habits and signaling requirements into a form that allows for the calls to be routed properly. Therefore, the systems must not only provide for many localized ingress requirements but also yield a single globalized form of any destination pattern.

Localized Call Ingress on Phones

Calls originating on endpoints such as phones or video terminals are typically dialed by users accustomed to a certain set of local dialing habits. Enterprise users in the US are used to dialing 9 1 408 526 4000 to reach Cisco's world headquarters in San Jose, California, whereas users in the UK would dial 9 00 1 408 526 4000 and users in France would dial 0 00 1 408 526 4000. Each of those three dialing forms features an enterprise off-net access code (9 for the US and UK, 0 for France), an international access code (00 for the UK and France, none needed for the US because the destination is intra-country), and a representation of the destination number, including the country code (1). Each of those three groups of users are dialing the same globalized destination number (+1 408 526 4000), but each with their own local dialing habits. In each of the three cases, + can be used as a global abstraction of the local dialing habits.

An enterprise telephony system must allow for the local dialing customs of users to be interpreted correctly. In all three cases above, the users are using a local dialing form to reach a common destination. The system must be configured to recognize user input and then route and deliver the call to the proper destination. Because the call can originate in many different forms, the system must provide pattern recognition to match each of those different forms.

Unified CM's translation patterns are used to convert localized user input as dialed from phones, to the global form used to route the calls within the Unified Communications system. These patterns must allow all localized dialing habits to be recognized, including:

- Intra-site on-net dialing
- Off-net local, national, and international dialing
- Local services such as emergency calling, directory and operator services
- Carrier selection codes, and so forth

For the three example calls mentioned above, the following translation patterns would be configured in separate partitions and placed in the calling search space (CSS) of:

- US phones: 9.1!, strip pre-dot, prefix +
- UK phones: 900.!, strip pre-dot, prefix +
- French phones: 000.!, strip pre-dot, prefix +

In each case, the locally significant dialed string is translated into a globalized form of + 1 408 526 4000.

For on-net destinations, such as calls between two users in the same site or calls between users at different sites, translation patterns should be used to derive the globalized on-net form of the destination number. This is applicable whether on-net dialing is achieved using site codes or the fully qualified PSTN address of the phone is used as the on-net number.

For example, assume two users in the San Jose site use five-digit abbreviated dialing to call each other. User A calls User B by dialing 51234. A translation pattern specific to this site is configured to recognize any five-digit string beginning with 5 and to translate the called number to the globalized on-net form of 800151234. The translation pattern is configured as: 5XXXX, prefix 8001.

The translation pattern must be site-specific (included in the CSS of only the phones in site San Jose) to avoid confusion with extension 51234 at other sites in the system. In the example above, the on-net global form is implemented using an inter-site access code (8) and a site code (001). If the system used the fully qualified PSTN address of the phone as the on-net number, the translation pattern would instead prefix +140855, to yield a globalized on-net number of +1 408 555 1234.

Note Variable Length On-net Dialing (VLOD) with flat addressing is the recommended approach where possible, because it simplifies configuration. While VLOD with partitioned addressing is supported, it entails extra configuration complexity.

Phone Calling Party Number Globalization

The calling party number for calls originating from phones is set to the number configured as the directory number of the line from which the call originates. Following the concepts of a globalized dial plan design approach, the calling party information of all calls should be globalized. If the DN format is not identical to the format chosen for the globalized internal calling party information (typically +E.164), then the correct handling of calling party information can be achieved by either making sure that all calling party transformations implemented in the system accept both the directory number format and the globalized +E.164 format as input or by making sure that the calling party information for calls from phones is also correctly set to +E.164. This can be achieved by setting the external phone number mask on the line to +E.164 and then setting the **use external phone number mask** option in a translation pattern that is matched. Using the **use external phone number mask** option on a route pattern will not work because device-level calling party transformations using a calling party transformation calling search space on the gateway or the gateway's device pool override the calling party transformation configured on route patterns, so that the input to the calling party transformation calling search space would be the untransformed directory number.

Starting with Cisco Unified CM 9.0, a new incoming calls calling party transformation calling search space on the phone and phone's device pool can be used to globalize the calling party number for calls originating from phones. This is the recommended way to globalize the calling party information of calls from phones to +E.164, because this method also is compatible with URI-dialed call flows for which calling party transformations in translation patterns are not applicable.

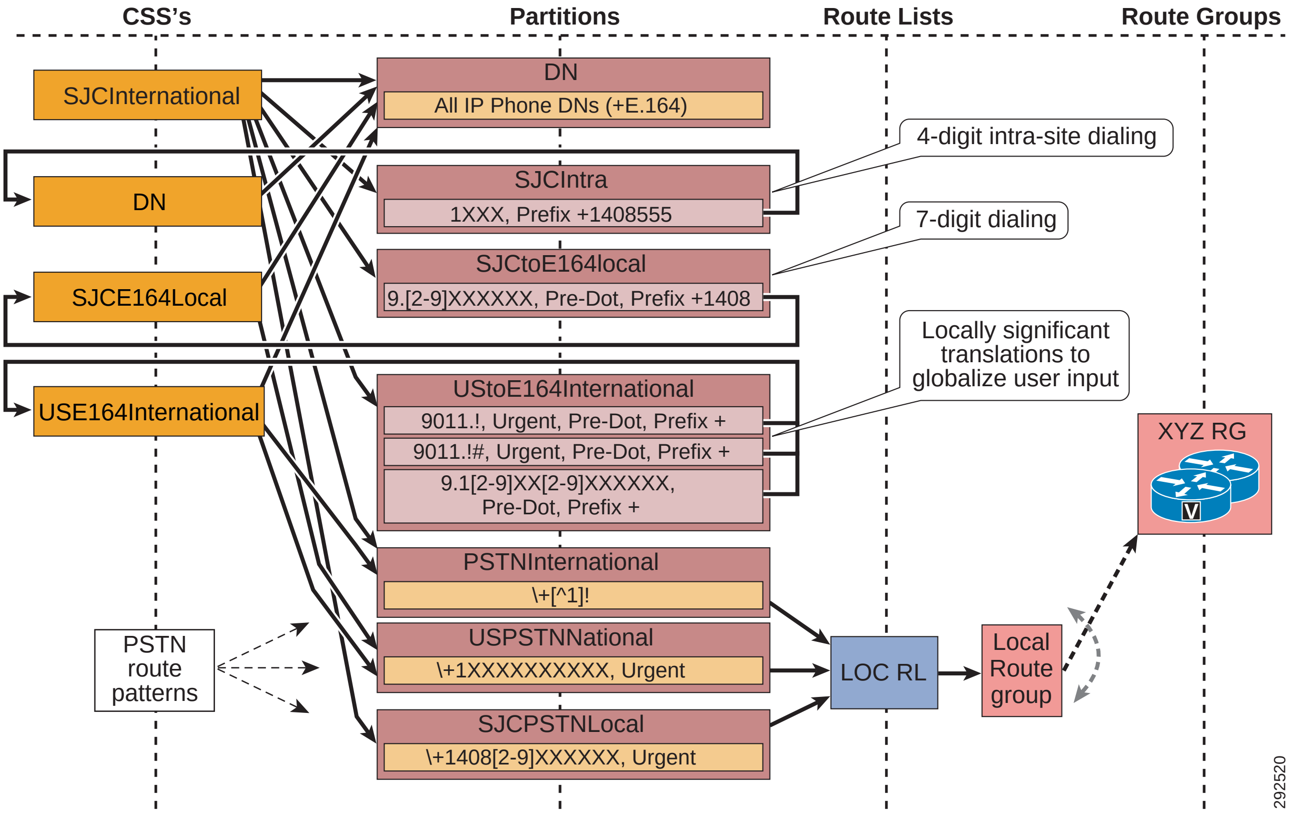
Allowing Call Ingress Using the Global Form

Phones can also provide dialed strings in the global form of the dialed number. In the case of software endpoints such as Cisco Unified Personal Communicator, + dialing can be accommodated directly from the Telephony User Interface (TUI) of the phone or can be derived from click-to-dial actions taken by the user. On Type-B IP phones, dialing + from the keypad can be achieved by pressing and holding either the * or 0 key, depending on the phone model. Also, the missed and received calls directories can contain entries where the number includes a +. As the user dials from those directories, the resulting call into Unified CM will have a called number beginning with +.

Note For definitions of Type-A and Type-B phones, see Dial Plan Elements, page 9-84.

To allow such calls to be handled properly by the phone's dial plan, you must ensure that not only the localized form of dialed numbers is allowed, but also the globalized form. Figure 9-2 illustrates how to accomplish this.

Figure 9-2 Allowing Localized and Globalized TUI



In Figure 9-2, a US IP phone user dials 9011496100773, connects to the destination in Germany, and then releases the call. The called party calls the US user back, connects, and then releases the call. The US user then goes into the Received calls directory, selects the entry for the last received call (+49 6100 773), and presses Dial. In this example, the US user initiates two separate calls to the same destination. For the first call, the form of the destination number localized for US dialing habits is used, and the corresponding translation pattern 9011.! is matched by the user's input. Once translated, the route pattern \+[^1]! is used to route the call. For the second call, the globalized form of the destination number is used and the route pattern \+[^1]! is used directly.

The calling search spaces configured for each site should generally allow for:

- Localized intra-site dialing habits of the site
- Localized off-net dialing habits of the users at the site
- Applicable local telephony services such as emergency calls, directory and operator services
- The globalized form of on-net and off-net numbers

Except for the first item in the list above, the localized patterns used to achieve call routing can typically be reused between sites in the same dialing domain. (All sites in France dial off-net numbers the same way, as do all sites in the UK, in the US, and so forth.) However, each site must be configured with its own intra-site abbreviated dialing translation patterns so that there will be no confusion when a user in San Jose, for example, dials 51234, compared to when a user in New York dials 51234. The translation from the abbreviated intra-site form of a number to the globalized on-net form of the same destination must be achieved with site-specific translation patterns, which requires that each site be configured with its own site-specific calling search space.

Localized Call Ingress on Gateways

The called and calling numbers delivered into the Unified Communications system by external networks (for example, the PSTN) are typically localized. The form of the numbers may vary, depending on the service provider's configuration of the trunk group. As a gateway is connected to a PSTN trunk group, the system administrator must work with the PSTN service provider to determine the applicable signaling rules to be used for this specific trunk group. As calls are delivered into the system from the trunk group, some of the information about the calling and called numbers will be provided explicitly and some of it will be implied. Using this information, the system must derive the calls' globalized calling and called party numbers.

The globalization of the called party number can be implemented through one of the following methods:

- In the gateway configuration, configure **Call Routing Information** > **Inbound Calls**, where the quantity of significant digits to be retained from the original called number and the prefix digits to be added to the resulting string are used to globalize the called number. The prefix digits should be used to add the applicable + sign and country, region, and city codes.
- Place translation patterns in partitions referenced by the gateway's calling search space. The translation patterns should be configured to match the called party number form used by the trunks connected to the gateway, and should translate it into the global form. The prefix digits should be used to add the applicable + sign and country, region, and city codes.
- On H.323 trunks and gateways, use the incoming call's called party transformation settings available on the gateway and on the gateway's device pool. There you can define strip and prefix digit instructions or alternatively configure a called party transformation calling search space per numbering type.

The globalization of the calling party number should be implemented by using the Incoming Calling Party Settings configured either on the gateway directly or in the device pool controlling the gateway.

Note If the administrator sets the prefix to **Default**, this indicates call processing will use the prefix at the next level setting (device pool or service parameter). Otherwise, the value configured is used as the prefix unless the field is empty, in which case there is no prefix assigned.

For example, assume a call is placed to Cisco's US headquarters (+1 408 526 4000) from a US number, and the call is delivered to a gateway located in San Jose, California. The called number provided to the gateway is 526 4000. This information is sufficient for the Cisco Unified Communications system to derive the full destination number for the call. A call delivered by the service provider on this specific trunk group should inherit an implied country code and area code based on the characteristics of the trunk group connected to the gateway, which presumes that all destination DID numbers handled by the trunk group are from the North American Numbering Plan country code (1) and from area code 408. Therefore, the derived global form of the number is +1 408 526 4000. The calling number provided to the gateway is 555 1234, with the numbering type set to Subscriber. The numbering type allows the system to infer the country code and area code from the configured characteristics of the trunk group. Thus, the system knows that the calling number is +1 408 555 1234.

On a different call, if the calling number is 33158405858 with numbering type International, this is an indication that the global form of the calling number should represented as +33158405858.

Globalized Call Routing

For the destination to be represented in a global form common to all cases, we must adopt a global form of the destination number from which all local forms can be derived. The + sign is the mechanism used by the ITU's E.123 recommendation to represent any global E.164 PSTN number in a global, unique way. This form is sometimes referred to as a fully qualified PSTN number.

The system can be configured with route patterns that match globalized called numbers including the + sign. These same route patterns can point to route lists and route groups featuring the Standard Local Route Group. This allows for the creation of truly global route patterns because the egress gateway can be determined from the calling endpoint’s device pool at the time of the call. All the necessary tasks of adapting the calls (both the calling and the called party numbers) to the local preferences and requirements are performed once a destination has been selected.

Localized Call Egress

When calls are routed to a destination using a global form of the called and the calling numbers, you might have to consider the following localization actions when the call is delivered to its destination.

Phone Calling Party Number Localization

As a call is delivered to a phone, the calling number will be in its global form, which might not be recognizable to the called party. Typically, users prefer to see calls from callers within their country presented with an abbreviated form of the caller's number.

For example, users in the US want to see incoming calls from US callers with a ten-digit national number, without the + sign or the country code (1). If a user whose global phone number is +1 408 555 1234 calls +1 408 526 4000, the called phone would like to receive 408 555 1234 as the calling party number while the phone is ringing. To achieve this, the system administrator should configure a Calling Party Transformation Pattern of: \+1.!, strip pre-dot. The Calling Party Transformation Pattern is placed in a partition included in the destination phone's Calling Party Transformation Pattern CSS, configured at the device-pool level. As a call from +1 408 555 1234 is offered to the phone, it matches the configured Calling Party Transformation Pattern, which removes the +1 and presents a calling party number of 408 555 1234 as the call rings in.

Note The calling party number stored in the missed and received calls directories of Type-B phones is left in its globalized form to allow one-touch dialing from the directories without requiring manual editing of the directory's stored number string. The calling party number stored in the missed and received calls directories of Type-A phones is the transformed calling party number. For Type-A phones, the transformed calling party number needs to be in the form of a supported dialing habit to make sure the user can call back from the missed and received calls directory. This behavior permits implementation of globalized dial plans even with older Type-A phones present in the network.

Note Many phone users are becoming accustomed to the globalized form of PSTN numbers, mainly due to the common use of mobile phones across international boundaries. The system administrator can forego the configuration of Calling Party Transformation Patterns for phones if displaying the global form of incoming numbers is preferred.

Gateway Calling Party Number Localization

As a call is delivered to a gateway, the calling party number must be adapted to the requirements of the PSTN service provider providing the trunk group to which the gateway is connected. Calling Party Number Transformation patterns can be used to change the calling party number digit string and numbering type. Typically, a calling party number featuring the gateway's country code should be changed to remove the + sign and the explicit country code, and they should be replaced with the national prefix. Also, the numbering type of the calling party number should be changed to National. If the gateway is connected to a trunk group featuring a specific area, region, or city code, the specific combination of + sign, country code, and local area code usually must be replaced by the applicable local prefix. Also, the numbering type must be adjusted to Subscriber.

For example, assume that a call from a San Francisco user (+1 415 555 1234) is routed through a route list featuring a San Francisco gateway as a first choice and a Chicago gateway as a second choice. The San Francisco gateway is configured with two Calling Party Transformation Patterns:

- \+1415.XXXXXXX, strip pre-dot, numbering type: subscriber
- \+1.!, strip pre-dot, numbering type: national

As the call is delivered to the San Francisco gateway, the calling party number matches both Calling Party Transformation Patterns. However, the first one is a more precise match and is selected to process the calling party number. Thus, the resulting transformed number is 5551234, with a calling party type set to Subscriber.

If the gateway had not been able to process the call (for example, if all ports were busy), the call would have been sent to the Chicago gateway to egress to the PSTN. The Chicago gateway is configured with the following two Calling Party Transformation Patterns:

- \+1708.XXXXXXX, strip pre-dot, numbering type: subscriber
- \+1.!, strip pre-dot, numbering type: national

As the call is delivered into the Chicago gateway, the calling party number matches only the second Calling Party Transformation Pattern. Therefore, the resulting calling party number offered to the gateway is 4155551234, with a calling party number type set to National.

Gateway Called Party Number Localization

As a call is delivered to a gateway, the called party number must be adapted to the requirements of the PSTN service provider providing the trunk group to which the gateway is connected. Called Party Number Transformation patterns can be used to change the called party number digit string and numbering type. Typically, a called party number featuring the gateway's country code should be changed to remove the + sign and the explicit country code, and they should be replaced with the national prefix. Also, the numbering type of the called party number should be changed to National. If the gateway is connected to a trunk group featuring a specific area, region, or city code, the specific combination of + sign, country code, and local area code usually must be replaced by the applicable local prefix. Also, the numbering type must be adjusted to Subscriber.

For example, assume that a call to a San Francisco user (+1 415 555 2222) is routed through a route list featuring a San Francisco gateway as a first choice and a Chicago gateway as a second choice. The San Francisco gateway is configured with two Called Party Transformation Patterns:

- \+1415.XXXXXXX, strip pre-dot, numbering type: subscriber
- \+1.!, strip pre-dot, numbering type: national

As the call is delivered to the San Francisco gateway, the called party number matches both of the Called Party Transformation Patterns. However, the first one is a more precise match and is selected to process the called party number. Thus, the resulting transformed number is 5552222, with a called party type set to Subscriber.

If the gateway had not been able to process the call (for example, if all ports were busy), the call would have been sent to the Chicago gateway to egress to the PSTN. The Chicago gateway is configured with the following two Called Party Transformation Patterns:

- \+1708.XXXXXXX, strip pre-dot, numbering type: subscriber
- \+1.!, strip pre-dot, numbering type: national

As the call is delivered into the Chicago gateway, the called party number matches only the second Called Party Transformation Pattern. Therefore, the resulting called party number offered to the gateway is 4155552222, with a called party number type set to National.

Note When a call egresses to a gateway, the calling and called party transformation patterns are applied to the calling and called numbers respectively.

Note SIP does not offer an indication of the numbering type. Therefore, SIP gateways will not be able to receive an indication of the called or calling party number type set by Unified CM.

Benefits of the New Design Approach

The benefits of the dial plan design approach enabled by the new globalization features in Unified CM 7.*x* include:

- Simplified configuration of call routing, especially when considering local egress to the PSTN
- Simplified configuration and enhanced functionality of system functions such as:
  - Automated Alternate Routing (AAR)
  - Emergency Responder (ER) site-specific failover
  - Call Forward Unregistered (CFUR)
  - Tail End Hop Off (TEHO)
  - Click-to-dial of E.164 numbers (including the + sign) from soft clients such as Cisco Unified Personal Communicator
  - Adaptive call routing for speed dials originating from roaming extension mobility users or roaming devices
  - One-touch dialing from phone directory entries, including dual-mode phones

- One-touch dialing from missed and received call lists in IP phone directories

Automated Alternate Routing (AAR)

If the AAR destination mask is entered in the globalized form, and if every AAR CSS is able to route calls to destinations in the globalized form, then the system administrator can forego the configuration of AAR groups because their sole function is to determine what digits to prefix based on the local requirements of the calling phone's PSTN access to reach the specific destination.

Furthermore, in most cases the sole function of the AAR CSS is to route the call to the calling phone's co-located gateway; therefore, it can be configured with only a single route pattern (\+!) pointing to a route list that contains the Standard Local Route Group. Because calls routed by this single route pattern will always be routed through the Local Route Group associated with the calling endpoint, that unique AAR CSS can be used by all phones at all sites, no matter in which region or country they are located.

Cisco Emergency Responder

Call routing to Cisco Emergency Responder (ER) is typically implemented by configuring a 911 CTI route point to connect to the primary ER server and a 912 CTI route point to connect to the backup ER server.

If both ER servers are unavailable, 911 calls can be directed to the PSTN egress gateway co-located with the calling phone by configuring:

- The 911 CTI route point to Call Forward No Answer (CFNA) and Call Forward Busy (CFB) to 912, through a calling search space that contains the partition of the 912 CTI route point
- The 912 CTI route point to CFNA and CFB to 911, through a calling search space that contains a global partition, itself containing a route pattern 911 pointing to a route list that contains the Standard Local Route Group

If both CTI route points become unregistered, calls to 911 will be forwarded through the local route group as determined by the calling phone's device pool. If Device Mobility is configured, roaming phones will be associated with the visited site's device pool, and thus associated with the visited site's Local Route Group.

Call Forward Unregistered (CFUR)

To allow calls handled by the Call Forward Unregistered function to use a gateway co-located with the calling phone, configure the CFUR destination of phones using the globalized + form of their PSTN number. The CFUR CSS can be configured with only a single route pattern (\+!) pointing to a route list that contains the Standard Local Route Group. Because calls routed by this single route pattern will always be routed through the Local Route Group associated with the calling endpoint, the same CSS can be used as the CFUR CSS by all phones at all sites, no matter in which region or country they are located.

Tail End Hop Off (TEHO)

To reduce PSTN connectivity charges, system administrators might want to route calls to off-net destinations by using the IP network to bring the egress point to the PSTN as close as possible to the called number. At the same time, if the call's preferred TEHO route is not available, it might be necessary to use the calling phone's local gateway to send the call to the PSTN. This can be achieved by allowing all phones partaking in TEHO routing for a given type of number to match the same route pattern that matches the specific destination number and that points to a route list containing the TEHO egress gateway-of-choice as the first entry and the Standard Local Route Group as the second entry.

Call Control Discovery

In environments where multiple call clusters are deployed, the dial plan must be engineered to route calls between clusters over the IP network where possible, and to use the PSTN as a backup route where required.

Configuration of a cluster to allow intercluster call routing requires the addition of sets of patterns describing the DN ranges hosted in remote clusters. For each remote DN range, the local cluster must be configured with:

- Number range pattern(s) to be recognized as hosted on a remote cluster
- The primary route to reach each remote cluster destination number range, along with associated trunks and protocols
- Secondary route(s) to the destination number ranges, with their associated digit manipulation to transform the destination number to a form acceptable by the PSTN carrier.

This configuration can be done manually, by using static dial plan entries such as route patterns. When done manually, the configuration effort increases with the quantity of remote ranges to be routed. This increase is linear when all clusters are pointed to a centralized dial plan resolution platform, such as a gatekeeper, a SIP proxy, or Cisco Unified CM Session Management Edition. However, this introduces a dependency on a central control point.

The alternative is to create a full mesh, where each cluster pair is configured with intercluster trunks referenced by route patterns defining the remote DN ranges of the cluster on the other side. In this mode, no central point controls the intercluster dial plan resolution, but the configuration effort grows exponentially with the quantity of clusters because a full mesh of trunks and associated DN ranges is required to link all cluster pairs.

Cisco Unified Communications Manager offers the ability for clusters to automatically exchange the DN ranges they host by subscribing to a network-based Service Advertisement Framework (SAF) Call Control Discovery (CCD) service. SAF CCD enables clusters to advertise their own hosted DN ranges into the network as well as to subscribe to advertisements generated by other call agents in the network. The main benefits of using SAF CCD are:

- Automated distribution of call routing information between call agents participating in the same SAF CCD network, thus avoiding incremental configuration work when new call agents are added or when new DN ranges are added to a call agent
- No reliance on a centralized dial plan resolution control point
- Automated recovery of inter-call agent call routing information when routing changes occur, including when multiple Unified CM clusters are combined

The section on Dial Plan Elements, page 9-84, presents some fundamental systemic functionality aspects of SAF CCD to complement the product information available in the latest version of the *Cisco Unified Communications Manager Administration Guide*, available at

http://www.cisco.com/en/US/products/sw/voicesw/ps556/prod_maintenance_guides_list.html

For more detailed information on the Service Advertisement Framework and Call Control Discovery, see Call Routing and Dial Plan Distribution Using Call Control Discovery for the Service Advertisement Framework, page 5-52.

This section of the chapter is not meant to present the exhaustive product configuration of Unified CM for participation in the SAF CCD service. Rather, it focuses on the fundamental system implications of using Unified CM as a call agent in a network offering the SAF CCD service. Design considerations for the dial plan aspects of the SAF CCD service are presented in the following section.

SAF CCD Design Considerations

SAF CCD allows for the exchange of directory number (DN) information between call agents such as Cisco IOS Gateways, Unified CME, and Unified CM. To ensure optimal performance, the system should be designed with the following criteria in mind:

- Advertising Globalized Numbers, page 9-25
- Learning Remote DN Ranges Through the SAF CCD Requesting Service, page 9-30
- Placing Calls to SAF CCD Learned DN Ranges, page 9-31

Advertising Globalized Numbers

Because the DN ranges exchanged between the call agents participating in the SAF CCD service are sent to all call agents, with no regard to site-specific dialing habits, the form of the DNs exchanged through the SAF CCD service should be a global one; that is, a form that is unique among all call agents. This form can be used from any device, on any call agent, in any location in the network. Cisco recommends that all patterns advertised into a SAF CCD service be globally unique within the enterprise.

For example, assume user Paul in Liverpool, England, can be reached by calling the following numbers:

- 1234 if called by coworker John, also located in Liverpool, England
- 85551234 if called by co-worker Wolfgang from Vienna, Austria, or by any other coworker in any on-net enterprise office location worldwide, no matter what call agent controls their phone

User Brian in Hawthorne, CA, can be reached by calling the following numbers:

- 1234 if called by coworker Carl, also located in Hawthorne, CA
- 84441234 if called by coworker Bono located in Dublin, Ireland, or by any other coworker in any on-net enterprise office location worldwide, no matter what call agent controls their phone

In this example, the localized, four-digit abbreviated intra-site form of the number associated with Paul cannot be used as a global identification to be sent to all call agents because it conflicts with that of user Brian. If Paul's call agent were to send a SAF CCD advertisement into the network for DN 1234, it would conflict with that advertised by Brian's call agent in the same SAF CCD network. If user Carl were to dial 1234, there would be some conflict as to whom (Paul or Brian) he is trying to reach.

To avoid such conflicts, call agents should always advertise numbers in a form that is global; that is, a form which does not rely on a context specific to a site or a cluster. It should be a form that can be dialed from any phone on the network, and that uniquely identifies the destination DN in the entire network. The following two main forms of global DNs can be used for this purpose:

- Site-Code Based On-Net Form, page 9-25
- +E.164 Based On-Net Form, page 9-26

Site-Code Based On-Net Form

In systems where the majority of inter-site calls are dialed by using an on-net scheme based on site codes, it is better to advertise DNs in their site-code form, along with a set of rules to allow their failover to the PSTN.

Each DN is globally reachable within the system by dialing an inter-site access code, followed by a site code, followed by a local extension. For example, user Paul in Liverpool can be reached from anywhere on the enterprise network by dialing inter-site access code 8, followed by site code 555 (Liverpool, England), followed by the local extension 1234. Combined, these parts yield a global DN of 85551234, which is globally unique within the network.

In systems using site codes implemented with Variable Length On-Net Dialing (VLOD) with flat addressing, the DN ranges advertised by the cluster to the rest of the network directly match the DN form configured on the lines. This ensures that, when calls are received from other call agents into a SAF CCD trunk, no digit manipulation is required to adapt the called number to the form used internally on the lines. For example, if the cluster advertised 855512XX and it receives a call for 85551234 on a SAF CCD trunk, a match can be made directly into the single partition containing the phones.

In systems using site codes implemented with Variable Length On-Net Dialing (VLOD) with partitioned addressing, the DN ranges advertised by the cluster to the rest of the network do not directly match the DN form configured on the lines. When calls are received from other call agents into a SAF CCD trunk, the called number must be translated from the globalized form to the site-specific abbreviated form used internally on the lines. For example, if the cluster advertised 855512XX and it receives a call for 85551234 on a SAF CCD trunk, a match must first be made into the single partition containing a series of translation patterns that can adapt the called number from the incoming global form to the localized form 1234 used in the destination phone's site-specific partition.

+E.164 Based On-Net Form

In systems where the majority of inter-site calls are dialed by using the PSTN form of DNs, it is better to advertise DNs in their associated +E.164 form. The +E.164 form carries in it all the information that allows any user in any system (on-net or off-net) to reach the destination DN across any network. Cisco recommends that the DN ranges learned from the SAF CCD service be stored as-is, in their +E.164 form, and that local user input be globalized to match it.

For example: user Paul in Liverpool uses a phone whose line DN is defined as 1234 in the Liverpool partition in the English cluster. However, any coworker in a different site calls Paul by dialing the locally significant form of Paul's +E.164 form DID (+44 15 4555 1234). For Wolfgang in Austria, it is 0 00 44 15 4555 1234, and for Elvis in Memphis, TN, it is 9 011 44 15 4555 1234. User Ringo, calling from a different site in Liverpool, dials 9 0154 555 1234 to reach Paul. For user Edge, on the road somewhere in the world, the call to Paul takes on the form of a click-to-call action from a laptop, to +44 15 4555 1234.

The literal form in which Paul's +E.164 is advertised is not necessarily used as-is by users in the other clusters on the network. All but user Edge in the example above used a localized form of Paul's number. But in every case, the local form dialed can be globalized to arrive at the global form advertised by Paul's cluster.

Each cluster must accept user input in a local, habitual form. For example, for user Elvis in the United States, the habitual manual enterprise user input to reach a user in another country involves dialing an off-net access code (9), followed by an international routing code (011), followed by the E.164 number of the destination (44 15 4555 1234). In this case, the globalization of the user input requires the matching of a pattern such as 9011.!, strip pre-dot, prefix +. This one translation pattern can be used for all calls from any USA user to any destination outside the NANP country code 1.

For all users in all clusters, the locally-significant globalization rules required to globalize habitual user input to an +E.164 form are few. They must cover the globalization of local PSTN calls, national calls, and international calls. In many countries, there is only one form for all intra-country, national calls.

Note Advertising DN ranges in the +E.164 form does not require that the DNs themselves be defined as +E.164 numbers in their host cluster.

In systems using the +E.164 form implemented with Variable Length On-Net Dialing (VLOD) with flat addressing, the DN ranges advertised by the cluster to the rest of the network directly match the DN form configured on the lines. This ensures that, when calls are received from other call agents into a SAF CCD trunk, no digit manipulation is required to adapt the called number to the form used internally on the lines. For example, if the cluster advertised +4415455512XX and it receives a call for +441545551234 on a SAF CCD trunk, a match can be made directly into the single partition containing the phones.

In systems using the +E.164 form implemented with Variable Length On-Net Dialing (VLOD) with partitioned addressing, the DN ranges advertised by the cluster to the rest of the network do not directly match the DN form configured on the lines. When calls are received from other call agents into a SAF CCD trunk, the called number must be translated from the globalized form to the site-specific abbreviated form used internally on the lines. For example, if the cluster advertised +4415455512XX and it receives a call for +441545551234 on a SAF CCD trunk, a match must first be made into the single partition containing a series of translation patterns that can adapt the called number from the incoming global form to the localized form 1234 used in the destination phone's site-specific partition.

When Both Forms Are Required

In some systems, users reach each other by both approaches listed above. For these situations, the host clusters should advertise both the site-code form of the DN ranges as well as the +E.164 form. Because the two forms are differentiated by the use of the + sign, no overlap situation can occur between the two DN ranges advertised for a given group of phones.

Special Considerations for Non-DID Numbers

When the system requires the inter-cluster reachability of non-DID numbers, the non-DID DN ranges can be configured to use SAF CCD. However, the PSTN failover functionality will not work in the same manner as for DNs associated with a DID. For example, if a non-DID DN range such as 800033XX is advertised and there is no associated DID range to route calls through the PSTN to the lines in the host cluster, you can do either of the following:

- Configure the PSTN failover digits to yield a call to an annunciator message in the calling cluster, indicating that the call should be re-attempted later due to network congestion, or
- Configure the PSTN failover digits to yield a call to an IVR, reception phone, or other device.

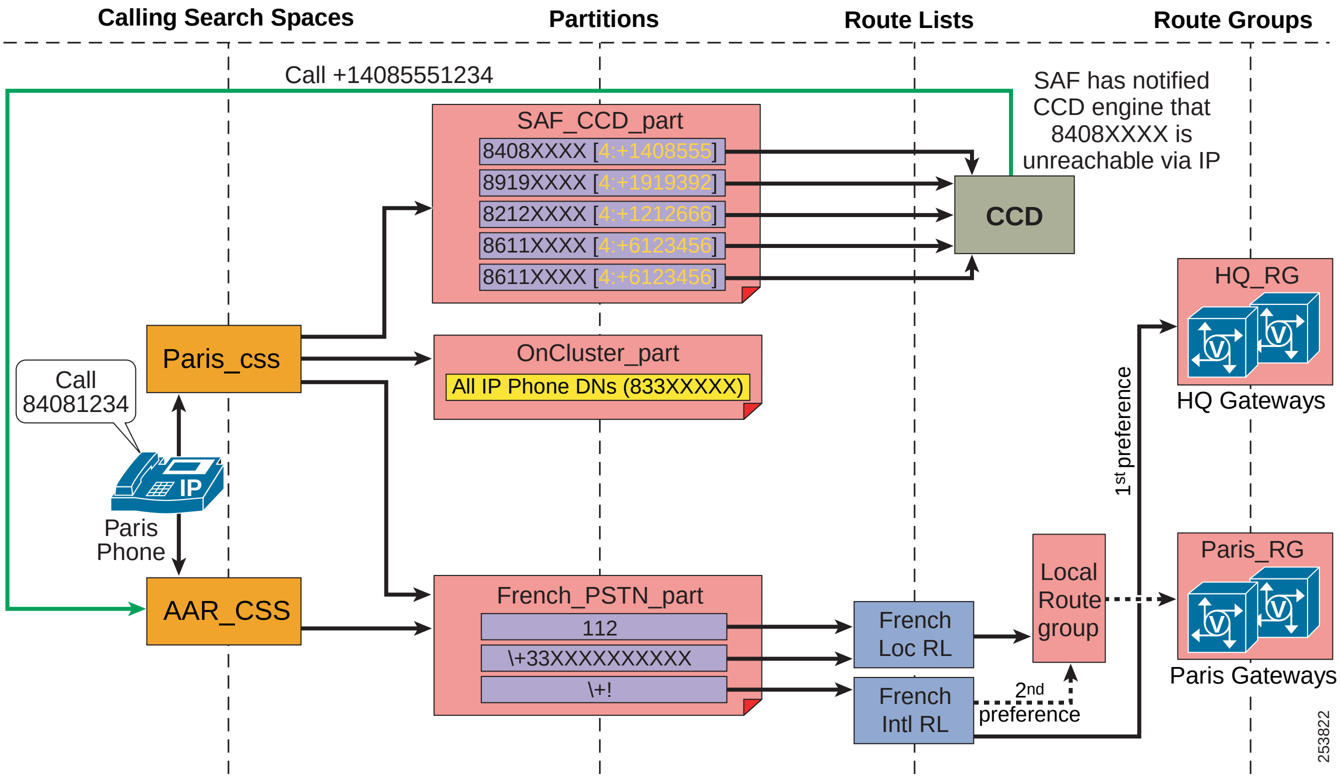
Note The +E.164 format allows the use of the +0 range to designate non-DID numbers. The +0 range can thus be used to route the calls in the +E.164 form on-net only; the PSTN will not route calls to country code 0.

Tip When configuring non-DID ranges, subdivide the +0 range into the actual country, area, and/or city codes where the DNs are hosted. For example, a non-DID range in Chicago, IL, could start with +01708XXXXXXX, allowing for 10 million non-DID numbers. In Frankfurt, Germany, the range could be +04969XXXXXXX, and so forth.

PSTN Failover Considerations for SAF CCD Outgoing Calls

When a call is placed to a SAF CCD discovered number, the call is routed through one of the SAF CCD trunks associated with the requesting service, as configured under **Call Control** > **Call Control Discovery** > **Requesting Service** in Unified CM. (See Figure 9-3.) If the trunk cannot accept the call (for example, if call admission control denies the call or if the trunk is down), then the call will be sent to the PSTN through the AAR calling search space (CSS) of the calling device. Because the destination number might not be in a form directly compatible with the PSTN, the destination number must first be adapted.

Figure 9-3 SAF CCD and PSTN Failover



The DN range records injected into the SAF CCD service by each call agent must carry the rules required to adapt the on-net form of the number to a form acceptable to the PSTN. The rules include what quantity of digits to strip from the left of the range (PSTN Failover Strip Digits), what digits to prefix to the post-strip called number (PSTN Failover Prepend Digits), and whether to use the DN range as-is when rerouting calls to the PSTN (Use HostedDN as PSTN Failover checkbox).

For example, in Figure 9-3 a Unified CM cluster discovers routes to other clusters. The discovered routes are placed into the partition named SAF_CCD_part. If the Paris user dials 84081234, the best-match routing logic will route the call using the SAF CCD discovered pattern 8408XXXX. If the IP route is not available, the dialed number will be combined with the ToDID information, which instructs Unified CM to strip the left-most four digits (in this case, 8408) and to prefix +1408555. This yields +14085551234, which is used to find a match through the calling phone's AAR calling search space. The call then matches the \+! route pattern and is routed through the French INtl RL route list; the first attempt will be to route the call through the HQ_RG route group, followed by an attempt to use the calling phone's local route group (in this case, Paris_RG).

These rules are configured for each advertised DN range under **Call Routing** > **Call Control Discovery** > **Hosted DN pattern** in Unified CM. They can also be configured for groups of DN ranges under **Call Routing** > **Call Control Discovery** > **Hosted DN group** in Unified CM.

Cisco recommends configuring the PSTN failover digits at the **Hosted DN pattern** level because it provides more granular control of the failover digits for each range and avoids another configuration step at the Hosted DN group level.

When Advertising DN Ranges Using Site Codes

For instance, the users in Liverpool can be reached on-net by dialing a number in the 855512XX range. Nothing in this form inherently defines the associated DID number required to route a call to this destination across the PSTN. To transform this site-code form into the +E.164 form required by the PSTN, (+44 15 4555 12XX), the advertised DN range should be stripped of its first (left-most) digit and prefixed with +44154. This is sometimes represented as S:PP, where S represents the quantity of digits to be stripped from the left and PP represents the literal digits to be prefixed to the post-strip called number. In this example, the transformation would be 1:+44154.

Note The cluster advertising the DN range must provide the SAF CCD records with the appropriate information to fail-over to the PSTN. If this information is not provided as the routes are injected into the SAF CCD service, each of the CCD client clusters would have to be configured to adapt the PSTN failover characteristics of the learned routes. This creates an additional configuration burden and requires multiple clusters be modified if any changes are required.

Once the called number form is changed to the +E.164 form, the call is routed through the calling device's AAR CSS. A match must be made on the +E.164 form of the number. Once a route pattern is matched, the call is routed through a route list, a route group, and eventually a trunk or gateway, where called party transformation patterns are used to adapt the number from the +E.164 to the localized form required by the PSTN carrier.

When Advertising DN Ranges Using the +E.164 Form

In this case, the DN ranges are advertised directly in the +E.164 form, therefore no PSTN failover digit configuration is required. It is best to check the **Use HostedDN as PSTN Failover** checkbox under the Hosted DN Range configuration to prevent any failover PSTN digit configuration done at the Hosted DN group level from taking precedence. This may be useful when a system requires both the site-code and the +E.164 forms of a number range to be advertised through the same Hosted DN Group, to accommodate both dialing forms between clusters. In such a case, the Hosted DN group PSTN Failover configuration would apply to the site-code DN ranges and would be ignored for the +E.164 DN ranges.

Note In SAF CCD PSTN failover digit configuration, the + sign is used to perform two main tasks: it allows the adapted PSTN failover number to avoid overlap with any other intra-site valid range in any other cluster (for instance, a cluster where 4415 is a valid intra-site extension would not overlap with +441545551234), and it allows the use of + as a differentiator in matching called party transformation patterns in situations where local calls could overlap with some country codes (for example, India country code 91 overlapping with local ten-digit dialing in Raleigh, NC, Morocco country code 212 overlapping with local ten-digit dialing in New York, NY, and so forth).

Configuring Multiple Advertising Services

Hosted DN groups are a collection of hosted DN patterns that you group together in Cisco Unified Communications Manager Administration. You assign a hosted DN group to a CCD advertising service in Unified CM Administration, and the CCD advertising service publishes all the hosted DN patterns that are a part of the hosted DN group.

You can configure multiple advertising services in Unified CM. Each advertising service establishes a unique relationship between Hosted DN ranges and the group of call processing nodes that will advertise themselves as responsible for the reception of calls to the DNs in those ranges.

An advertising service is associated with a Hosted DN group, which itself is common to a set of Hosted DN ranges. The advertising service is also associated with one SIP SAF trunk and/or one H.323 SAF trunk. Each of those trunks is associated with a device pool, which itself is associated with a Unified CM server group. The constituent members of the Unified CM server group will be advertised as responsible for calls to any of the DN ranges in the Hosted DN group. In systems where the call processing servers are deployed as a cluster across the WAN (clustering over the WAN), Cisco recommends that the Unified CM server groups used to serve the phones be used also to advertise the DN ranges corresponding to the lines configured on the phones. This will ensure that calls made by remote SAF CCD clients to those phones are sent to the Unified CM servers co-located with the phone's controlling servers.

Learning Remote DN Ranges Through the SAF CCD Requesting Service

The SAF CCD Requesting Service is used to learn the DN ranges hosted in other call agents participating in the SAF CCD service. The Requesting Service is associated with SAF CCD trunks, and it is used to select the trunks associated with the Requesting Service when calls are placed to SAF CCD learned DN ranges.

Each DN range is associated with multiple attributes, such as:

- DN Range — For example: 8555XXXX, +1408555XXXX
- ToDID info — Representing the PSTN failover information provided by the advertising call agent. For example: 1:+1408
- IP address — Representing the signaling destination of the call agent hosting the advertised DN range. This field carries the address of the trunk used by the advertising cluster to inject the DN range into the SAF CCD service. For example: 10.0.0.1.
- Protocol — Representing the signaling protocol required to contact the call agent responsible for the hosted DN range. For example: SIP, H.323

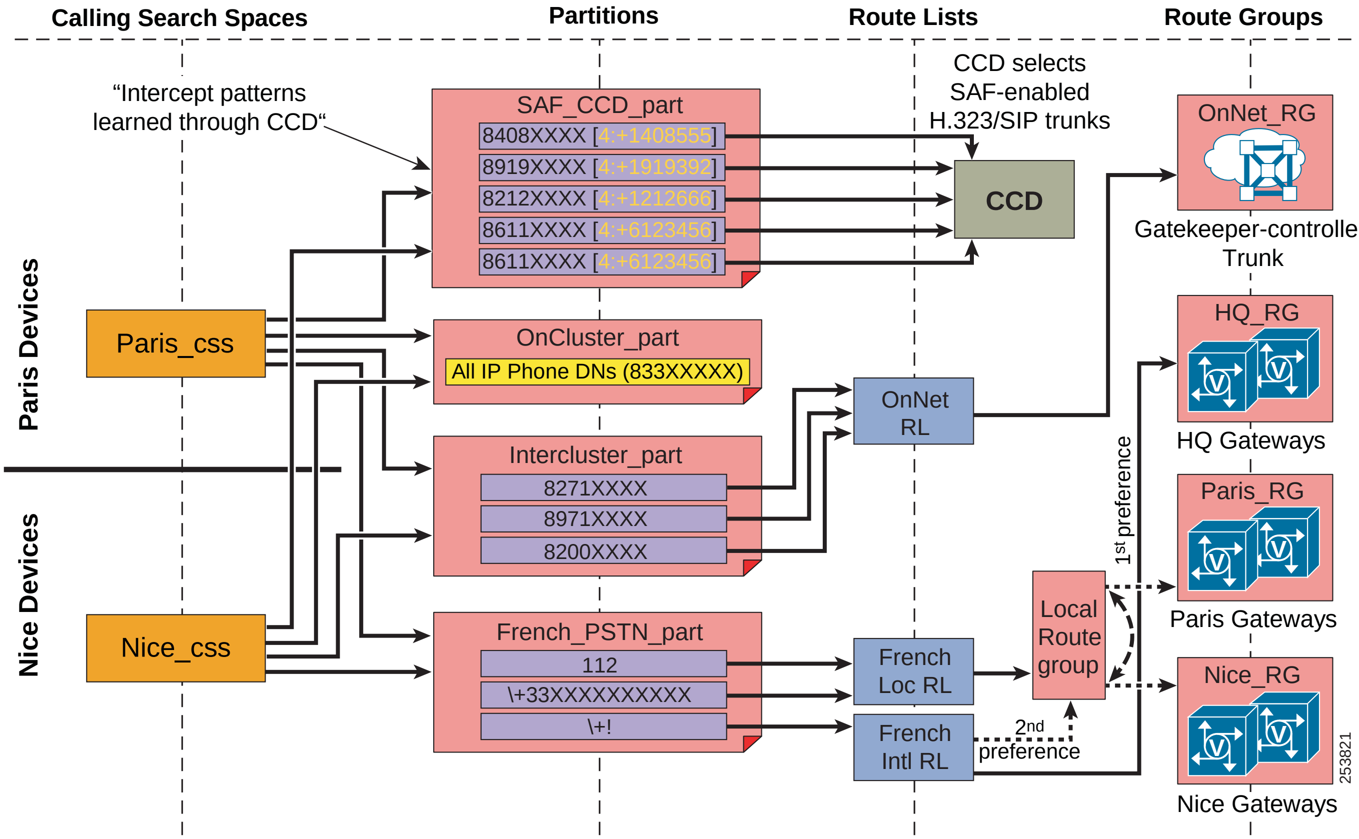
Note If an advertising service is associated with a trunk whose device pool features a Cisco Unified Communications Manager Group with more than one member, one SAF CCD record is advertised per member. This means that, if a single hosted DN range is advertised through a trunk with three Unified CM group members, three SAF CCD records are advertised. Load balancing is used by the calling cluster between all the records advertising the same hosted DN range.

The SAF CCD Partition

In a cluster, a single SAF CCD partition is configured and is used to contain all learned patterns, no matter the source of the advertised DN range, the required protocol, or the DN range form (site-code or +E.164) used in the advertisement. (See Figure 9-4.)

Note The SAF CCD partition is not listed for searches under **Call Routing** > **Class of Control** > *partition*.

Figure 9-4 Integration of SAF Call Control Discovery with Static Routing



The fact that all learned patterns are placed in the single call control discovery partition means that a phone cannot be given access to only a subset of learned patterns. Access to all the patterns is effected by inclusion of the Call Control Discovery Partition in a CSS used by a phone or in the CSS of a translation pattern used to adapt the dialed, localized form of a number into the globalized form advertised into the SAF CCD service.

For example, the Paris user in Figure 9-4 dials 84081234. None of the statically configured patterns matches the dialed string. However, the SAF_CCD_Part partition has been populated with DN ranges learned from the SAF CCD requesting service. The pattern 8408XXXX matches the dialed string directly and allows the user's call to be placed through a SAF CCD-enabled IP trunk. Note that the Paris and Nice users in this example have access to all the patterns in the SAF_CCD_part partition.

Placing Calls to SAF CCD Learned DN Ranges

In general, calls placed to SAF CCD learned DN ranges should match the range either directly, if the user dialed the advertised globalized number, or through a globalization translation pattern, if the user dialed the number in a local form. SAF CCD learned patterns are always learned as non-urgent. This might cause partial overlaps when the SAF CCD learned patterns are global +E.164 patterns and there is an alternative match on a \+! PSTN route pattern. In this case dialing the global on-net destination learned through SAF CCD either directly or by dialing the habitual local form that is transformed by appropriate translations will be subject to inter-digit timeout (T302).

The calling party number should be sent as-is if the DN of the caller is already in a global form (site-code or +E.164). If the calling party number is in a local form, it should be globalized before egressing the local cluster. This is best done through the use of calling party transformation patterns on the SAF CCD trunk used to place the call.

When a user dials a number corresponding to a DN range advertised by a remote call agent, the following events occur:

1. The dialed string is processed through the calling phone's effective CSS. The best-match process is used, as usual. This means that, if a call is placed to a destination matching several SAF CCD learned routes and/or patterns locally configured in the cluster, the most precise match will be chosen to match the call. In cases where multiple equal-precision matches are found, the order in which the associated partitions are listed in the effective CSS will be used. In the special case where multiple Unified CM nodes belonging to the same cluster advertise the same route, multiple equal-precision patterns will be found in the SAF CCD partition of all clusters participating in the SAF CCD service. In this case, the calls to this pattern are load-balanced between all the equal-precision matches.
2. Either a match is found directly on the dialed pattern (for example, the user dials 84081234 and this matches a pattern found in the SAF CCD partition as included in the phone's CSS), or a match is made with a translation pattern used to adapt the local form to the global form advertised in the SAF CCD partition (for example, the user in Memphis, TN, dials 9011441545551234, matches a translation pattern that adapts the called number to +441545551234, and then matches the +441545551234 found in the SAF CCD partition located in the translation pattern's CSS).
3. The call is extended through the IP trunk used to learn the pattern in the local cluster, to the trunk used to advertise the pattern in the remote cluster.
4. Upon egressing the calling cluster, the called number is in the form advertised by the remote cluster.
5. Upon egressing the calling cluster, the calling number should be provided in a global form. If the local cluster's DNs were not provisioned in a global form, calling party transformation patterns should be used to adapt the local form to the global form to be sent to the remote cluster. This is especially important to allow remote users to use the dial function from the missed and received calls lists. Note also that globalizing the calling party number should be done by the calling cluster to simplify configuration. The alternative requires configuring all remote clusters with the rules required to recognize calls coming from other clusters and globalize the calling party numbers.
6. If the IP route between the calling and the called cluster is available, the call will be received at the remote cluster into the trunk associated with the advertising service used to inject the DN range into the SAF CCD service.
7. The remote cluster's SAF CCD trunk receiving the call will look for a match in the trunk's Inbound Calls CSS.
8. If the DNs as configured in the cluster are in the same global form as that advertised into the SAF CCD service, a match is found by including the DN partitions into the SAF CCD trunk's Inbound Calls CSS. The call is offered to the called line.
9. If the DNs as configured in the cluster are in a form different than the global form advertised into the SAF CCD service, a match is found by including, in the SAF CCD trunk's Inbound Calls CSS, a partition containing translation patterns matching the global form to the local form configured on the DNs of the lines. The call is offered to the called line.
10. In step 6., if the IP route between the two clusters is not available, the PSTN failover digit transformation rules (ToDID rules) are applied to the called number, and the resulting destination number will be used to find a match in the calling device's AAR CSS.
11. At this point, the called number should be in +E.164 form and should be used to match a route pattern pointing to a route list, route group, and gateway (or trunk) combination.

12. Once the call egresses on a gateway or trunk, transformation patterns should be used to adapt both the calling and called party numbers to the form required by the PSTN carrier. At this point, the call is launched into the PSTN.
13. Once the call reaches the remote cluster, the call is processed as usual for incoming PSTN calls.

Note If the design intent were to route all calls to SAF CCD learned routes through the PSTN, such as is the case when deploying the VoPSTN approach, simply put the SAF requesting service's associated trunk(s) in a call admission control static location configured with 1 kbps of bandwidth. This will force all calls to be routed through the calling device's AAR CSS.

Dial Plan Considerations for the Intercompany Media Engine

The Cisco Intercompany Media Engine (IME) allows participating enterprises to route calls over the Internet between them. When a phone participates in the IME and places calls through trunks or gateways marked as **PSTN Access**, a record of the call is sent to the enterprise's IME server, including the calling and called numbers. This record serves to flag the pairing of numbers in two different IME-participating companies for future call routing over the Internet; if another call between these same two numbers is detected by the IME server, it instructs Cisco Unified Communications Manager (Unified CM) to route the call over the Internet.

One challenge in such pairings is ensuring the consistent normalization of numbers in all the participating IME-enabled Unified CM clusters. Because the calls are initially placed over the PSTN and because these calls can traverse the boundaries of different cities, provinces or states, or even countries, the form of the number dialed by the user will vary greatly. Similarly, different companies might have adopted different approaches when assigning DNs to lines.

Cisco recommends using the +E.164 form to identify the calling and the called numbers for calls participating in the IME service. The +E.164 form affords the normalization (for example, every number begins with + followed by the country code of the associated DID number) and the globalization required to ensure consistent results.

When calls are placed toward trunks or gateways marked as PSTN Access, the trunk or gateway's associated IME E.164 transformation profile is applied. This in turn applies a set of transformation patterns to the calling party number, as well as a set of transformation patterns to the called party number. For outgoing calls, the post-transformation numbers are used in the records sent to the IME server.

Note The IME E.164 Transformation Profile is configured in Unified CM under **Advanced Features** > **Intercompany Media Services** > **E.164 Transformation**.

The Intercompany Media Services E.164 Transformation configuration allows the application of transformation patterns to outgoing calls, segregated between calling party and called party. For each, a CSS can be configured, which itself contains the partitions in which the calling/called party transformation patterns are contained. The transformations are applied to either the original number (the form the number was in as it matched the route pattern) or the routing transformed number (the form the number was in once route list transformations were applied).

For the calling party number:

The original number is the DN of the phone as it matched a route pattern, when considering the calling party settings. The routing transformed number is the DN of the phone after any calling party digit manipulations are performed through route lists.

For example, a DN configured as 85551234 is placing a call to the PSTN, to 91415551000. The call is placed through a translation pattern 91[2-9]XX[2-9]XXXXXX. The translation pattern is configured to modify the called party number to +14155551000, while changing the calling party number to +14085551234. This call then matches a route pattern \+!, configured with a route list that modifies the calling party number to 408 555 1234 and the called party number to 415 555 1000.

If the Outgoing Calling Party Settings are set to apply the transformations to the original number, then +14085551234 will be used to match a calling party transformation pattern in the Outgoing Party E.164 Transformation CSS.

If the Outgoing Calling Party Settings are set to apply the transformations to the routing transformed number, then 4085551234 will be used to match a calling party transformation pattern in the Outgoing Party E.164 Transformation CSS.

For the called party number:

The original number is the dialed number as it matches the route pattern, when considering the called party settings. In the example above, the original number is +14155551000.

The routing transformed number is the destination of the phone after any digit manipulations are performed through route lists. In the example above, it is 4155551000.

No matter to which form of the number the transformations are applied, they must yield a normalized, globalized number to be sent to the IME service.

When calls are received from trunks and gateways marked as PSTN Access, the form in which the calling and called numbers are received must be normalized and globalized before they are sent to the IME service. The Incoming Transformation Profile Settings can be used to adapt the incoming form of the calling and called numbers to the +E.164 form required by the IME service. It features an Incoming Calling Party Transformation Profile as well as an Incoming Called Party Transformation Profile. For each, a CSS can be configured, which itself contains the partitions in which the calling/called party transformation patterns are contained.

The transformations for the calling number are applied to the routing transformed number; that is, the number as processed through the Incoming Calling Party Settings at the trunk or gateway level.

The transformations for the called number are applied to the number as it is presented into the gateway or trunk's CSS for inbound calls.

Design Guidelines for Multisite Deployments

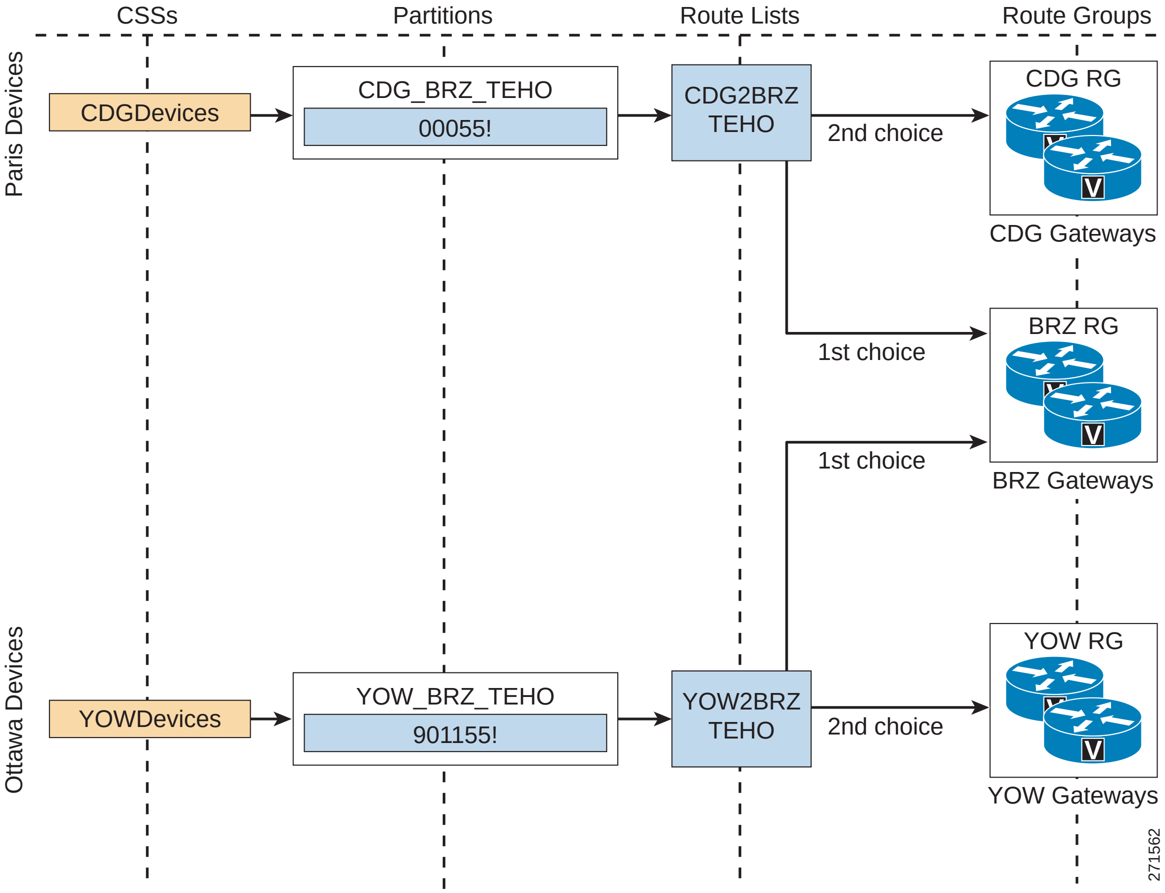
The following guidelines and best practices apply in common to all multisite IP Telephony deployments. For deployments that involve more than one Unified CM cluster, also refer to the section on Additional Considerations for Multi-Cluster Systems, page 9-37.

- To prevent routing loops, make sure the calling search spaces of all PSTN gateways do not include any partitions that contain external route patterns assigned to route lists and route groups that can send calls out the same gateway.
- When choosing DID ranges with your Local Exchange Carrier (LEC), try to select them so that no overlap occurs within a site. For example, if you use four-digit dialing within a site and you need two blocks of 1000 DIDs, the blocks (408)555-1XXX and (408)444-1XXX would overlap when reduced to four-digit numbers and would introduce additional complexity due to inbound and outbound translations.
- Allow for multiple ways of dialing emergency numbers. For example, in North America, configure both 911 and 9.911 as emergency route patterns within Unified CM.
- When automated alternate routing (AAR) is deployed, ensure that the external phone number mask or AAR Destination Mask configured on the IP phones is compatible with all the prefixes added by the various AAR groups. For example, in a multi-national deployment, do not include national access codes such as 0 in the mask unless they are part of the global E.164 address. The simplest way to configure AAR relies on the configuration of the AAR destination mask as the full E.164 address of the phone, including the + sign.
- You can force calls to on-net destinations, but dialed as PSTN calls, to be routed on-net within the cluster by adding translation patterns that match the E.164 DID ranges for each site and that manipulate the digits so they match the destination's internal extension. For example, if a DN is reachable on-net by dialing 1234, but someone within the system dials this same destination as 9 1 415 555 1234, you can force the call to be kept on-net by creating a translation pattern 9 1 415 555.1XXX, which removes all digits pre-dot and routes the call on-net to the resulting number. However, remember to configure the AAR calling search space to exclude the partition containing the "force on-net" translation patterns but to include a partition containing the regular route patterns pointing to the PSTN, so that automated PSTN failover is possible when the IP WAN is out of bandwidth.
- Within a centralized call processing cluster with N sites, you can implement Tail End Hop Off (TEHO) using one of the following methods:
  - TEHO with centralized failover

    This method involves configuring a set of N route patterns in a global partition, with each pattern pointing to a route list that has the appropriate remote site route group as the first choice and the central site route group as the second choice.
  - TEHO with local failover

    This method involves configuring N sets of N route patterns in site-specific partitions, with each pattern pointing to a route list that has the appropriate remote site route group as the first choice and the local site route group as the second choice. For the example in Figure 9-5, in order to implement local failover TEHO routes to Brazil, a site in Paris, France would require a dedicated route pattern and route list to route the calls to the TEHO gateways in Brazil as a first choice or to the Paris gateways as a second choice. Because the pattern is linked to a site-specific route list, it cannot be reused at any other site. Likewise, the site in Ottawa, Canada requires its own dedicated route pattern pointing to an Ottawa-specific route list to allow local failover to a gateway in Ottawa.

Figure 9-5 TEHO Without Local Route Group

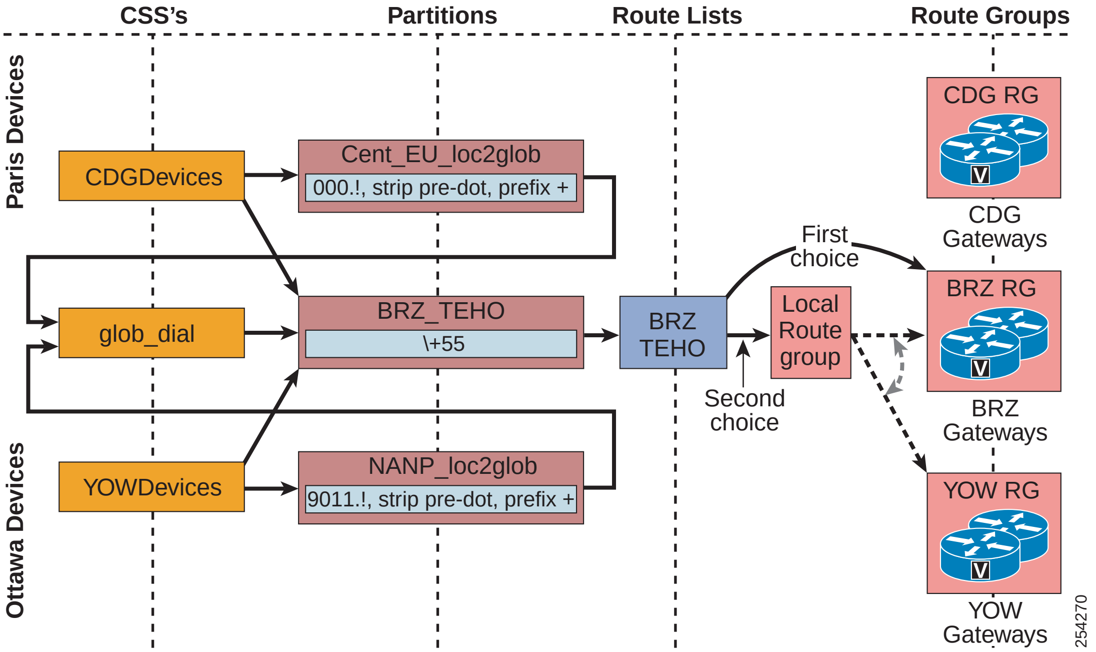


While this second approach allows for an optimal failover scenario when the remote gateway or the IP WAN is unavailable, it also introduces a high level of complexity into the dial plan because it requires a minimum of N^2 route patterns and N^2 route lists, as opposed to the N route patterns and N route lists needed with the first approach.

– TEHO with local failover with Local Route Group

The Local Route Group allows for the local failover of TEHO routes to be implemented without having to create route patterns for each site. For the example in Figure 9-6, a single TEHO pattern and route list is used by both the Paris and Ottawa sites. Because the user input for these two sites is not the same (French users dial Brazilian destinations differently that Canadian users do), the configuration relies on translation patterns to globalize the user input. The global form is then used to match a single, cluster-wide route pattern pointing to a route list whose first entry is the Brazil route group and whose second entry is the Standard Local Route Group. The local route group is resolved to the Paris route group when the calling device is in a Paris device pool, and to an Ottawa route group when the calling device is in an Ottawa device pool.

Figure 9-6 TEHO With Local Route Group



- When appropriate for your national numbering plan, you may configure an additional translation pattern at each site to catch local PSTN calls dialed as long-distance calls and to translate them into the proper abbreviated form. This translation pattern should be accessible only from phones located within the site. Such a configuration also helps simplify the AAR configuration (see Special Considerations for Sites Located Within the Same Local Dialing Area, page 9-121).
- Do not use the multilevel precedence and preemption (MLPP) feature to assign higher priority to emergency calls. An emergency-related call might not appear as such to the IP Telephony system, and you would risk terminating an existing emergency call to place another call to the main emergency service routing number. For example, an emergency situation might prompt someone to place a call to a regular ten-digit number to reach a medical professional. Preemption of this call would abort the ongoing emergency communication and could delay handling of the emergency. Also, incoming calls from emergency service personnel would be at risk of preemption by MLPP.

Note A Unified CM cluster with a very large dial plan containing many gateways, route patterns, translation patterns, and partitions can take an extended amount of time to initialize when the Cisco CallManager Service is first started. If the system does not initialize within the default time, there are service parameters that you can increase to allow additional time for the configuration to initialize. For details on the service parameters, refer to the online help for Service Parameters in Unified CM Administration.

Additional Considerations for Multi-Cluster Systems

In addition to the considerations made in the previous section, observe the following best practices when designing a dial plan for a multisite deployment involving multiple Unified CM clusters:

- Avoid splitting DID ranges across multiple Unified CM clusters. This practice would make intercluster routing very difficult because summarization would not be possible. Each DID range should belong to a single Unified CM cluster.

- Avoid splitting devices within a remote site between multiple Unified CM clusters using call admission control based on static locations. Static locations-based call admission control is significant only within a cluster, and having devices belong to different clusters at the same remote site would lead to poor utilization of the IP WAN bandwidth because you would have to split the available bandwidth between the clusters. Each remote site should belong to a single Unified CM cluster. Locations can be configured in Unified CM to use RSVP as the call admission control mechanism, which allows the efficient sharing of a single site's total WAN bandwidth between phones belonging to different clusters. To take full advantage of the efficiency of RSVP-based call admission control, all phones within the remote site must be configured to use RSVP.
- Use gatekeeper-controlled intercluster trunks to route calls between Unified CM clusters. This practice enables you to add or modify clusters easily in your network without reconfiguring all other clusters.
- Implement redundancy in the connection between Unified CM and the gatekeeper by using a gatekeeper cluster and by assigning the intercluster trunk to a device pool that uses a Unified CM group with multiple servers configured.
- When sending calls to the gatekeeper, expand the called number to the full E.164 address. This practice simplifies PSTN failover when the IP WAN is not available because no additional digit manipulation is required to reroute the call via a PSTN gateway. Also, this practice eliminates the need to configure the local (calling) Unified CM with dial length information for each remote site.
- Within the gatekeeper, configure one zone per Unified CM cluster. For each cluster/zone, add zone prefix statements to match all DN ranges owned by that cluster.
- You can implement Tail-End Hop-Off (TEHO) across multiple Unified CM clusters by following these guidelines:
  - Add specific route patterns for the relevant E.164 ranges to the source (originating) Unified CM cluster, and point them to a route list that has the IP WAN route group as the first choice and the Standard Local Route Group as the second choice
  - Within the Cisco IOS gatekeeper configuration, add zone prefix statements for all the relevant E.164 ranges and point them to the appropriate Unified CM cluster
  - Ensure that the intercluster trunk calling search space in the destination Unified CM cluster includes partitions featuring route patterns that match the local PSTN numbers, and that digit manipulation is applied as needed (for example, stripping the area code before sending the call to the PSTN) by using appropriate Called Party Number Transformation Patterns.

For details on how to configure a Cisco IOS gatekeeper for distributed call processing deployments, see Gatekeeper Design Considerations, page 8-37.

Choosing a Dial Plan Approach

As introduced in Planning Considerations, page 9-4, there are two main approaches to a dial plan for internal destinations within an IP Telephony system:

- Uniform on-net dial plan, where each internal destination is dialed in the same way regardless of whether the caller is located in the same site or in a different site.
- Variable-length on-net dial plan, where internal destinations are dialed differently within a site than across sites. Typically, this approach uses four or five-digit abbreviated dialing for calls within a site and either full E.164 addresses or an on-net access code followed by a site code and the extension for calls across sites.

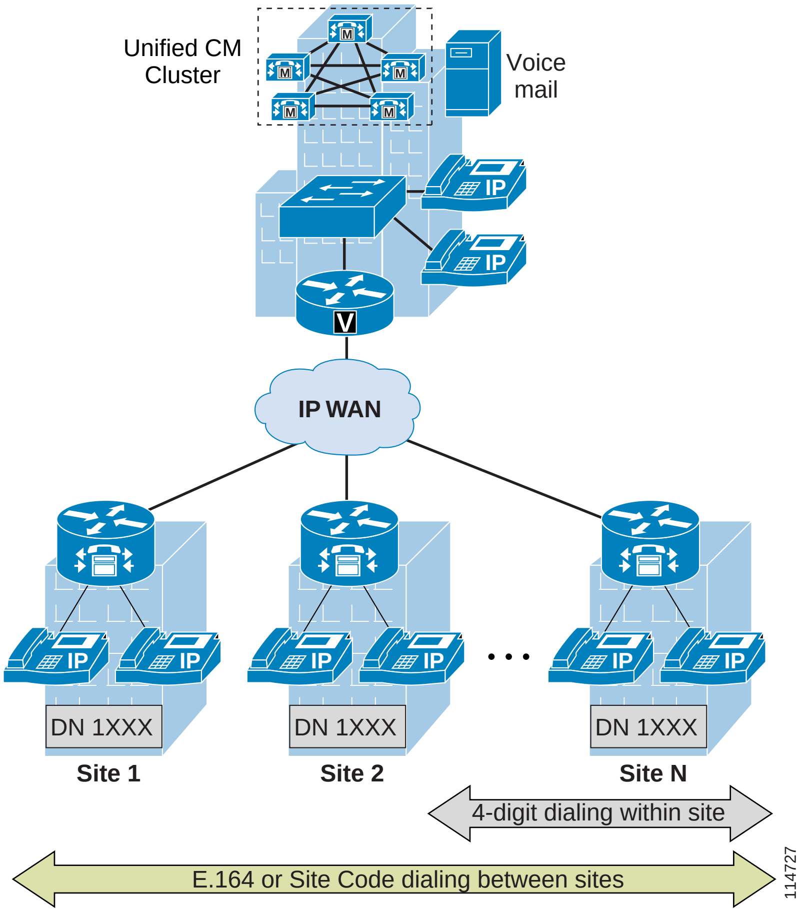
To help you decide which approach is best suited for your needs, consider the following high-level design questions:

- How many sites will eventually be served by the IP Telephony system?
- What are the calling patterns between sites or branches?
- What do users dial within a site and to reach another site?
- Are there any calling restrictions applied to on-net inter-site calls?
- What transport network (PSTN or IP WAN) will be used for most inter-site calls?
- What (if any) CTI applications are being used?
- Is there a desire for a standardized on-net dialing structure using site codes?

Uniform on-net dial plans are the easiest to design and configure; however, they work best for small to medium deployments, and they become impractical when the number of sites and users increases. They are described and analyzed in detail in the section on Deploying Uniform On-Net Dial Plans, page 9-40.

Variable-length on-net dial plans are more scalable but also more complex to design and configure. Figure 9-7 shows the typical requirements for a large-scale deployment using the variable-length on-net dial plan approach.

Figure 9-7 Typical Dialing Requirements for Large Multisite Deployments



With Unified CM, the main method for implementing a variable-length on-net dial plan relies on flat addressing. In this method, all internal extensions are placed in the same partition. This method is typically based on on-net site codes for inter-site calls and is analyzed in detail in the section on

Deploying Variable-Length On-Net Dial Plans with Flat Addressing, page 9-43. In some cases it is possible to use this approach even when using full E.164 addresses for inter-site calls, as described in the section on Special Considerations for Deployments Without Site Codes, page 9-50.

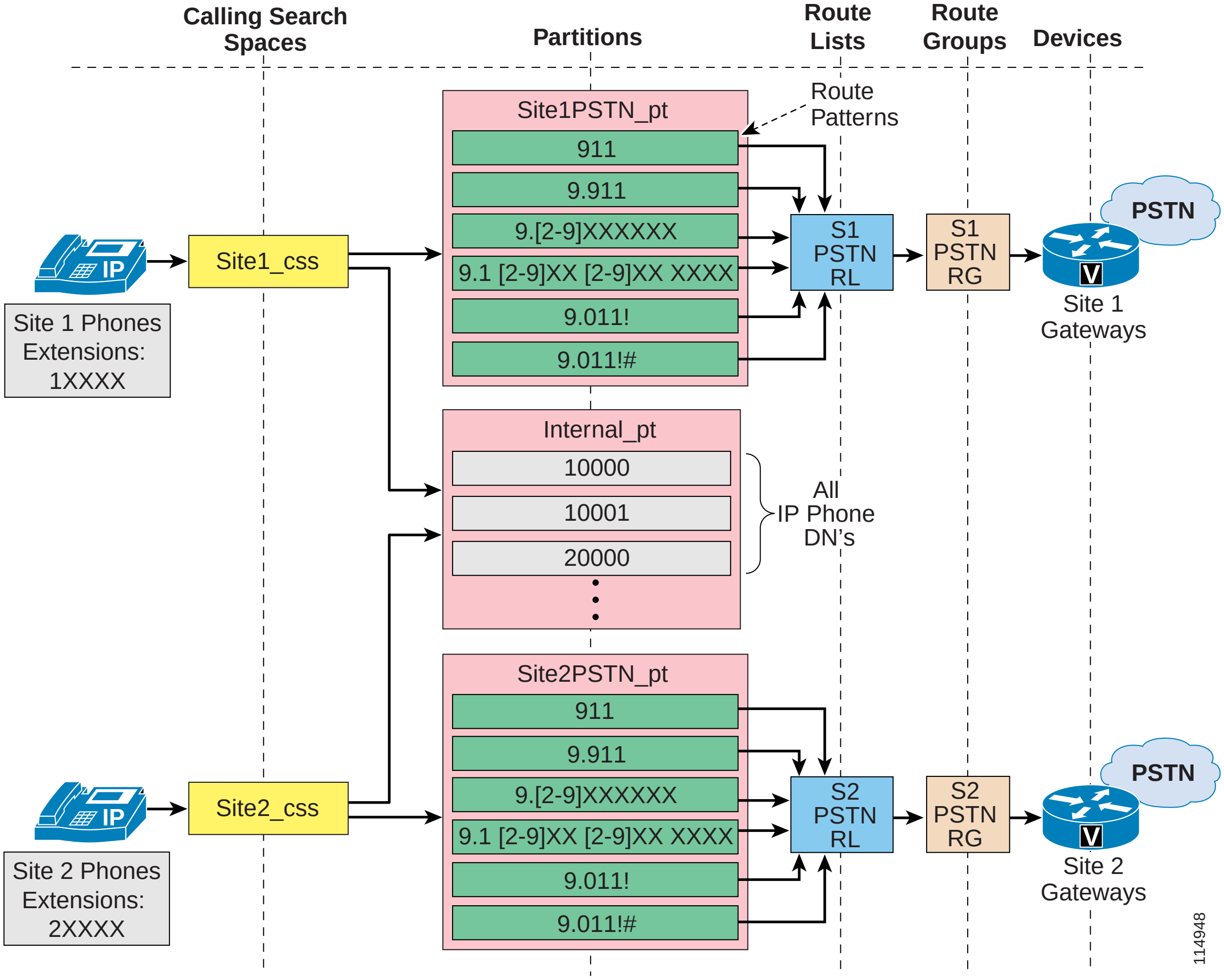
Deploying Uniform On-Net Dial Plans

You can implement a uniform on-net dial plan by following these guidelines:

- Uniquely identify all phones with an abbreviated extension.
- Place all the phone DNs in a single partition.
- At each site, place PSTN route patterns in one or more site-specific partitions, according to the chosen class-of-service approach.

Figure 9-8 shows an example implementation for a single Unified CM cluster deployment.

Figure 9-8 Example of Uniform On-Net Dial Plan Deployment



Use this approach if both of the following conditions apply:

- The DID ranges available do not overlap when considering the number of digits chosen to identify internal extensions.
- The number of sites covered by the IP Telephony system is not expected to grow significantly over time.

The following sections analyze implementation details and best practices for different types of calls within the framework of uniform on-net dial plans:

- Inter-Site Calls Within a Cluster, page 9-42
- Outgoing PSTN and IP WAN Calls, page 9-42
- Incoming Calls, page 9-42
- Voicemail Calls, page 9-42

Inter-Site Calls Within a Cluster

Because all internal DNs are directly reachable from every device's calling search space, all on-net calls (intra-site and inter-site) are automatically enabled and need no additional configuration within Unified CM.

Outgoing PSTN and IP WAN Calls

PSTN calls are enabled via the site-specific partitions and route patterns, so that emergency and local calls can be routed via the local branch gateway. Long-distance and international calls may be routed via the same branch gateway (as shown in Figure 9-8) or via a centralized gateway, depending on company policy. This second option would simply require an additional route list per site, with the first-choice route group pointing to the central site gateway and, optionally, a second-choice route group pointing to the local branch gateway. To allow the reuse of route patterns between sites for PSTN calls while still allowing site-specific routing of the calls, route lists referencing the Standard Local Route Group can be used.

Abbreviated dialing to another Unified CM cluster or Cisco Unified Communications Manager Express (Unified CME) is also possible via a gatekeeper. For these IP WAN calls, Cisco recommends that you expand the abbreviated string to the full E.164 via a translation pattern before sending it to the gatekeeper.

Emergency Calls

If Cisco Emergency Responder is used for managing emergency calls, the partition containing the CTI route point used to send calls to Cisco Emergency Responder should be part of the calling search space of all phones in all branches instead of the site-specific 911 patterns as illustrated in Figure 9-8. Cisco Emergency Responder will be able to identify the calling phone because there is no duplicity of DNs allowed in the internal partition. For more information on Cisco Emergency Responder considerations, refer to the chapter on Emergency Services, page 10-1, and to the Cisco Emergency Responder product documentation available at

http://www.cisco.com

Incoming Calls

Incoming PSTN calls simply require stripping the excess digits in order to match the extension length configured in Unified CM. This can be done within the gateway configuration or, alternatively, via a translation pattern included in the gateway's calling search space.

Voicemail Calls

Because every extension is unique within the system, the extension itself can be used to configure voicemail boxes within the voicemail system. No translations are necessary to send calls to the voicemail system or to enable a Message Waiting Indicator (MWI) within Unified CM.

However, when users access the voicemail system from the PSTN, they need to be trained to enter their abbreviated extension to access their voicemail boxes.

Deploying Variable-Length On-Net Dial Plans with Flat Addressing

You can implement a variable-length on-net dial plan with flat addressing by defining phone DNs as unique strings containing an on-net access code, a site code, and the extension (for example, 8-123-1000). You can place all these DNs in the same global partition, thus enabling inter-site calls using the site code, and you can define translation patterns in site-specific partitions (one translation pattern and one partition per site) to enable abbreviated dialing within a site.

This internal structure can be hidden from the end users by configuring the Line Text Label parameter within the Directory Number configuration page with the four or five-digit number that users are accustomed to dialing within the site. The external phone number mask should also be provisioned with the corresponding PSTN number in order to enable AAR and to give users a visual indication of their own DID number on the IP phone display.

Table 9-4 illustrates the basic relationship between calling search spaces and partitions at each site, without taking into account additional elements required to implement classes of service:

Table 9-4 Calling Search Spaces and Partitions for Variable-Length Dial Plans with Flat Addressing

Calling Search Space	Partition	Partition Contents
Site1_css	Site1Translations_pt	Translation patterns for Site 1's abbreviated dialing
	Site1PSTN_pt	PSTN route patterns for Site 1 (more partitions may be needed based on classes of service)
	Internal_pt	All IP phone DNs (unique form)
…	…	…
SiteN_css	SiteNTranslations_pt	Translation patterns for Site N's abbreviated dialing
	SiteNPSTN_pt	PSTN route patterns for Site N (more partitions may be needed based on classes of service)
	Internal_pt	All IP phone DNs (unique form)

Use this approach if one or more of the following conditions apply:

- No dialing restrictions are needed for on-net inter-site calls.
- A global on-net numbering plan using site codes is desired.
- Inter-site calls are normally routed over the IP WAN.
- CTI-based applications, such as Cisco Emergency Responder, are used across sites.

Note If dialing restrictions need to be applied to on-net inter-site calls, or an on-net numbering plan using site codes is not desired, refer to the section on Special Considerations for Deployments Without Site Codes, page 9-50, for a variant of this approach that can accommodate these needs.

The following considerations apply to this approach:

- The destination numbers of intra-site four-digit calls get expanded to the unique internal DN on the IP phone display.
- The Placed Calls directory will display the original four-digit string as dialed by the user.

- Calling number, and numbers in the Missed Calls and Received Calls directories, appear as the unique internal DN.
- To preserve the four-digit dialing feature when the IP WAN is unavailable and the branch phones are in SRST mode, you need to apply a translation rule to the **call-manager-fallback** configuration within the SRST router.
- When the branch phones are in SRST mode, the Line Text Label that masks the unique internal DN as a four-digit number on the IP phone display is not available, so the users will see their full internal DN appear instead.

To better understand how to deploy the flat addressing approach, consider again the hypothetical customer network shown in Figure 9-9. In this case, it has been decided that a variable-length on-net dial plan is required, with four-digit dialing within each site (the 1XXX extension range is chosen at each site) and inter-site dialing with an eight-digit string consisting of an on-net access code (8 in this example), a three-digit site code, and the four-digit extension. The three-digit site code is derived from the NANP area code for the sites located in the United States, and from the E.164 country code followed by a site identifier for the sites located in Europe. Table 9-5 summarizes the site codes chosen.

Table 9-5 Site Codes for the Customer Network in Figure 9-9

	San Jose	New York	Dallas	London	Paris	Milan
Site Code	408	212	972	442	331	392

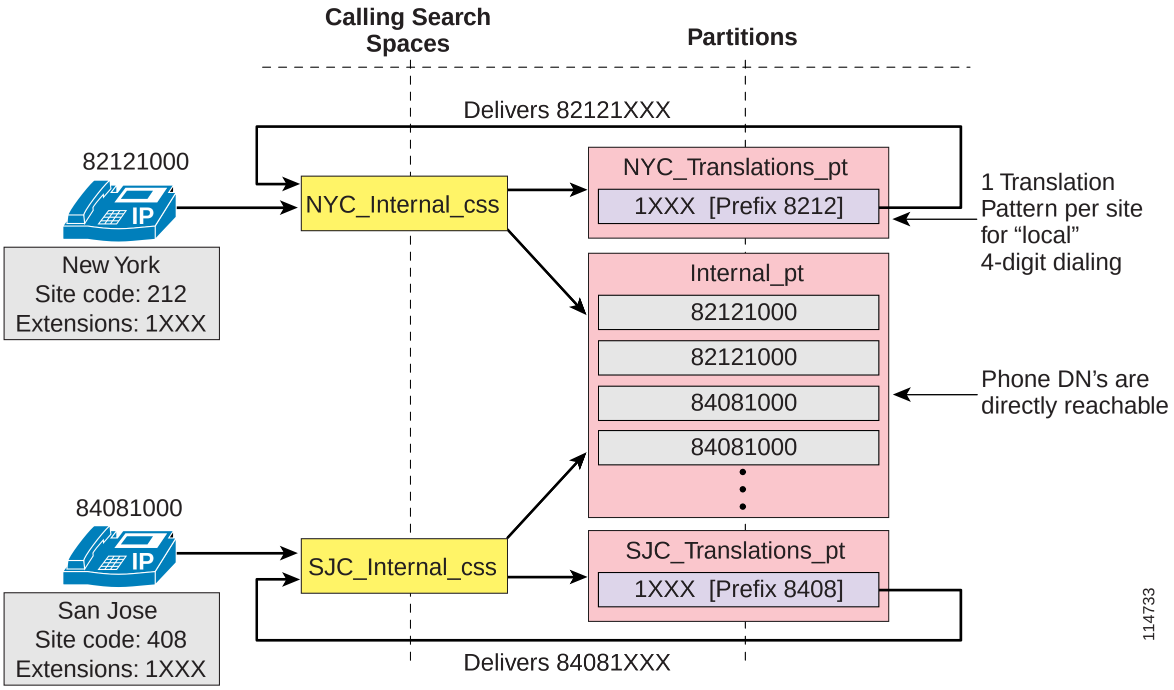
Using the US cluster from this example, the following sections analyze implementation details and best practices for different types of calls within the framework of the flat addressing approach:

- Inter-Site Calls Within a Cluster, page 9-45
- Outgoing PSTN and IP WAN Calls, page 9-46
- Incoming Calls, page 9-49
- Voicemail Calls, page 9-49
- Special Considerations for Deployments Without Site Codes, page 9-50

Inter-Site Calls Within a Cluster

Figure 9-9 shows an example configuration for inter-site calls within the US cluster.

Figure 9-9 Inter-Site Calls Within a Cluster for the Flat Addressing Method



To provide connectivity between sites and partitions, use the following guidelines:

- Place all unique DNs, including the on-net access code 8, in a global partition (named Internal_pt in this example).
- Create one partition per site, each containing a translation pattern that expands four-digit numbers into the fully qualified eight-digit number for that site, thus enabling abbreviated dialing within the site.
- For each site, include both the Internal_pt partition and the local translation partition in the phone's calling search space.

The inclusion of the on-net access code in the DN configured in Unified CM enables you to place all internal extensions in a partition directly accessible by all phones, and at the same time ensures that all call directories on the IP phones are populated with numbers that can be directly redialed.

Note You must ensure that the on-net access code and site code combinations do not overlap with the local abbreviated dialing range at any site.

Outgoing PSTN and IP WAN Calls

Depending on how the various types of PSTN calls need to be routed (centralized gateways versus distributed gateways), the configuration may vary.

To provide on-net connectivity for inter-site calls to the Europe (EU) cluster, the following options are possible:

Option 1: Eight-Digit Numbers Only

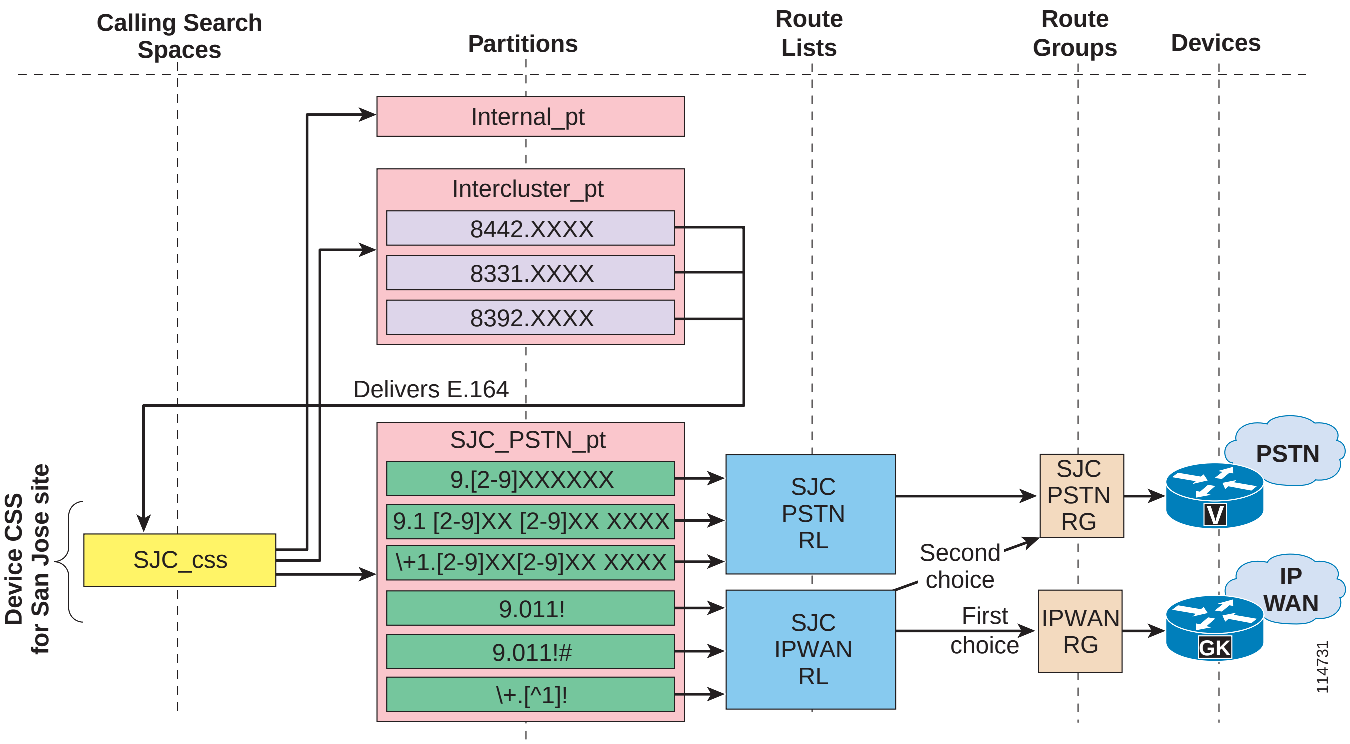
This option relies on a single route pattern that matches all eight-digit ranges (8XXXXXXX) and points to a route list or route group that contains only a gatekeeper-controlled intercluster trunk. The gatekeeper is configured to use the site codes as zone prefixes.

This solution is simple and easy to maintain because no information is needed about other clusters' site codes or E.164 ranges. However, no automatic PSTN failover is provided when the IP WAN is unavailable, and users are expected to redial manually using the PSTN access code and the E.164 address of the destination.

Option 2: Eight-Digit Numbers and E.164 Addresses with Centralized PSTN Failover

This option, illustrated in Figure 9-10, uses a global set of translation patterns that match the Europe eight-digit ranges and translate them into the corresponding E.164 numbers. The translation patterns use the central site's calling search space (in this case, San Jose), and the call then matches the international PSTN route pattern within the central site's PSTN partition. At each site, the international PSTN route pattern points to a route list with the IP WAN route group as a first choice and the local PSTN route group as a second choice. The gatekeeper is configured to use the E.164 ranges as zone prefixes.

Figure 9-10 Outgoing PSTN and IP WAN Calls for the Flat Addressing Method with Centralized PSTN Failover for IP WAN Calls



Note The configuration example in Figure 9-10 assumes that the line/device approach to building classes of service is being used, but the same considerations apply when using the traditional approach.

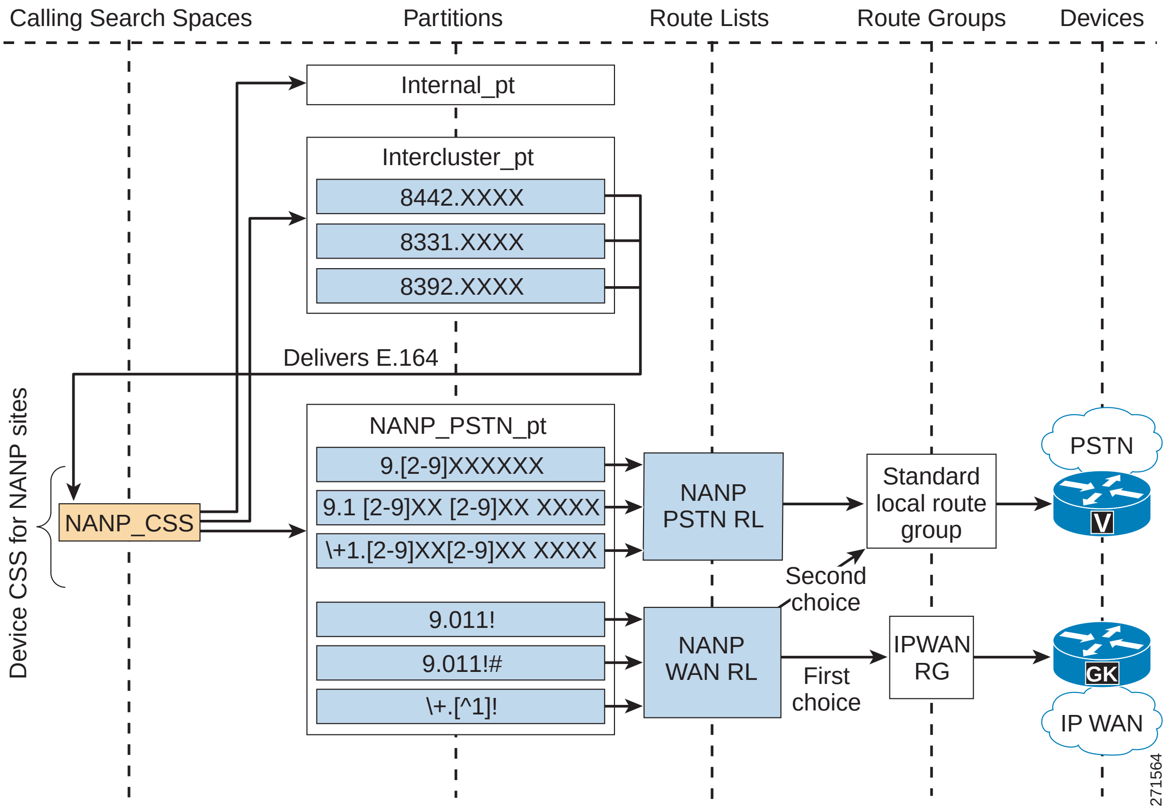
This solution requires a little more configuration and maintenance than that outlined in Option 1 because it requires that you configure and maintain information about other clusters' site codes and E.164 ranges. On the other hand, it provides automatic PSTN failover when the IP WAN is unavailable. PSTN failover is provided using the central site's gateway only, so the IP WAN bandwidth usage is not optimal.

Also observe that calls to the Europe sites dialed as PSTN calls will automatically be placed on-net if the IP WAN is available, with an automatic PSTN failover that in this case uses the local gateway.

Option 3: Eight-Digit Numbers and E.164 Addresses with Distributed PSTN Failover

This option, illustrated in Figure 9-11, uses a global set of translation patterns matching the Europe eight-digit ranges and translating them into the corresponding E.164 numbers. The translation patterns use a global calling search space (used by all sites within the North American Numbering Plan), and the call then matches the international PSTN route pattern within the NANP's PSTN partition. The international PSTN route patterns point to a route list with the IP WAN route group as a first choice and the Standard Local Route Group as a second choice. The gatekeeper is configured to use the E.164 ranges as zone prefixes.

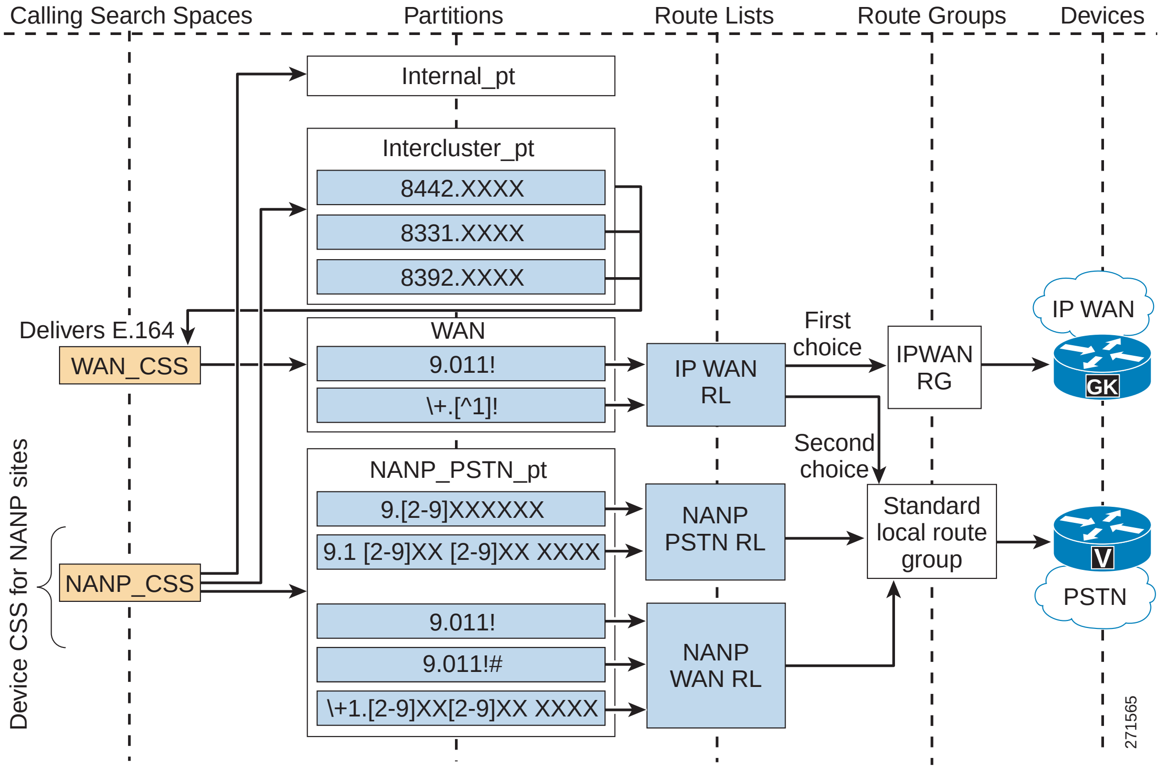
Figure 9-11 Outgoing PSTN and IP WAN Calls for the Flat Addressing Method with Distributed PSTN Failover for IP WAN Calls



This solution provides automatic PSTN failover when the IP WAN is unavailable, using the local site's gateway so that the IP WAN bandwidth usage is optimal. Because of the advent of the Local Route Group construct, this approach practically supersedes Option 2, as it requires the same level of configuration but provides local PSTN failover.

Also in this case, as for Option 2, calls to the Europe sites dialed as PSTN calls will automatically be placed on-net if the IP WAN is available, with an automatic PSTN failover using the local gateway. This in effect offers a form of TEHO functionality to all off-net European calls originating in the NANP sites. If only the calls dialed as on-net destinations are to be sent to the IP WAN, the approach can be modified to send calls to the IP WAN only if they were originally dialed as on-net inter-cluster destinations. Figure 9-12 illustrates this approach.

Figure 9-12 ***IP WAN Access for Inter-Cluster Calls Only***



Incoming Calls

Incoming PSTN calls require that the E.164 number be manipulated to obtain the eight-digit internal number in order to reach the destination phone. You can implement this requirement in any of the following ways:

- Configure the Num Digits and Prefix Digits fields within the Gateway Configuration page in Unified CM to strip and then prefix the needed digits.
- If you have configured translation patterns to force on-net inter-site calls within the cluster, you can reuse these patterns by simply prefixing the PSTN access code to the incoming called number on the gateway.
- If you are using an H.323 gateway, you can use translation rules within the gateway to manipulate the digits before sending the call to Unified CM.

The third approach has the advantage that the translation rules you configured can be reused to provide incoming PSTN connectivity to the IP phones when the branch is in SRST mode.

Voicemail Calls

Because every eight-digit extension is unique within the system, the extension itself can be used to configure voicemail boxes within the voicemail system. No translations are necessary to send calls to the voicemail system or to enable Message Waiting Indicators (MWIs) within Unified CM. Note that users are required to use their eight-digit on-net number when prompted for the mailbox number.

Special Considerations for Deployments Without Site Codes

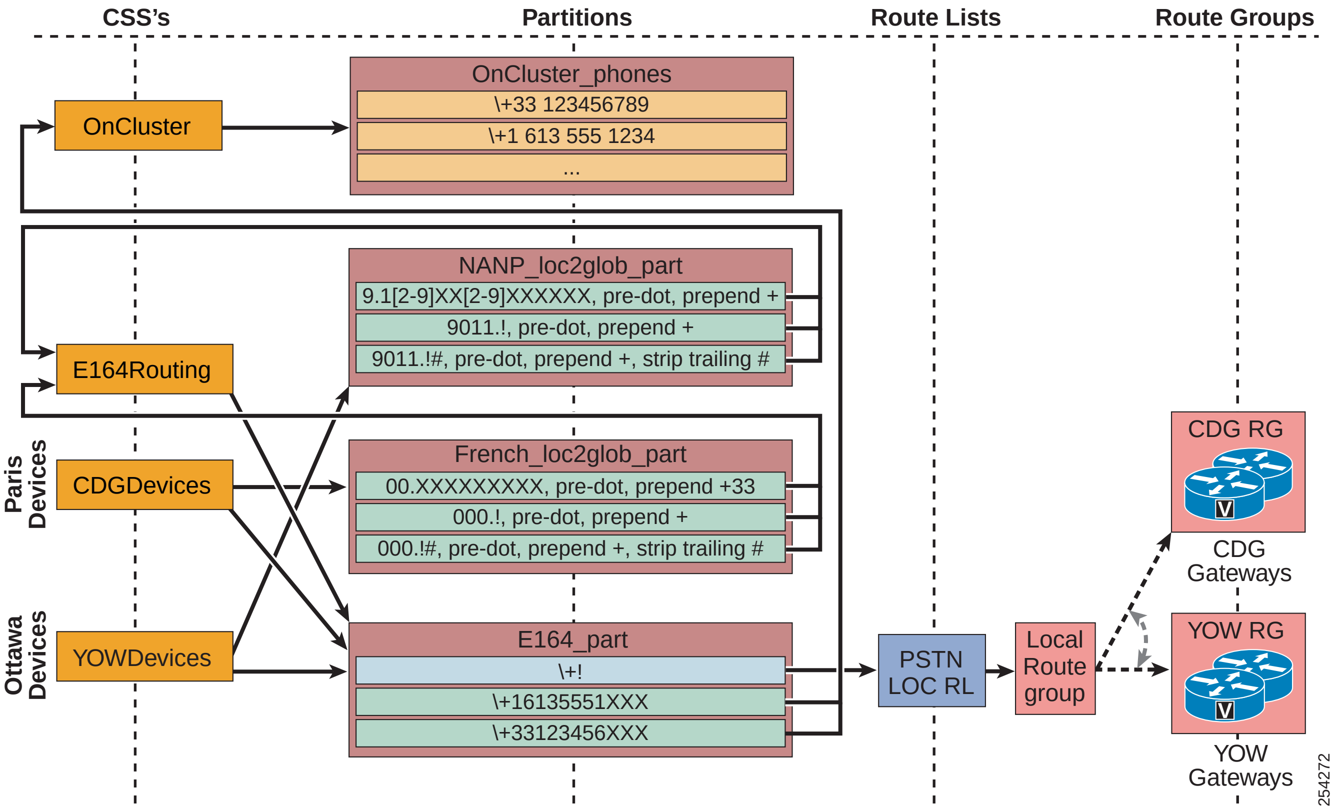
This scenario is a variant of the flat addressing approach that does not rely on the definition of an on-net numbering plan based on site codes. In this scenario, intra-site calls are still dialed as four-digit numbers, while inter-site calls are dialed as regular PSTN calls and are then intercepted and routed across the IP WAN by Unified CM.

To implement this mechanism, follow these guidelines, illustrated in Figure 9-13:

- Define the phone DNs as the full E.164 addresses and place them all in the same partition (named OnCluster_phones in this example).
- Configure translation patterns to accept localized user input and globalize it so that a full E.164 number is obtained. The resulting globalized numbers are routed through the CSS E164Routing. In this example, only two device calling search spaces are required: one accepts localized user input from the Paris site, but could be reused across all French sites. The other accepts user input from the Ottawa site, but could be reused across all NANP sites.
- Configure an E.164 routing partition (E164_part in this example). Create the appropriate set of route patterns and route lists to route PSTN calls. In this example, we rely on a single, cluster-wide route pattern \+!, which is able to route all globalized destination PSTN calls to the local route group. In addition, create translation patterns that match the existing on-cluster E.164 prefixes and route calls back to the OnCluster_phones partition.

As a user in Paris dials a number, it is globalized through the translation patterns located in the French_loc2glob_part partition, and the resulting number is then routed through the E164Routing CSS. If the destination number is an on-cluster DN, it will simultaneously match the generic pattern \+! and the urgent translation pattern with the respective specific site prefix in the E164_part partition. The more specific site prefix will be selected, and the call will be extended to the on-cluster DN by means of the OnCluster calling search space. This two-step routing is required to avoid the T.302 interdigit timeout when dialing on-net destinations. If the dialed destination is not a phone on the cluster, the globalized number routed through the E164Routing CSS will match only the \+! pattern in the E164_part partition, and the resulting call will be routed to the PSTN.

Figure 9-13 Variable-Length Dial Plans with Flat Addressing Without Site Codes



This configuration variant allows you to simplify the dialing rules to be followed by the users. If the destination is located within the site, use abbreviated dialing (omitted from Figure 9-13 for clarity). If the destination is outside the site, whether it is on-net or off-net, dial it in the off-net PSTN form.

- Because you are effectively forcing on-net PSTN calls, remember to configure AAR so that calls can still be placed across the PSTN when the IP WAN bandwidth is not sufficient.
- The placed-calls directory displays digit strings as they were dialed by the user. For example, if the user dialed 1000 and a call was placed to phone +16135551000, the placed-calls directory would display 1000, thus allowing for direct redialing of the number without having to edit the dial string.
- The missed-calls and received-calls directories display phone numbers as they appeared when the call was offered to the phone. Because the DNs are configured as E.164 numbers with +, one-touch dialing is possible. In Figure 9-13, note that the device CSS of the phones can route calls directly to the DNs using the globalized E.164 form, including the + sign.

Deploying Dialed Pattern Recognition in SIP Phones

The dialed pattern recognition capabilities of SIP phones need to take into account the typical dialing habits to be expected from users within the enterprise. Typically any combination of the following patterns may be in use in most enterprises:

- An abbreviated dialing pattern for calls within the same site (In the case of uniform on-net dial plans, the abbreviated dialing pattern could be used for inter-site calls.)
- An inter-site dialing pattern, typically used in variable on-net dial plans when using site codes and an on-net access code such as 8
- An off-net dialing pattern for local calls
- An off-net dialing pattern for long-distance calls
- Emergency call patterns, with and without the off-net access code
- An off-net dialing pattern for international calls

Table 9-6 and Table 9-7 show an example of the SIP dial rules that could be employed in an enterprise with the following dial plan characteristics:

- Abbreviated dialing is four digits (irrespective of whether abbreviated dialing is used for inter-site calls or not)
- Inter-site calls use 8 as an on-net access code, followed by seven digits representing the site code and DN
- Emergency dialing is allowed as 911 and as 9911
- Local seven-digit calls use 9 as an off-net access code, followed by the seven digits
- Local ten-digit calls use 9 as an off-net access code, followed by the ten digits
- Long-distance calls are dialed as 91 and ten digits
- International calls are dialed as 9011 followed by a variable quantity of digits, and dialing can be terminated by #.

Pattern recognition is concerned only with automating the collection of user digit input, to be forwarded automatically to Unified CM without requiring inter-digit timeout or pressing the Dial key. All enforcement of class of service is handled by the various calling search spaces chosen from within Unified CM. That is why all phones are configured with SIP dial rules allowing the recognition of international dialing, for example, even if not all phones will be assigned to an unrestricted class of service.

The dial plan characteristics listed above are representative of the variable-length on-net dial plan with flat addressing (see Deploying Variable-Length On-Net Dial Plans with Flat Addressing, page 9-43). From the standpoint of pattern recognition, this dial plan is compatible with the uniform on-net dial plan and the variable-length on-net dial plan with partitioned addressing (see Deploying Uniform On-Net Dial Plans, page 9-40).

For each pattern in Table 9-6 and Table 9-7, the description provides the pattern in equivalent Unified CM notation. The tables provide the SIP dial rules for both the 7905_7912 and 7940_7960_OTHER cases.

Note The 7905_7912 SIP dial rules are limited to 128 characters, and the 7940_7960_OTHER SIP dial rules are limited to 8K (8,192) characters.

Table 9-6 7940_7960_OTHER Dial Rule

Description	Pattern	Timeout	Effect
Abbreviated 2XXX	2...	0	The combination of these six ranges represents the four-digit abbreviated dialing patterns that could be used at any site. As any string matching [2-7]XXX is dialed, it is sent to Unified CM immediately (timeout = 0).
Abbreviated 3XXX	3...	0	
Abbreviated 4XXX	4...	0	
Abbreviated 5XXX	5...	0	
Abbreviated 6XXX	6...	0	
Abbreviated 7XXX	7...	0	
Inter-site dialing 8.XXXXXXX	8,.......	0	Upon recognition of 8, secondary dial tone is played and seven more digits are collected, followed by immediate forwarding to Unified CM (timeout = 0).
Emergency 911	9,11	0	Upon recognition of 9, secondary dial tone is played and the digits 11 are collected, with immediate forwarding to Unified CM (timeout = 0).
Emergency 9.911	9,911	0	Upon recognition of 9, secondary dial tone is played and the digits 911 are collected, with immediate forwarding to Unified CM (timeout = 0).
Local PSTN 7-digits	9,.......	3	Upon recognition of 9, secondary dial tone is played and seven more digits are collected. Timeout of 3 seconds allows the user to continue dialing when local PSTN ten-digits dialing is configured.
Local PSTN 10-digits	9,..........	0	Upon recognition of 9, secondary dial tone is played and ten more digits are collected, with immediate forwarding to Unified CM (timeout = 0).
Long Distance	9,1..........	0	Upon recognition of 9, secondary dial tone is played and ten more digits are collected, with immediate forwarding to Unified CM (timeout = 0).
International with 6 seconds inter-digit timeout	9,011*	6	Upon recognition of 9, secondary dial tone is played, then 011 and a variable quantity of digits are collected. Timeout of 6 seconds allows for user to pause dialing without triggering a call to an incomplete string.
International with # as end of dialing	9,011*#	0	Upon recognition of 9, secondary dial tone is played, then 011 and a variable quantity of digits are collected, terminated by #. Immediate forwarding to Unified CM (timeout = 0).
Operator	0	0	As soon as 0 is detected, immediate forwarding to Unified CM (timeout = 0).

Table 9-7 7905_7912 Dial Rule

Description	Pattern	Effect
Abbreviated 2XXX	2...t0	The combination of these six ranges represents the four-digit abbreviated dialing patterns that could be used at any site. As any string matching [2-7]XXX is dialed, it is sent to Unified CM immediately (t 0).
Abbreviated 3XXX	3...t0	
Abbreviated 4XXX	4...t0	
Abbreviated 5XXX	5...t0	
Abbreviated 6XXX	6...t0	
Abbreviated 7XXX	7...t0	
Inter-site dialing 8.XXXXXXX	8.......t0	The digit 8 and seven more digits are collected, followed by immediate forwarding to Unified CM (t0).
Emergency 911	911t0	The digits 911 are collected, with immediate forwarding to Unified CM (t0).
Emergency 9.911	9911t0	The digits 9911 are collected, with immediate forwarding to Unified CM (t0).
Local 7-digits and LD	9.......t4>#....t1	The digit 9 and seven more digits are collected, with 4 seconds allowed for up to four other digits to be dialed. If another four digits are entered, they are sent to Unified CM after 1 second. The # would be recognized as the terminating character after 9 and seven digits are entered.
International	9011>#t6-	The digits 9 011 and a variable quantity of other digits are collected. Timeout of 6 seconds allows for user to pause dialing without triggering a call to an incomplete string. The # is allowed as terminating character.
Operator	0	As soon as 0 is detected, immediate forwarding to Unified CM (timeout = 0).

Building Classes of Service for Unified CM

Unified CM offers two main approaches to defining and applying classes of service to users and devices: the traditional approach and the line/device approach. The fundamental elements addressed for each approach include the types of calls to be allowed (for example, local, national, or international) and the path taken by the calls (for example, IP network, local gateway, or central gateway). Both elements depend on the calling search space configuration. The following sections describe the two main approaches used in Unified CM systems to implement classes of service. Both approaches are based on the fundamental functionality of the line and device calling search space.

The device calling search space can be determined dynamically based on where in the network the phone is physically located, as determined by the phone's IP address, if Device Mobility is configured. See Device Mobility, page 9-122, for more details.

Building Classes of Service for Unified CM with the Traditional Approach

With Unified CM, you can define classes of service for IP Telephony users by combining partitions and device calling search spaces with external route patterns, as follows:

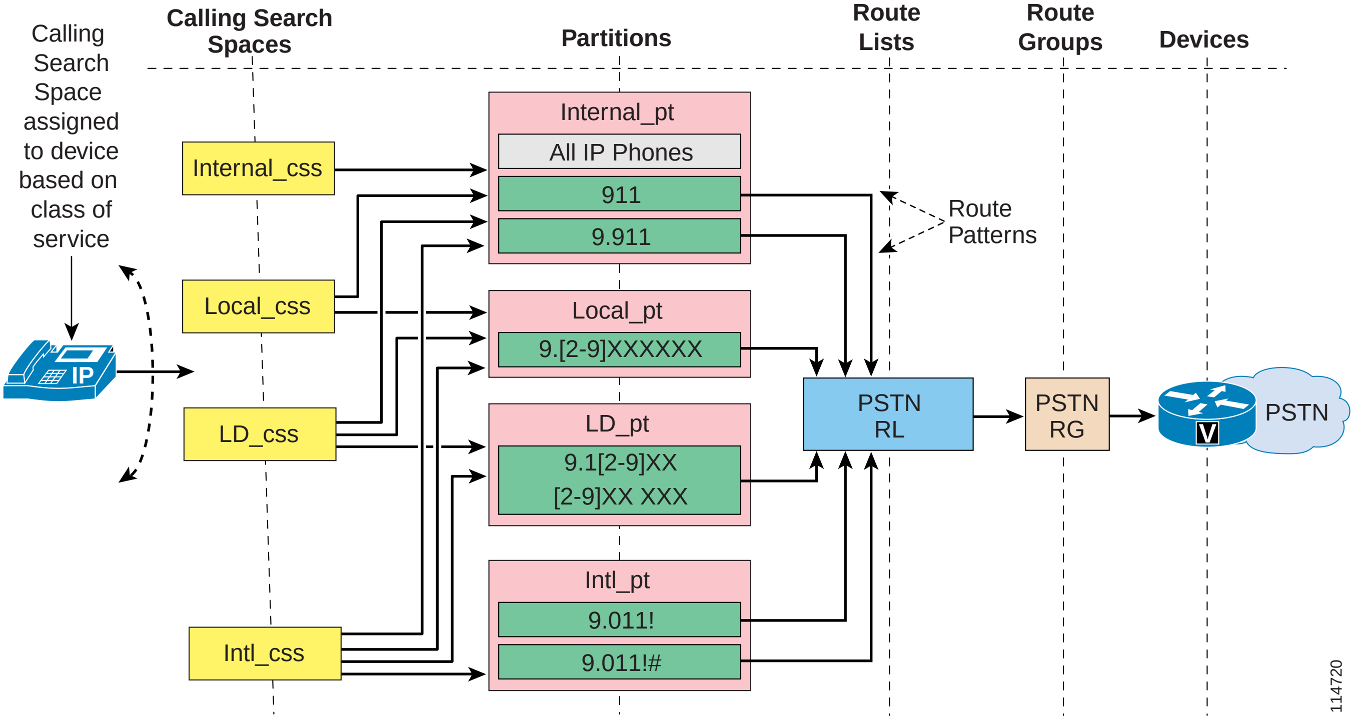
- Place external route patterns in partitions associated with the destinations they can call. While you could place all route patterns in a single partition, you can achieve more refined call restriction policies by associating the route patterns with partitions according to the destinations they can call.

For example, if you place local and international route patterns in the same partition, then all users can reach both local and international destinations, which might not be desirable. Cisco recommends that you group route patterns in partitions according to the reachability policies for the various classes of service.

- Configure each calling search space to be able to reach only the partitions associated with its call restriction policy. For example, configure the local calling search space to point to the internal and local partitions, so that users assigned to this calling search space can place only internal and local calls.
- Assign these calling search spaces to the phones by configuring them on the Unified CM device pages. In this way, all lines configured on the device automatically receive the same class of service.

Figure 9-14 shows a simple example for a single-site deployment.

Figure 9-14 Basic Example of Classes of Service Using the Traditional Approach



With this approach, the device calling search space performs two distinct logical functions:

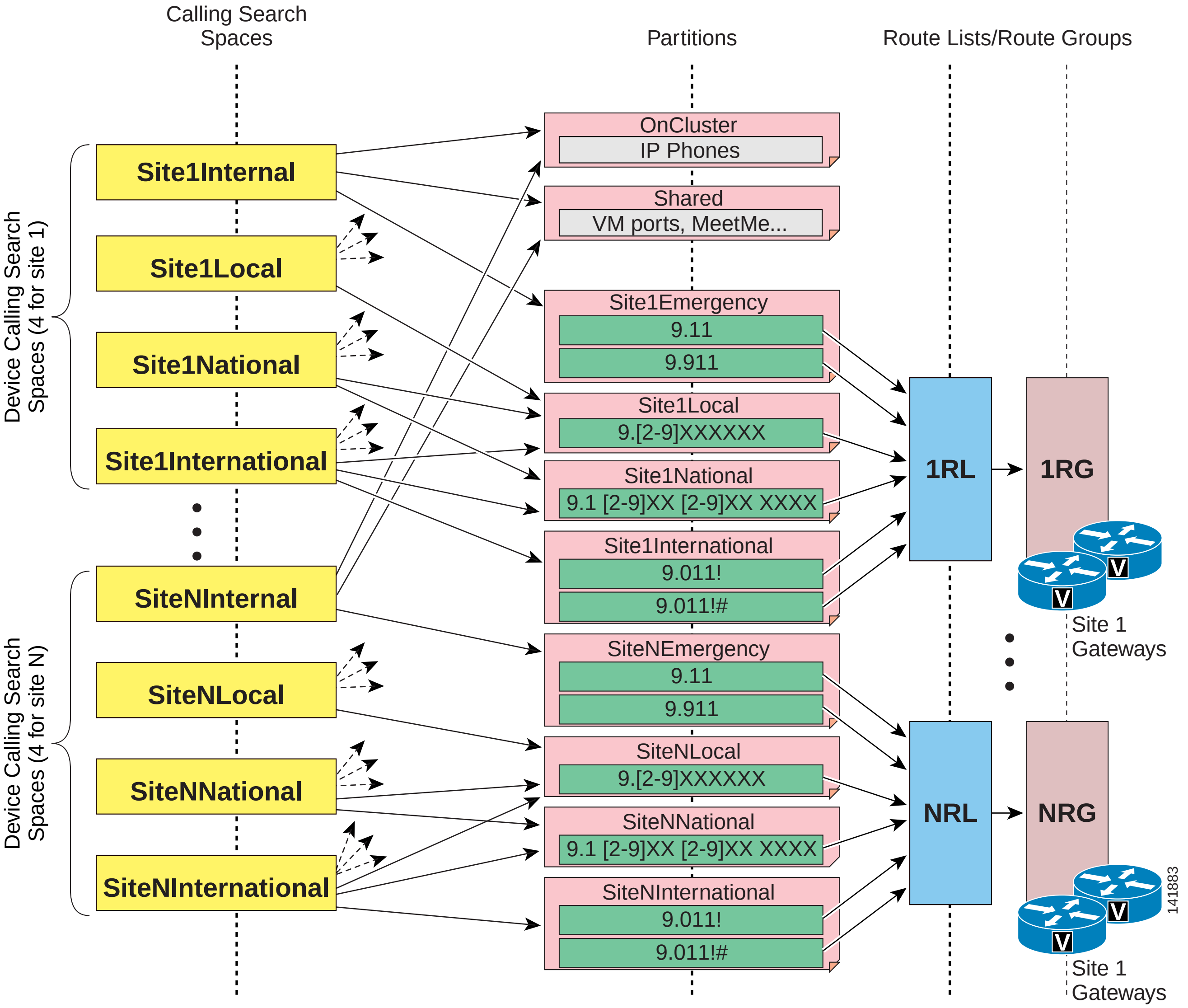
- Path selection

  The calling search space contains specific partitions, which in turn contain specific route patterns that point to specific PSTN gateways through route lists and their associated route groups.
- Class of service

  By selectively including certain partitions and not others in the device calling search space, you effectively apply calling restrictions to certain groups of users.

As a consequence, when you apply this approach to a multisite deployment with centralized call processing, you have to replicate partitions and calling search spaces for each site because for each site you have to create classes of service and, at the same time, route the PSTN calls out of the local branch gateways, as illustrated in Figure 9-15. Alternatively, you can configure the route patterns to point to route lists referencing the Standard Local Route Group, thus allowing the actual egress gateway to be determined by the calling phone's device pool. This allows for the pattern to be reused between sites while retaining site specificity of call routing.

Figure 9-15 Calling Search Spaces and Partitions Needed with the Traditional Approach



To facilitate on-net dialing between sites when applying this dial plan approach to a multisite deployment with centralized call processing, place all IP phone DNs in an on-cluster or internal partition that can be reached from the calling search spaces of all sites. Note that this is not possible if the IP phone DNs overlap.

Extension Mobility Considerations with the Traditional Approach

When using the Extension Mobility feature, the dialing restrictions of a phone are a function of the logged-in (or logged-out) status of the phone. Typically, a logged-out phone should be restricted to calling other phones and services (such as 911), but access to local or toll calls through the PSTN are restricted. Conversely, a phone where a user is logged-in should allow calls according to that user's dialing privileges and should route those calls to the appropriate gateway (for example, a co-located branch gateway for local calls).

With the traditional approach for building classes of service, consider the following guidelines to apply calling restrictions when using Extension Mobility:

- At each site, configure the device calling search space for all IP phones to point to only PSTN emergency services (using the local gateway).
- Configure the line calling search spaces for IP phones used for Extension Mobility in a logged-out state to point to internal numbers only.
- For each Extension Mobility user, configure the line calling search space within the device profile to point to internal numbers and the additional PSTN route patterns allowed for their specific class of service (again, using an appropriate gateway according to company policy).

Note that, when an Extension Mobility user who is normally based in Site 1 logs into an IP phone in Site 2, the path selection for PSTN calls will change as follows:

- Emergency calls will be correctly routed using Site 2's PSTN gateway because the emergency services are provided by the device calling search space of the IP phone at Site 2.
- All other PSTN calls will be routed according to the Extension Mobility user's profile (more specifically, the line calling search space configured in the device profile). Typically, this means that these PSTN calls will traverse two WAN links and use Site 1's gateway to access the PSTN.

You can use one of the following methods to modify this behavior and ensure that PSTN calls are always routed via the local PSTN gateway even when Extension Mobility users roam across different sites:

- Include local PSTN route patterns in the device calling search space and remove them from the line calling search space within the device profile. This method ensures that local PSTN calls will be routed via the co-located branch gateway, but it also means that users will be able to dial these calls even without logging into the IP phones. Long-distance and international calls will still be routed according to the Extension Mobility user's device profile, so this solution is suitable only if these calls are usually routed via a centralized gateway.
- Define multiple device profiles for each user, one for each site to which they usually roam. Each device profile is configured so that its line calling search space points to PSTN route patterns that use the local gateway for that site. This method might prove burdensome to configure and manage if a significant number of users roam to a significant number of sites.
- Implement the line/device approach described in the next section on Building Classes of Service for Unified CM with the Line/Device Approach, page 9-57.

Note When Cisco Emergency Responder is used, the site-specific calling search space configured on the device should include the partition containing the 911 CTI route point pointing to Cisco Emergency Responder. This same partition can also contain a translation pattern 9.911 pointing to the same 911 CTI route point, to allow users to dial 9911 when trying to reach emergency services.

Building Classes of Service for Unified CM with the Line/Device Approach

The traditional approach outlined in the preceding section can result in a large number of partitions and calling search spaces when applied to large multisite deployments with centralized call processing. This configuration is required because the device calling search space is used to determine both the path selection (which PSTN gateway to use for external calls) and the class of service.

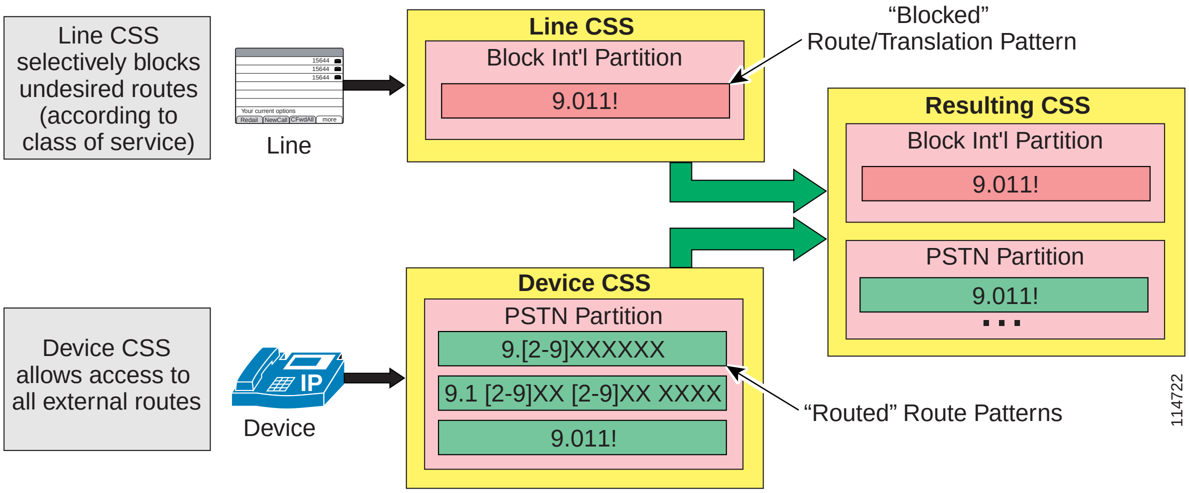
It is possible to significantly decrease the total number of partitions and calling search spaces needed by dividing these two functions between the line calling search space and the device calling search space, in what is called the *line/device approach*.

Keeping in mind how the line calling search space and the device calling search space for each given IP phone are combined within Unified CM, and how the line calling search space partitions appear first in the resulting calling search space (see Calling Privileges in Unified CM, page 9-110), you can apply the following general rules to the line/device approach:

- Use the device calling search space to provide call routing information (for example, which gateway to select for PSTN calls).
- Use the line calling search space to provide class-of-service information (for example, which calls to allow).

To better understand how to apply these rules, consider the example shown in Figure 9-16, where the device calling search space contains a partition with route patterns to all PSTN numbers, including international numbers. The route patterns point to a PSTN gateway via the route list and route group construct.

Figure 9-16 Key Concepts in the Line/Device Approach



At the same time, the line calling search space contains a partition with a single translation pattern that matches international numbers and that has been configured as a blocked pattern.

The resulting calling search space therefore contains two identical patterns matching international numbers, with the blocked pattern in the line calling search space appearing first. The result is that international calls from this line will be blocked.

It is possible to use route patterns instead of translation patterns to block calls within the line calling search space. To configure a blocked route pattern, first create a "dummy" gateway with an unused IP address and place it into a "dummy" route list and route group construct. Then point the route pattern to the dummy route list. The main difference between using a route pattern and a translation pattern to block calls is the end-user experience when trying to dial a blocked number, as follows:

- When a route pattern is used, the end users will be able to dial the entire number and only then will they hear a fast-busy tone.

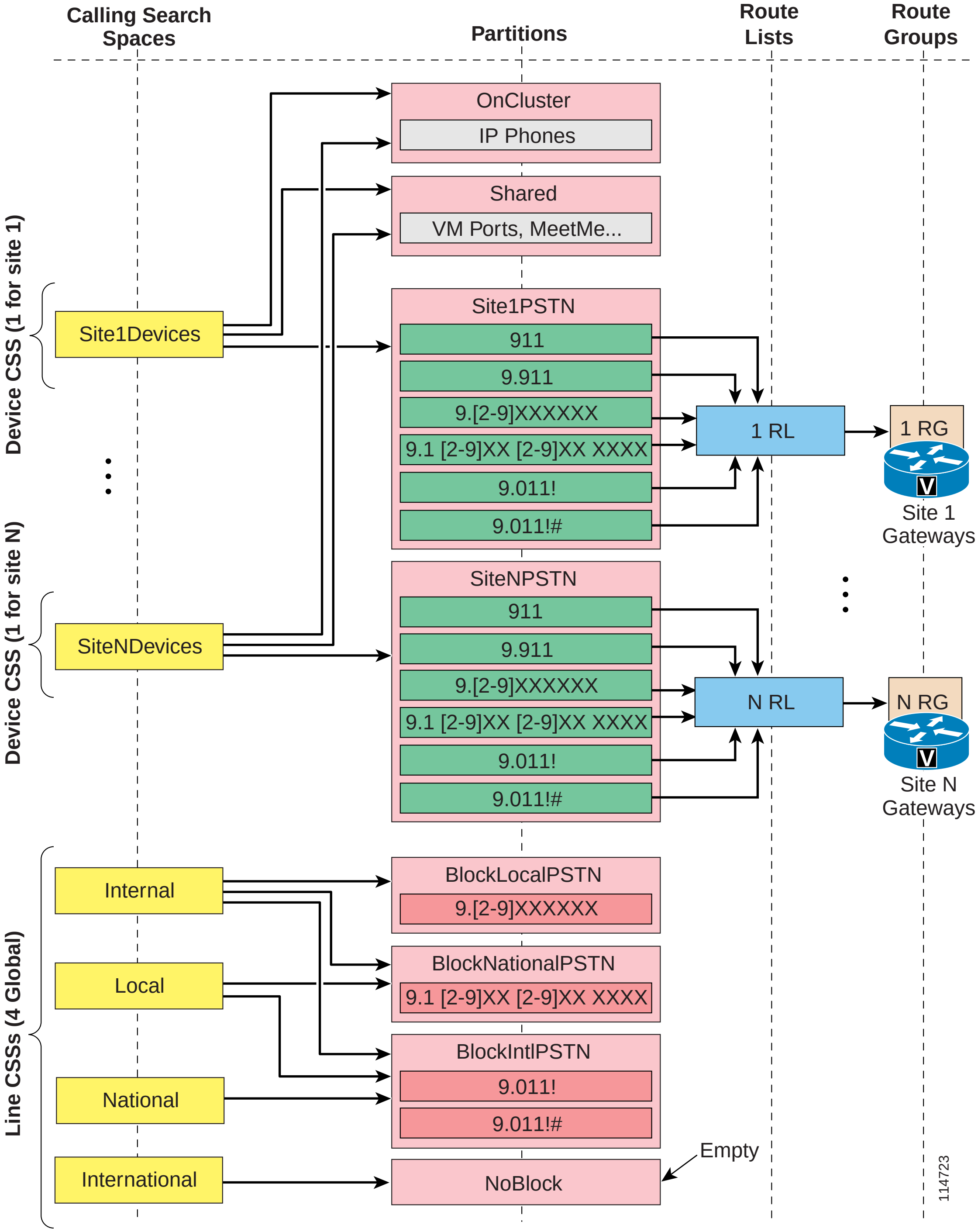
- When a translation pattern is used, the end users will hear a fast-busy tone as soon as the number they are dialing can no longer match any allowed pattern. This behavior assumes an IP phone running SCCP, or an Type-B IP phone running SIP with no SIP dial rules configured in the phone.

Follow these guidelines to implement the line/device approach in a multisite deployment with centralized call processing:

- Create an unrestricted calling search space for each site and assign it to the phone's device calling search space. This calling search space should contain a partition featuring route patterns that route the calls to the appropriate gateway for the phone's location (for example, a co-located branch gateway for emergency services and a centralized gateway for long-distance calls).
- Create calling search spaces containing partitions featuring blocked translation/route patterns for those types of calls not part of the user's dialing privileges, and assign them to the user's lines. For instance, if a user has access to all types of calls except international, that user's line (or lines) should be configured with a calling search space that blocks the 9.011! route pattern.

Figure 9-17 shows an example of how these guidelines can apply to a multisite deployment with N sites.

Figure 9-17 Calling Search Spaces and Partitions Needed with the Line/Device Approach



This approach has the significant advantage that only a single site-specific, unrestricted calling search space is required for each site (that is, one per branch). Because the dialing privileges are implemented through the use of blocked route patterns (which are not site-dependent), the same set of blocking calling search spaces can be used in all branches.

Consequently, you can use the following formulas to calculate the total number of calling search spaces and partitions needed:

Total Partitions = (Number of classes of service) + (Number of sites) + (1 Partition for all IP phone DNs)

Total Calling Search Spaces = (Number of classes of service) + (Number of sites)

Note These values represent the minimum numbers of partitions and calling search spaces required. You may need additional partitions and calling search spaces for special devices and applications, as well as for on-net patterns for other call processing agents.

Note If Cisco Emergency Responder is used, the 911 CTI route pattern and 9.911 translation pattern can be placed in the global On-Cluster partition.

When applied to centralized call processing deployments with large numbers of sites, the line/device approach drastically reduces the number of partitions and calling search spaces needed. For example, a deployment with 100 remote sites and 4 classes of service requires at least 401 partitions and 400 calling search spaces with the traditional approach but only 105 partitions and 104 calling search spaces with the line/device approach.

However, the line/device approach relies on the fact that you can globally identify the types of PSTN calls that must be restricted for certain classes of service (for example, local, long-distance, and international calls). If the national numbering plan of your country does not allow this global identification of the different types of calls, the efficiency of this approach (in terms of configuration savings) is lower than that indicated in the formulas above.

For example, in France the numbering plan is based on five area codes, from 01 to 05 (plus the 06 area code for cellular phones), followed by eight digits for the subscriber number. The key characteristic is that each PSTN destination is reached by dialing exactly the same number (for example, 01XXXXXXXX for Paris numbers, 04XXXXXXXX for Nice numbers, and so on), whether calling from the same local area or from a different area. This means that it is not possible to block access to long-distance calls with a single partition and a single route pattern because the concept of "long-distance call" changes depending on the area where the calling party is located. (For example, a call to 014455667788 is a local call if the caller is in Paris but a long-distance call is the caller is in Nice or Lyon.)

In such cases, you will have to configure additional sets of blocking calling search spaces and partitions, one for each area within which local and long distance calls are dialed the same way. In the example of France, you would have to defining five additional blocking calling search spaces and partitions, one for each area code, as shown in Table 9-8:

Table 9-8 The Line/Device Approach Applied to the French National Numbering Plan

Calling Search Space	Partition	Blocked Route Patterns
Internal_css	BlockAllNational_pt	0.0[1-6]XXXXXXXX
	BlockIntl_pt	0.00!, 0.00!#
Local01_css	BlockLD01_pt	0.0[2-6]XXXXXXXX
	BlockIntl_pt	0.00!, 0.00!#
Local02_css	BlockLD02_pt	0.0[13-6]XXXXXXXX
	BlockIntl_pt	0.00!, 0.00!#

Table 9-8 The Line/Device Approach Applied to the French National Numbering Plan

Calling Search Space	Partition	Blocked Route Patterns
Local03_css	BlockLD03_pt	0.0[124-6]XXXXXXXX
	BlockIntl_pt	0.00!, 0.00!#
Local04_css	BlockLD04_pt	0.0[1-356]XXXXXXXX
	BlockIntl_pt	0.00!, 0.00!#
Local05_css	BlockLD05_pt	0.0[1-46]XXXXXXXX
	BlockIntl_pt	0.00!, 0.00!#
LD_css	BlockIntl_pt	0.00!, 0.00!#
Intl_css	NoBlock_pt	none

Guidelines for the Line/Device Approach

Consider the following guidelines when using the line/device approach:

- For this approach to work, the blocked patterns configured within the line calling search space must be at least as specific as the route patterns configured within the device calling search space. Wherever possible, Cisco recommends that you configure the blocked patterns as more specific than the routed ones, to avoid any possibility of error. Use extra care when dealing with the @ wildcard because the patterns defined within it are very specific.
- AAR is triggered when on-net DNs are dialed. Access to these DNs can be controlled by the same processes described previously. AAR uses a different calling search space for rerouted calls. In most cases, the AAR calling search space can be the same as the site-specific, unrestricted device calling search space because it can never be dialed directly by end users.
- Refer to the section on Call-Forward Calling Search Spaces, page 9-113, for guidance on using the line/device approach for Call Forward All actions.

Note The priority order between line and device is reversed for CTI devices (CTI route points and CTI ports). For these devices, the partitions in the device calling search space are placed before the line calling search space in the resulting calling search space. Therefore, the line/device approach cannot be applied to CTI devices such as Cisco IP SoftPhone unless you are careful not to rely solely on the concatenation order for pattern selection but instead ensure that the desired blocked pattern's precision is greater in all cases than that of the permitted pattern(s).

Globalized Numbers and Class of Service

System administrators using the line-device approach for calling search spaces should be aware that the blocked patterns used in the line CSS of endpoints might have to block not only the localized form but also the globalized form of calls. While the localized form of a number lends itself to classification as local, regional, or national, the globalized form does not. This can lead to class-of-service disparities, where direct user dialing is subjected to a class of service while one-touch dialing from the missed and received calls lists is not.

For example, consider creating a local class of service for the city of Ottawa, Ontario, Canada. All local calls in Ottawa fall into the area codes 613 and 819, and local calling is implemented using 10-digit dialing. If only localized user input is allowed on a phone in Ottawa, the class of service "local" can be imposed on a phone by allowing only calls made in the form 9[2-9]XX[2-9]XXXXXX. Any call made to a national (long distance) destination would start with a different dialing form (off-net access code 9 followed by the national steering code 1, followed by the number), as would international calls (9 followed by 011). The form of the call defines its class.

If one-touch redial is to be implemented, the global form of local numbers is to be allowed in the phone's dial plan. For a dial plan based on line-device approach, where class-of-service is implemented through blocked patterns addressed by line calling search spaces, this means that a series of blocked patterns has to be implemented to allow only calls to area codes 613 and 819. This could be achieved, for example, by the following blocked patterns:

```
\+1[^68]!
\+16[^1]!
\+161[^3]!
\+18[^1]!
\+181[^9]!
```

Typically we require a blocked pattern per significant digit and digit string to be allowed. The above set of blocked patterns would allow local calls to be one-touch dialed from the missed or received calls list.

The situation becomes more complicated, however, because not all 613 and 819 area code destinations are local calls. While the localized patterns will permit a user to initiate a call only to local destinations (by dialing 9 819 or 9 613 as the beginning of the dial string), the globalized patterns will allow a user to receive a call from a non-local number in area code 613 or 819, go to the received calls list, and one-touch dial the number back, matching the globalized patterns. In such instances, the global form blocked patterns should be refined to represent exactly the local calling area. For the example above, this would entail defining the exact subset of area codes 613 and 819 that are within the local calling area for Ottawa. The set of refined blocked patterns can become very complex, depending on the structure of the local calling area.

Also, these +E.164 blocked patterns, in contrast to the localized form, are site-specific and cannot be reused for other sites. The line calling search space inherits this site specificity, so that the line calling search space implementing class-of-service "local" for Ottawa in the above example is specific to the site in Ottawa. This fundamentally breaks the whole idea of the line-device approach to decrease the number of required calling search spaces by creating classes of services through site-unspecific calling search spaces that can be used universally for all sites.

Extension Mobility Considerations for the Line/Device Approach

When using the Extension Mobility feature, the line/device approach provides a natural way to implement the dialing restrictions of a phone as a function of the logged-in (or logged-out) status of the phone. Typically, a logged-out phone should be restricted to calling other phones and services (such as 911), but access to local or toll calls through the PSTN are restricted. Conversely, a phone where a user is logged-in should allow calls according to that user's dialing privileges and should route those calls to the appropriate gateway (for example, a co-located branch gateway for local calls).

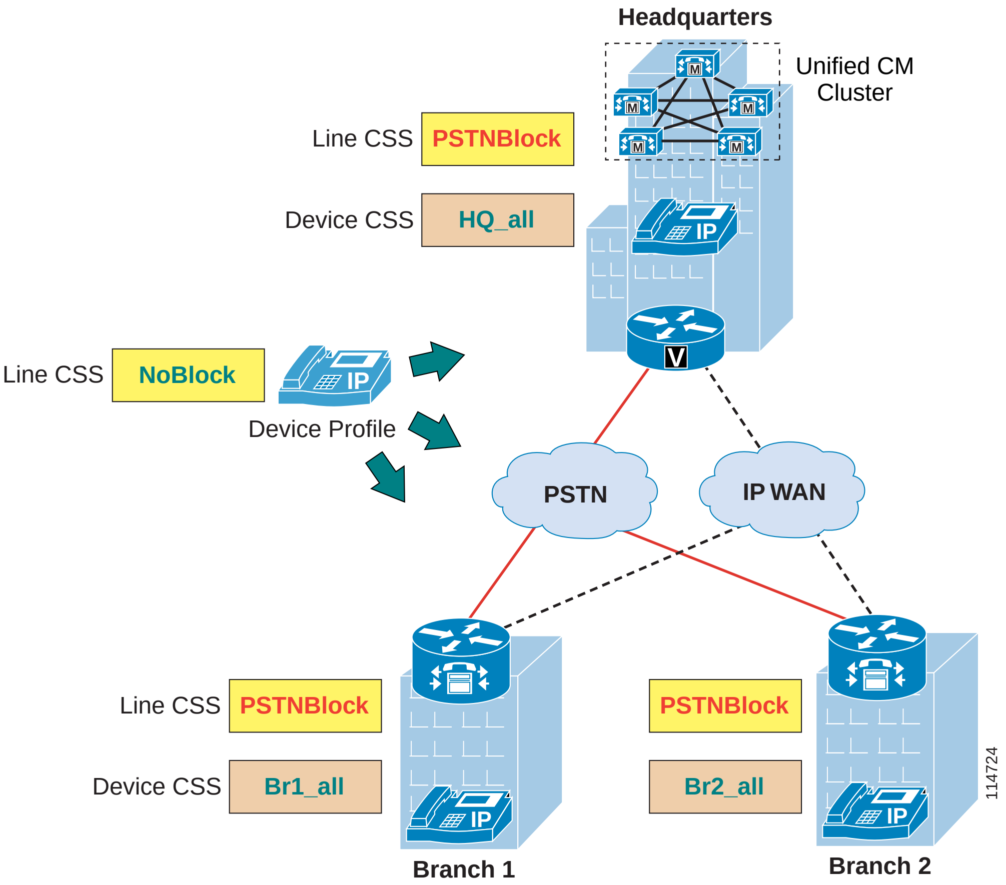
With the line/device approach for building classes of service, you can simply apply the same rules described in the previous section to the Extension Mobility device profile construct. Consider the following guidelines to apply calling restrictions when using Extension Mobility:

- At each site, configure the device calling search space for all IP phones to point to a site-specific partition that contains all possible PSTN route patterns and that routes them appropriately (for example, using the local gateway for emergency and local calls and a centralized gateway for long distance calls).

- Configure the line calling search space for all IP phones (or the line calling search space for the default logout device profile) to point to a global calling search space featuring blocked translation/route patterns that block all calls except those to be allowed when no user is logged in (for example, internal extensions and emergency services).
- For each Extension Mobility user, configure the line calling search space within the device profile to point to a global calling search space featuring blocked translation/route patterns to selectively block the PSTN calls that are not to be allowed for their specific class of service (for example, block only international calls). If some users must have unrestricted calling privileges, assign them to a line calling search space featuring an empty partition.

The key advantage of using the line/device approach for extension mobility is that, in a multisite deployment with centralized call processing, appropriate call routing is ensured even when users log in to IP phones located at branch sites different from their home site, as illustrated in Figure 9-18.

Figure 9-18 Extension Mobility with the Line/Device Approach



As described previously in this chapter, the line calling search space configured within the device profile replaces the line calling search space configured on the physical IP phone when a user logs in through Extension Mobility. Because the call routing is correctly determined by the device calling search space, the login operation is used merely to "unlock" the phone by removing the phone's line calling search space (which contains blocked patterns) and replacing it with the device profile's line calling search space (which does not contain blocked patterns in this simplified example).

Because all the call routing is done within the device calling search space while the line calling search space only introduces blocked patterns, whenever a user logs in at a different site from their home site, they will automatically inherit all the local dialing habits for that site. For example, assume that phone DNs are defined as eight-digit numbers, but four-digit dialing is provided within each site by local translation patterns. In this case, a user roaming to a different site will not be able to dial a colleague at the home site by using only four digits because the four-digit number will now be translated according to the rules of the host site where the user logged in.

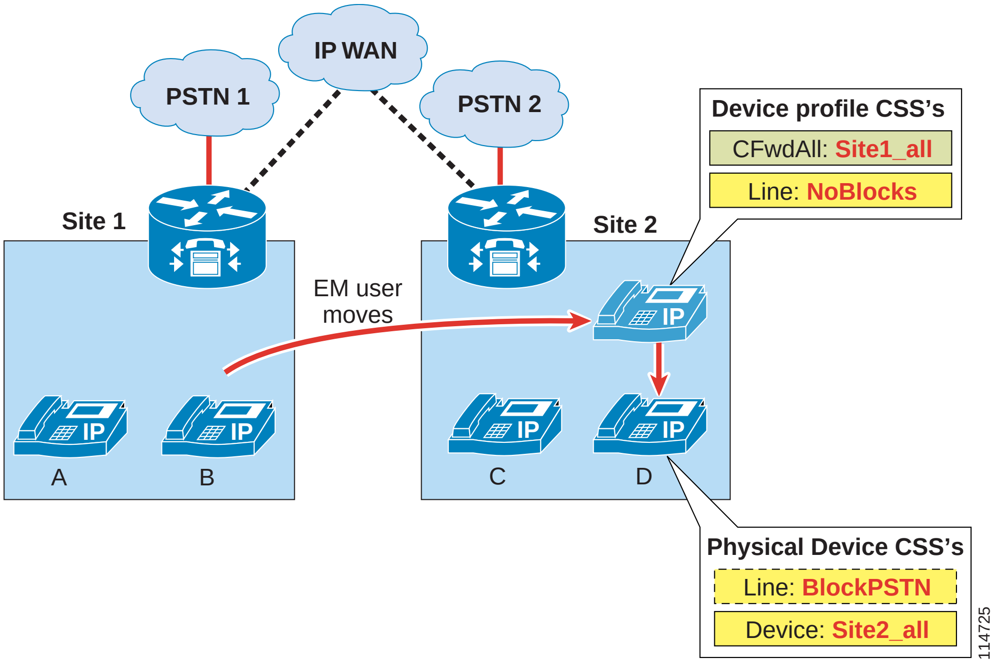
In summary, when you use the line/device approach for Extension Mobility, end-users have to adopt the dialing behavior of the site where they logged in.

Call Forwarding Considerations

When applying the Line/Device calling search space approach to a centralized call processing environment with Extension Mobility, you should be aware of the call forwarding behavior if users need to be allowed to forward all their calls to external PSTN numbers.

In Figure 9-19, an Extension Mobility user is normally based in Site 1, with a device profile allows the user to place unrestricted PSTN calls and to forward all incoming calls to any PSTN number.

Figure 9-19 Call Forwarding Considerations for Extension Mobility with the Line/Device Approach



As described in the section on Call-Forward Calling Search Spaces, page 9-113, the Forward All calling search space is not concatenated with the line or the device calling search spaces and therefore needs to be set to Site1_all, which includes all PSTN routes using the Site 1 gateway.

When this user moves to Site 2 and logs into phone D, the user's device profile applies its line calling search space and Forward All calling search space(s) to the physical device. For direct PSTN calls, the calling search space used is the concatenation of the line and device calling search space, and phone D's device calling search space (Site2_all) correctly points to the Site 2 gateway.

If the user now configures the phone to forward all calls to a PSTN number, any forwarded call will use the Site1_all calling search space, which still points to the Site 1 gateway. This condition results in the following behavior:

- Incoming PSTN calls enter the IP network at the Site 1 gateway and are hairpinned back into the PSTN within the same gateway.
- Calls originating from Site 1 phones (such as phone A) are correctly forwarded to the PSTN via the Site 1 gateway.
- Calls originating from Site 2 phones (such as phone C) traverse the WAN to Site 1 and access the PSTN via the Site 1 gateway. The same behavior applies to calls originating from other sites within the same Unified CM cluster.

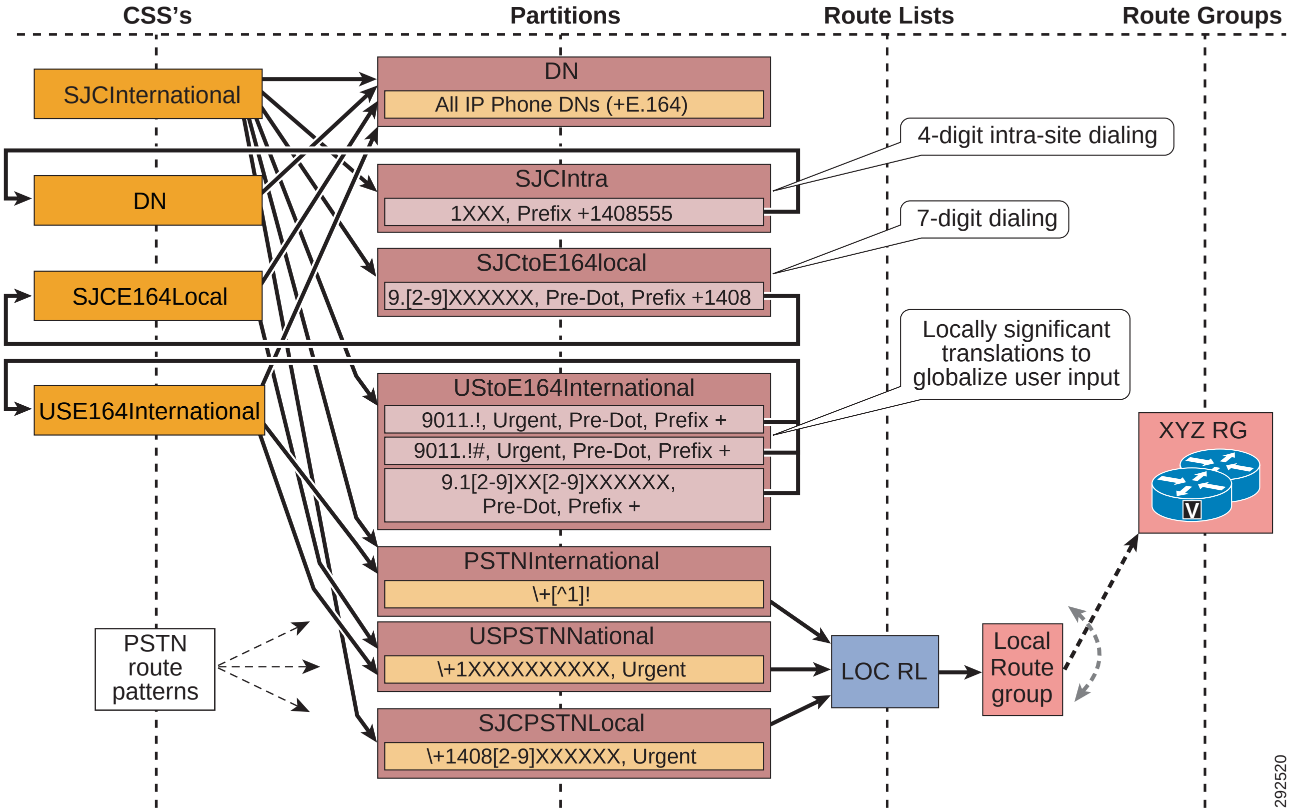
Keep this behavior in mind when designing the network and training the users.

Building Classes of Service for Unified CM for +E.164 Dial Plans with the Traditional Approach and Local Route Group

Trying to support globalized +E.164 dialing with the line-device approach has the problem that the required blocked patterns for some classes of service make the line calling search spaces site-specific, which negates the design goal of the line-device approach to minimize the number of required calling search spaces by using site-unspecific line calling search spaces addressing the appropriate blocked patterns. With that in mind, the advantage of the line-device approach over the traditional approach seems to be minimal, especially since the introduction of the concept of local route groups with Unified CM 7.0. The use of local route groups allows the administrator to move the site-specific egress gateway selection from the route pattern to the local group selection specific to the calling device. This selection process uses the local route group configured in the calling device's device pool.

The main difference of this approach is that the effective class of service is not the result of combining route patterns on the device blocked patterns on the line, but is the direct result of any pattern addressed by the single class-of-service specific calling search space. (See Figure 9-20.)

Figure 9-20 Class of Service "International" for +E.164 Dial Plans



[Figure 9-20](#) shows an example of how to implement class of service "international" for a single site in the US. In this concept, the local habitual dialing is normalized to +E.164 through translation patterns in partitions SJCIntra, SJCtoE164local, and UStoE164International.

The translation in partition SJCIntra implements 4-digit intra-site dialing, assuming that all local DIDs of the site are in the range +1 408 555 1XXX. Local dialing (9+7) for the site in San Jose is implemented by the translation pattern in partition SJCtoE164local by again transforming the local habitual dialing to +E.164. The same is true for partition UStoE164International, which implements the globalization of US habitual PSTN dialing to international and national destinations.

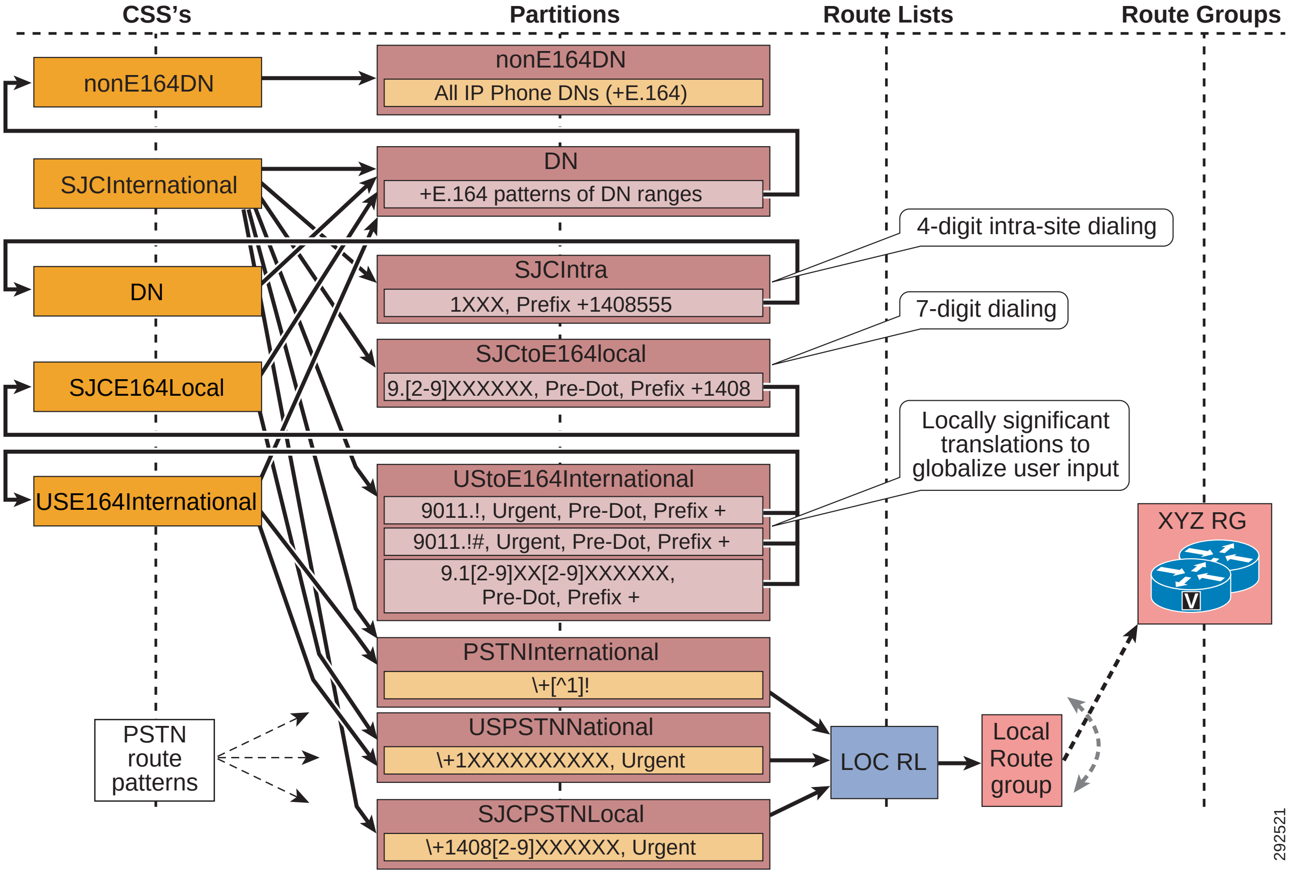
The naming convention used here helps to identify which pieces of the dial plan need to be replicated to support multiple classes of service, sites, and dialing domains. If the name includes the specification of a site (for example, SJC in partition name SJCE164Local), then this element needs to be replicated for every site. If the name includes the specification of a class of service (for example, International in USE164International), then it needs to be replicated for every class of service. If the name does not include the specification of a site (for example, partition USPSTNNational), then it can be reused for all sites.

The single calling search space creating the requested class of service can be used as a line or device calling search space. In deployments that support mobility features such as extension mobility or device mobility, the line calling search space has to be used to enable the user to keep his class of service when roaming.

Using Non +E.164 Directory Numbers

The above example assumes that all directory numbers are configured as +E.164 so that the directory numbers can be matched directly after normalizing the local habitual dialing to +E.164. In cases where directory numbers are not configured as +E.164, an intermediate routing step is required to again map from the internal +E.164 routing to the format of the configured DNs. (See Figure 9-21.)

Figure 9-21 Indirection to Support Non +E.164 Directory Numbers



The additional indirection step shown in Figure 9-21 makes sure that all directory numbers not in +E.164 format can be reached by dialing either the local habitual format or +E.164. Partition DN in this case does not hold the actual directory numbers but a set of urgent translation patterns matching all on-net DID ranges. If, for example, a site uses DID range +1 408 555 1XXX and the directory numbers are configured as E.164 without the leading +, then an urgent translation pattern \+.14085551XXX needs to be created in partition DN with the digit discard instruction set to discard pre-dot. A user with extension +1 408 555 1234 can now be reached from other users using the calling search space in the example by dialing:

- 1234: Translation pattern in partition SJCIntra transforms dialed digits to +14085551234, translation pattern +14085551XXX in partition DN transforms to 14085551234, and then we get a match on the directory number in partition nonE164DN.
- 95551234: Translation pattern in partition SJCtoE164local globalizes the dialed digits and then the call flow is the same as for 4-digit intra-site dialing.

- 914085551234: Translation pattern in partition UStoE164International globalizes the dialed digits and then the call flow is the same as for 4-digit intra-site dialing.
- +14085551234: Translation pattern +14085551XXX in partition DN transforms to 14085551234 and then we get a match on the directory number in partition nonE164DN.

Keep in mind that, when using non +E.164 directory numbers, you will have to make sure that the calling party number for calls originating from these lines is set to +E.164 to ensure that correct calling party information can be delivered for all call flows. This can be achieved by setting the external phone number mask to +E.164 for all line appearances and configuring the calling party transformations on the translation patterns in partition DN to use the external phone number mask. Another option to globalize non +E.164 directory numbers is to use the incoming call's calling party transformation calling search space on the phone or phone's device pool. This method is supported in Cisco Unified CM 9.0 and later releases, and is the recommended way to globalize directory numbers. Using globalization based on this calling search space has the advantage of also working with URI-dialed call flows for which number transformations configured on translation patterns are not applicable.

Overlaps Between Directory Numbers and International PSTN Dialing

Directory numbers in Unified CM always are stored in digit analysis as non-urgent patterns. This can lead to the situation that dialing an international on-net destination causes a partial overlap with the international PSTN route pattern \+[^1]!. For example, if a user in San Jose dials +496100773 and that happens to be a directory number configured on the system, then the call using the schema without the additional indirection step (Figure 9-20) will hit the interdigit time-out limit because, although we have a match on directory number \+496100773, there is another potential match on the variable length pattern \+[^1]! in partition PSTNInternational. One way to avoid this is to add specific, urgent fixed-length route patterns to the PSTNInternational partition for all countries where we have on-net directory numbers. The urgent route pattern in the PSTNInternational partition will terminate digit collection, but Unified CM will then still route the call to the best match, which in this case is the directory number in partition DN. This approach works only for countries with a fixed-length national numbering plan. If the national numbering plan is a variable-length numbering plan (such as in Germany), the only way to solve the overlap between directory numbers and the variable-length PSTN route pattern is to implement the intermediate routing step as shown in Figure 9-21. In this case the urgent translation patterns in partition DN serve the purpose of adding urgent patterns to digit analysis that will make sure that digit collection terminates as soon as a known on-net destination is dialed.

Other Classes of Service

Based on the example in Figure 9-20 and Figure 9-21, other classes of service such as "national" and "local" are created by removing the unnecessary PSTN route patterns and recreating the dialing normalization of the local habitual dialing habits. (See Figure 9-22.)

Figure 9-22 Class of Service "National" for +E.164 Dial Plans

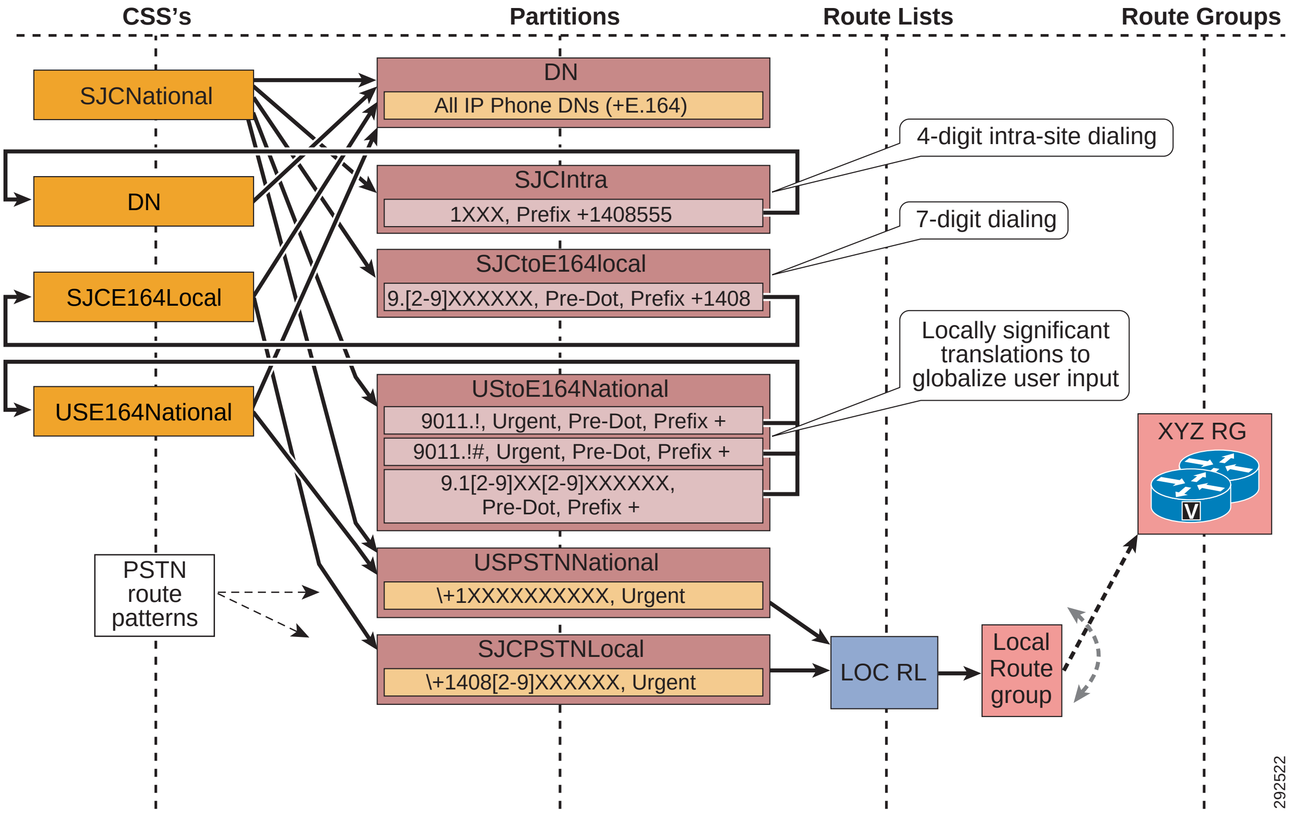
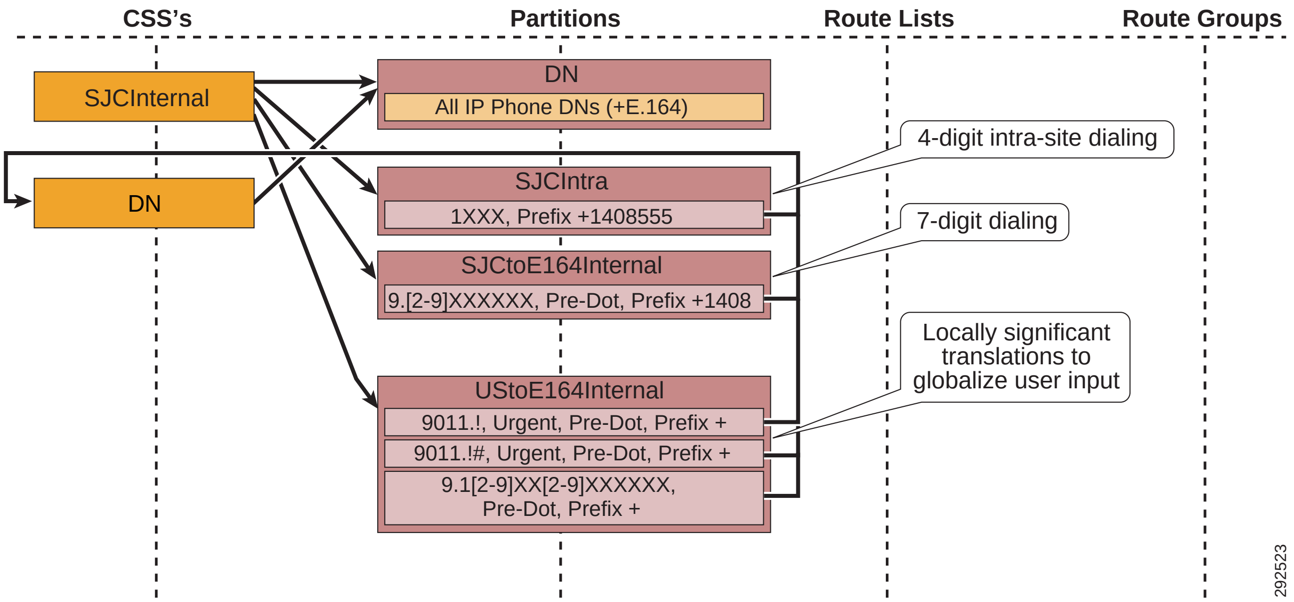

CSS's
Partitions
Route Lists
Route Groups
SJCNational
DN
SJCE164Local
USE164National
PSTN route patterns
DN
All IP Phone DNs (+E.164)
SJCIntra
1XXX, Prefix +1408555
SJCtoE164local
9.[2-9]XXXXXX, Pre-Dot, Prefix +1408
UStoE164National
9011.!, Urgent, Pre-Dot, Prefix +
9011.!#, Urgent, Pre-Dot, Prefix +
9.1[2-9]XX[2-9]XXXXXX, Pre-Dot, Prefix +
USPSTNNational
\+1XXXXXXXXXX, Urgent
SJCPSTNLocal
\+1408[2-9]XXXXXX, Urgent
4-digit intra-site dialing
7-digit dialing
Locally significant translations to globalize user input
LOC RL
Local Route group
XYZ RG
V
292522

Figure 9-22 shows how to build class of service "national." Comparing this schema to class of service "international" in Figure 9-20, we see that partitions SJCIntra, SJCtoE164local, USPSTNNational, and SJCPSTNLocal can be reused and so can calling search spaces DN and SJCE164Local. The dialing normalization in USToE164National has to be recreated because using the dialing normalization as defined by the patterns in partition UStoE164International in Figure 9-20 would also grant access to the international PSTN route pattern \+[^1] in partition PSTNInternational. This is a result of the fact that, in the definition of a translation pattern, we define a single calling search space that has to be used after applying the called and calling party transformations defined in the translation pattern. Because a calling search space is equivalent to a class of service, the translation patterns inherit the class-of-service specificity of the egress calling search space defined in the translation patterns and thus also become class-of-service specific.

Dialing normalization in partition UStoE164National does include dialing normalization patterns for the local habitual international dialing 9011 because we need to support international dialing to international on-net destinations (directory numbers in partition DN outside the US).

Classes of service "local" and "internal" can be created using the same approach. Partition SJCPSTNLocal is not really required to build class of service "international", but providing this differentiated PSTN access from the beginning permits reuse of partitions SJCPSTNLocal and SJCtoE164Local for class of service "local". Keep in mind that partition SJCPSTNLocal cannot be reused for class of service "internal," which should not provide access to the PSTN. For this class of service the dialing normalization of the local dialing has to be replicated as shown in Figure 9-23.

Figure 9-23 Class of Service "internal" for +E.164 Dial Plans


CSS's
Partitions
Route Lists
Route Groups
SJCInternal
DN
DN
All IP Phone DNs (+E.164)
SJCIntra
1XXX, Prefix +1408555
SJCtoE164Internal
9.[2-9]XXXXXX, Pre-Dot, Prefix +1408
UStoE164Internal
9011.!, Urgent, Pre-Dot, Prefix +
9011.!#, Urgent, Pre-Dot, Prefix +
9.1[2-9]XX[2-9]XXXXXX, Pre-Dot, Prefix +
4-digit intra-site dialing
7-digit dialing
Locally significant translations to globalize user input
292523

Emergency Calls

Access to emergency services has to be granted to all users. This can be achieved either by adding the partition with the emergency number route patterns to each calling search space or by enabling access to the emergency number route patterns through the device-level calling search space. If access to emergency numbers is granted through the device calling search space, then in roaming scenarios (for example, extension mobility) the user has to dial emergency services using the habitual dialing of the visited site, while access to emergency numbers through the line calling search space would allow the user to dial emergency services using the habitual dialing of the home site. This differentiation obviously is important only if the habitual dialing of emergency services differs between home and visiting site as, for example, in the case of a European user (emergency number 112) logging into an US phone (emergency number 911).

Building Classes of Service in Cisco IOS with H.323

The following scenarios require you to define classes of service within Cisco IOS routers running the H.323 protocol:

- Unified CM multisite deployments with centralized call processing
- Cisco Unified Communications Manager Express (Unified CME) deployments

Under normal conditions in Unified CM multisite deployments with centralized call processing, classes of service are implemented using partitions and calling search spaces within Unified CM. However, when IP WAN connectivity is lost between a branch site and the central site, Cisco SRST takes control of the branch IP phones, and all the configuration related to partitions and calling search spaces is unavailable until IP WAN connectivity is restored. Therefore, it is desirable to implement classes of service within the branch router when running in SRST mode.

Similarly, in Unified CME deployments, the router needs a mechanism to implement classes of service for the IP phones.

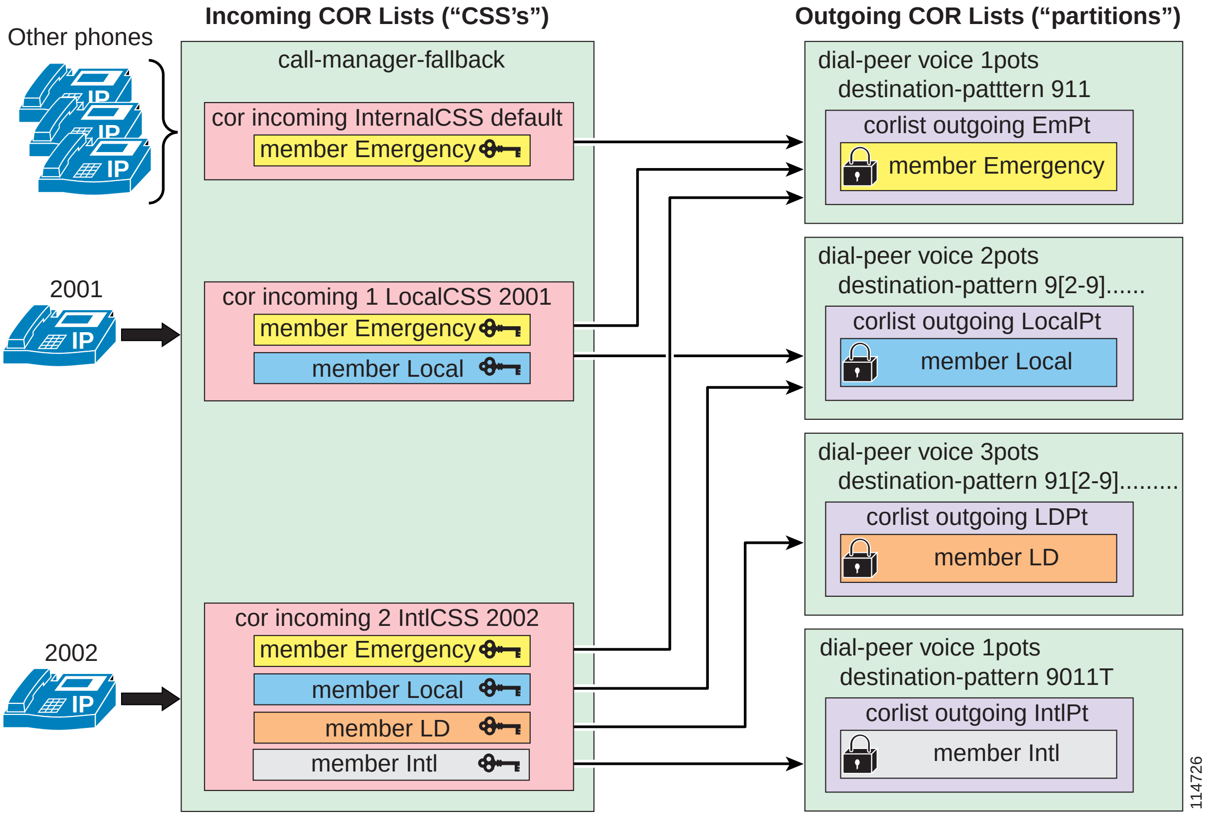
For both of these applications, define classes of service in Cisco IOS routers by using the class of restriction (COR) functionality (refer toCalling Privileges in Cisco IOS with H.323 Dial Peers, page 9-150, for details on COR).

You can adapt the COR functionality to replicate the Unified CM concepts of partitions and calling search spaces by following these main guidelines:

- Define tags for each type of call that you want to distinguish.
- Assign "basic" outgoing corlists (equivalent to partitions), containing a single member tag, to the respective POTS dial peers that route each type of call.
- Assign "complex" incoming corlists (equivalent to calling search spaces), containing subsets of the member tags, to the IP phones that belong to the various classes of service.

Figure 9-24 illustrates an implementation example based on SRST, where the IP phone with DN 2002 is configured to have unrestricted PSTN access, the IP phone with DN 2001 is configured to have only local PSTN access, and all other IP phones are configured to have access only to internal numbers and emergency services.

Figure 9-24 ***Building Classes of Service for Cisco SRST using COR***



The following steps provide examples and guidelines for implementing a Cisco IOS solution like the one shown in Figure 9-24.

Step 1 Using the **dial-peer cor custom** command, define meaningful tags for the various types of calls (Emergency, VMail, Local, LD, Intl):

```
dial-peer cor custom
  name Emergency
  name VMail
  name Local
  name LD
  name Intl
```

Step 2 Using the **dial-peer cor list** command, define basic corlists to be used as partitions, each containing a single tag as a member:

```
dial-peer cor list EmPt
  member Emergency

dial-peer cor list VMailPt
  member VMail

dial-peer cor list LocalPt
  member Local

dial-peer cor list LDPt
  member LD
```

```
dial-peer cor list IntlPt
  member Intl
```

Step 3 Using the **dial-peer cor list** command, define more complex corlists to be used as calling search spaces, each containing a subset of the tags as members according to the classes of service needed:

```
dial-peer cor list InternalCSS
  member Emergency
  member VMail

dial-peer cor list LocalCSS
  member Emergency
  member VMail
  member Local

dial-peer cor list LDCSS
  member Emergency
  member VMail
  member Local
  member LD

dial-peer cor list IntlCSS
  member Emergency
  member VMail
  member Local
  member LD
  member Intl
```

Step 4 Using the command **corlist outgoing** *corlist-name*, configure the basic "partition" corlists as outgoing corlists assigned to the corresponding POTS dial peers:

```
dial-peer voice 100 pots
  corlist outgoing EmPt
  destination-pattern 911
  no digit-strip
  direct-inward-dial
  port 1/0:23

dial-peer voice 101 pots
  corlist outgoing VMailPt
  destination-pattern 914085551234
  forward-digits 11
  direct-inward-dial
  port 1/0:23

dial-peer voice 102 pots
  corlist outgoing LocalPt
  destination-pattern 9[2-9]......
  forward-digits 7
  direct-inward-dial
  port 1/0:23

dial-peer voice 103 pots
  corlist outgoing LDPt
  destination-pattern 91[2-9]..[2-9]......
  forward-digits 11
  direct-inward-dial
  port 1/0:23

dial-peer voice 104 pots
  corlist outgoing IntlPt
  destination-pattern 9011T
  prefix-digits 011
  direct-inward-dial
  port 1/0:23
```

Step 5 Using the **cor incoming** command within the **call-manager-fallback** configuration mode, configure the complex corlists acting as "calling search spaces" to be incoming corlists assigned to the various phone DNs:

```
call-manager-fallback
  cor incoming InternalCSS default
  cor incoming LocalCSS 1 3001 - 3003
  cor incoming LDCSS 2 3004
  cor incoming IntlCSS 3 3010
```

When deploying COR for SRST, keep in mind the following limitations:

- In SRST version 2.0, available on Cisco IOS Release 12.2(8)T or later, the maximum number of **cor incoming** statements allowed under **call-manager-fallback** is 5 (plus the default statement).
- In SRST version 3.0, available on Cisco IOS Release 12.3(4)T or later, the maximum number of **cor incoming** statements allowed under **call-manager-fallback** is 20 (plus the default statement).

Therefore, if the phone DNs of users with non-default privileges are not consecutive and the SRST site is relatively large, you might have to reduce the number of classes of service in SRST mode to accommodate all the DNs without exceeding these limitations.

Although the preceding example is based on Cisco SRST, the same concepts can be applied to a Cisco Unified Communications Manager Express (Unified CME) deployment, except for the following considerations:

- With Unified CME, the corlist expressing the class of service (equivalent to a calling search space) can be assigned directly to the individual IP phones by using the **cor** {**incoming** | **outgoing**} *corlist-name* command under the **ephone-dn** *dn-tag* configuration mode.
- According to COR general rules, all IP phones for which no corlist is configured have unrestricted access to all dial peers, regardless of their outgoing corlist. Unified CME has no mechanism equivalent to the **cor incoming** *corlist-name* **default** command, which applies default restrictions to all phones.

Deploying Call Coverage

Call coverage functionality is a key feature in many IP telephony deployments. Many customer-focused service companies have to route customer calls to the appropriate service representatives expeditiously. This section focuses on design guidelines for using the hunting mechanism based on hunt pilots, hunt lists, and line groups in Cisco Unified CM Release 4.1 to manage call distribution, and it covers the following main topics:

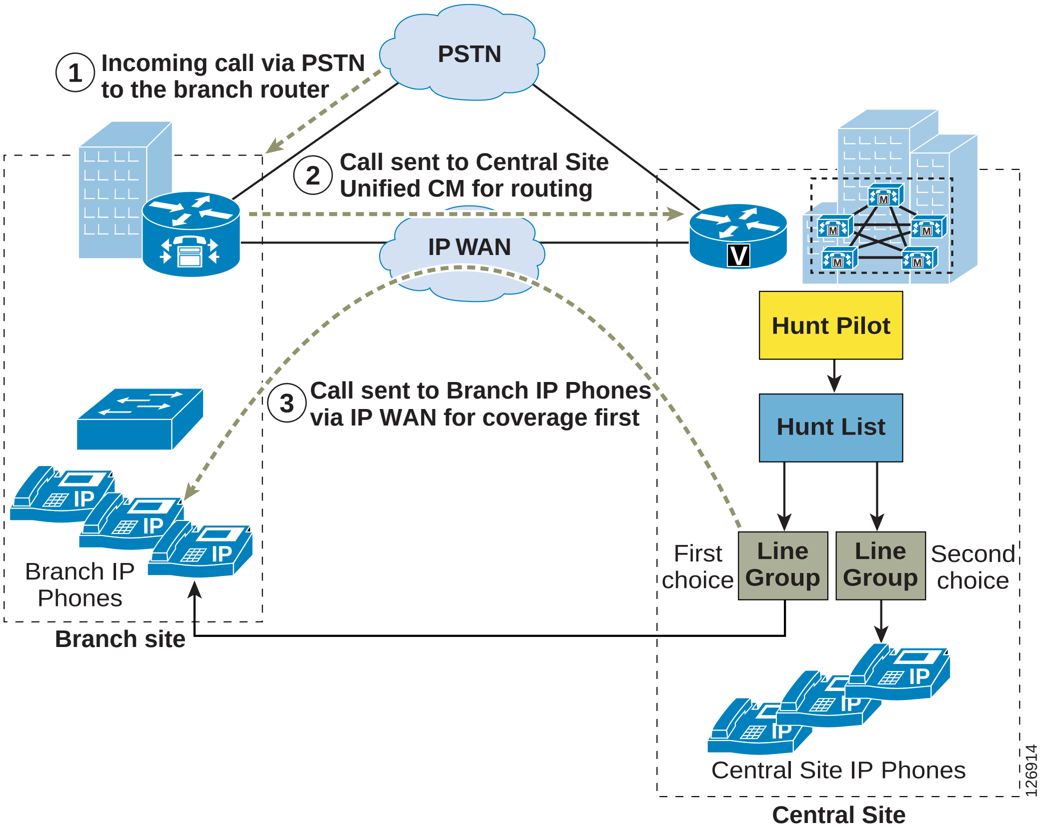
- Deploying Call Coverage in a Multisite Centralized Call Processing Model, page 9-76
- Deploying Call Coverage in a Multisite Distributed Call Processing Model, page 9-77

Note Call coverage functionality does not offer call queues per se, and the caller is presented with ringback tone until a destination is found for the call. To provide prompting, music on hold, and so forth, Cisco offers many contact center technologies such as the Cisco Unified Customer Voice Portal (CVP). For more information on the contact center technologies available from Cisco, refer to the documentation at http://www.cisco.com/go/ucsrnd.

Deploying Call Coverage in a Multisite Centralized Call Processing Model

Figure 9-25 shows an example of configuring hunt lists for a multisite centralized call processing deployment. This example assumes that the hunt pilot call will be distributed first through the remote office operators. If the call is not answered or is rejected due to call admission control, the call will then be routed to central-site operators or voicemail.

Figure 9-25 Call Coverage Between Multiple Sites in a Centralized Call Processing Deployment


1 Incoming call via PSTN to the branch router
PSTN
2 Call sent to Central Site Unified CM for routing
IP WAN
3 Call sent to Branch IP Phones via IP WAN for coverage first
Hunt Pilot
Hunt List
First choice
Line Group
Line Group
Second choice
Branch IP Phones
Branch site
Central Site IP Phones
Central Site
126914

In centralized IP telephony systems, features such as Automated Alternate Routing (AAR) and Survivable Remote Site Telephony (SRST) may be enabled for high availability. Consider the following guidelines when deploying call coverage functionality with AAR or SRST features enabled:

- Automated Alternate Routing (AAR)

  The line group members can be assigned in different locations and regions. Call admission control implemented through locations works as expected. However, a call being distributed from a hunt pilot will not use AAR to reroute a call if Unified CM blocks the call to one of its line group members due to insufficient WAN bandwidth. Instead, Unified CM distributes the call to the next available member or next available line group.

Note Cisco strongly recommends that you use only AAR to voicemail ports within line groups.

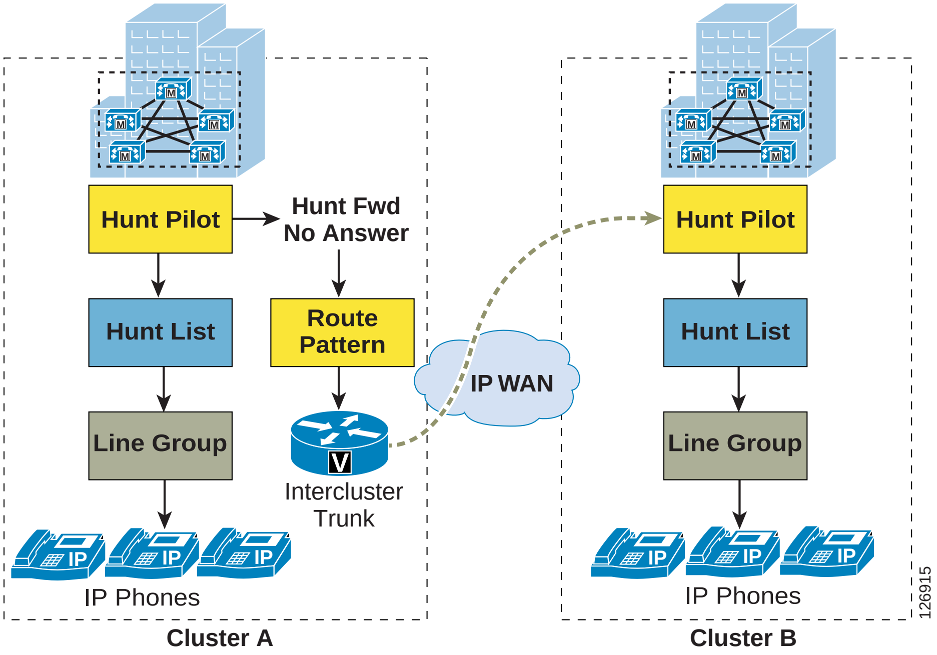
- Survivable Remote Site Telephony (SRST)
  - When Unified CM receives a call for the hunt pilot, and if some of its line group members are at the remote sites operating in SRST mode, then Unified CM skips those members and distributes the call to the next available line group member. From the perspective of Unified CM, the members operating in SRST mode are unregistered, and hunt pilot calls are not forwarded to unregistered members.
  - When a router operating in SRST mode receives a call for the hunt pilot, call coverage functionality is unavailable. The call fails if no configuration is added to reroute the call to a registered and available extension. You can use the **alias** or the **default-destination** command under the **call-manager-fallback** mode in Cisco IOS to reroute the call destined for the hunt pilot to an operator extension or to voicemail.

Deploying Call Coverage in a Multisite Distributed Call Processing Model

Beginning with Cisco Unified CM Release 4.1, route groups can no longer be added to hunt lists. Thus, a hunt list cannot be used to send the calls to other clusters or to a remote gateway. But the hunt option settings in Hunt Pilot, introduced in Cisco Unified CM Release 4.1, can be used to match a route pattern that in turn points to gateways or trunks.

Figure 9-26 shows an example of configuring hunt lists for a distributed call processing deployment with an intercluster trunk. This example assumes that the hunt pilot call is first distributed within Cluster A. If the call is not answered, it is rerouted to Cluster B for call distribution using the Forward Hunt No Answer setting, which matches a route pattern. The route pattern, in turn, points to an intercluster trunk to Cluster B.

Figure 9-26 Call Coverage Between Clusters in a Distributed Call Processing Deployment



Tip

In distributed call processing deployments, load sharing of incoming hunt pilot calls can be managed using Cisco VoIP gateways and gatekeepers. In the event that the call is not answered within one cluster, it can overflow to another cluster for service. Calls can also be sent through gateways or trunks to IVR treatment. Tool Command Language (TCL) IVR applications can be implemented on Cisco IOS gateways.

Guidelines

When deploying call coverage functionality in a distributed call processing model, if calls are distributed across multiple clusters, then the route patterns should be properly configured to account for any digit transformations that are done on the outbound or inbound route group devices. If digit transformations are not done, then the configured route patterns and hunt pilot should be the same on all clusters, otherwise the calls will not be distributed appropriately.

Hunt Pilot Scalability

Cisco recommends using the following guidelines when deploying call coverage using top-down, circular, and longest-idle algorithms:

- The Unified CM cluster supports a maximum of 15,000 hunt list devices.
- The hunt list devices may be a combination of 1500 hunt lists with 10 IP phones in each hunt list, or a combination of 750 hunt lists with 20 IP phones in each hunt list.

Note

When using the broadcast algorithm for call coverage, the number of hunt list devices is limited by the number of busy hour call attempts (BHCA). Note that a BHCA of 10 on a hunt pilot pointing to a hunt list or hunt group containing 10 phones and using the broadcast algorithm is equivalent to 10 phones with a BHCA of 10.

- Cisco recommends having a maximum of 35 directory numbers in a single line group configured to send the calls simultaneously to all DNs. Additionally, the number of broadcast line groups depends on the BHCC. If there are multiple broadcast line groups in a Unified CM system, the number of maximum directory numbers in a line group must be less than 35. The number of busy hour call attempts (BHCA) for all the broadcast line groups should not exceed 35 calls set up per second.

Deploying Directory URI Dialing

Starting with Cisco Unified CM 9.0, provisioning and dialing of alphanumeric directory uniform resource identifiers (URIs) is supported by Unified CM. When handling SIP URIs, Unified CM call routing differentiates between numeric SIP URIs and alphanumeric SIP URIs (see Routing of SIP Requests in Unified CM, page 9-107). For endpoints registered to Unified CM, this adds dialing a directory URI as a new dialing habit. Also, caller ID based on directory URIs is a new concept for calls originating from devices registered to Unified CM.

Case Sensitivity

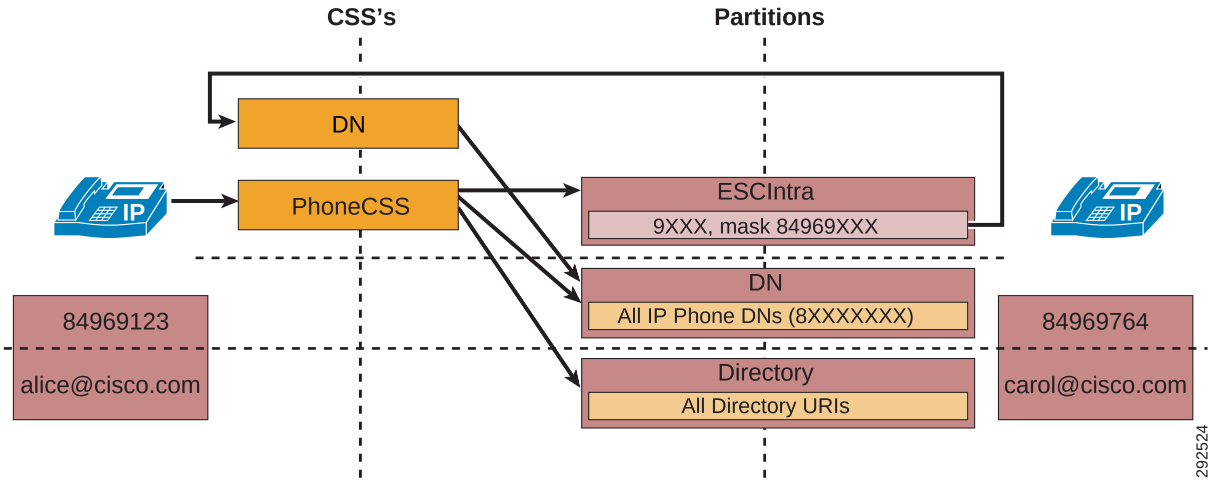
Per RFC 3261 (section 19.1.4, URI Comparison) comparison of the userinfo of SIP URIs has to be case-sensitive. According to this standardized behaviors, Alice@cisco.com and alice@cisco.com are not to be considered equivalent. When routing directory URIs, Unified CM respects this standard and looks for a case-sensitive full match of the user portion and a case-insensitive match of the host portion. To avoid confusion, Cisco highly recommends provisioning only directory URIs with all lowercase userinfo so that all directory URIs can reliably be dialed by entering all lowercase information.

Unified CM 9.1 and later releases can be configured to always use case-insensitive comparison of the user info portion of directory URIs. This can be achieved by configuring the enterprise parameter **URI Lookup Policy** accordingly. This setting applies to matching locally configured directory URIs and also to matching directory URIs for which an ILS lookup is done. The default setting of this enterprise parameter defines standard compliant case-sensitive matching of the user info portion of directory URIs.

Independent Call Routing

When a directory URI is routed by Unified CM, the dialed directory URI is first directly matched against the configured local directory URIs addressed by the calling device's calling search space. If a full case-sensitive match is found, then the call gets extended directly to the dialed destination. (See Figure 9-27.)

Figure 9-27 Independent Routing of Numeric Dialing and Directory URI Dialing



The example in Figure 9-27 assumes that directory numbers are configured using an 8-digit enterprise numbering plan consisting of an access code (8), a 3-digit site code (496), and a 4-digit extension (9123, 9764). To enable 4-digit intra-site dialing, a translation pattern 9XXX exists that will transform the called party number to the required 8-digit sequence by applying the called party transformation mask 84969XXX.

To reach the phone on the right in Figure 9-27, a user using the phone on the left might dial 9764 or carol@cisco.com. If the user dials 9764, the translation pattern will be matched, the called party will be transformed to 84969764 and, using calling search space DN, the call will be extended to directory number 84969764 in partition DN. On the other hand, if the directory URI carol is dialed, the call will be extended directly to directory number 84969764 because the directory URI carol@cisco.com in partition DirectoryURI will be found directly using calling search space PhoneCSS. This difference becomes important when the dialing normalization translation pattern 9XXX is used not only to transform the called party number but also to apply transformations to the calling party number. On translation pattern 9XXX, in a typical numeric dial plan we might also have a calling party number transformation to make sure that calling party information for this call is delivered in a globalized +E.164 form. To achieve this, typically the external phone number mask on the DNs would have been set and the calling party transformation on the translation pattern would have been configured as **use external phone number mask**. Under this condition the caller ID for a call from the left phone to the phone on the right would depend on the dialing habit used to place the call. If the call is placed by dialing 9764, the caller ID displayed on the right would be based on the correct +E.164 caller ID of the left phone, while a call dialed as carol@cisco.com would have a called ID based on 84969123 because the calling party transformations of the translation pattern would not be applied.

To avoid this, Cisco highly recommends moving the calling party normalization of calls originating from phones from translation patterns to the incoming calls calling party transformation calling search space on the phone or device pool.

The fact that calls dialed numerically or through a directory get routed differently has to be taken into account when enabling directory URI dialing as an additional dialing habit on Unified CM. The problem with inconsistent caller ID delivery described above can easily be addressed by using the incoming calls calling party transformation calling search space on the calling phone or calling phone's device pool (introduced in Unified CM 9.0). Translation patterns in an existing enterprise dial plan might exist for various other purposes that cannot be addressed as easily (for example, call blocking or destination-dependant caller ID masking). Cisco highly recommends checking an existing dial plan for the existence of intermediate routing steps such as translation patterns and understanding what is achieved by these intermediate routing steps, because for directory URI dialing it might not be possible to replicate this existing behavior created by these intermediate routing steps.

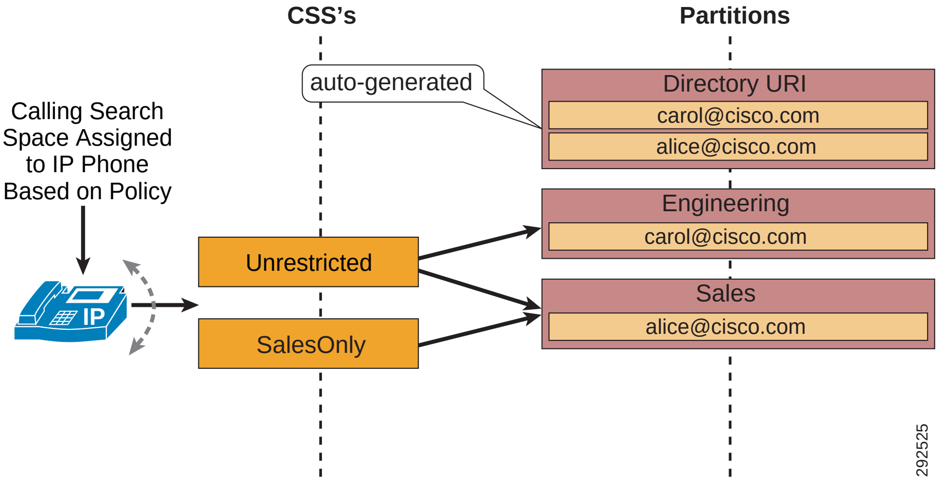
Building Class of Service for Directory URIs

A partition in Unified CM is used to group destinations belonging to the same class of reachability. For example, all directory numbers of the sales department might reside in one partition and all directory numbers of the engineering department might reside in a different partition if classes of service need to be implemented that differentiate between the ability to reach users of certain departments.

Manually configured directory URIs associated with directory numbers can be put in any partition; they do not have to reside in the same partition as the directory number they are associated with (although that definitely is an option).

All directory URIs that get associated to directory numbers automatically based on a directory URI and a primary extension configured for an end user are always created in a single partition **Directory URI**. If differentiated reachability is required, this Directory URI partition cannot be used. Instead the directory URIs of all users need to be provisioned in the correct partitions, which then can be used to create the required restricted classes of service. Figure 9-28 shows an example with two user groups, Engineering and Sales. Duplicate Directory URIs can exist in Unified CM as long as they do not reside in the same partition.

Figure 9-28 Building Class of Service for Directory URIs



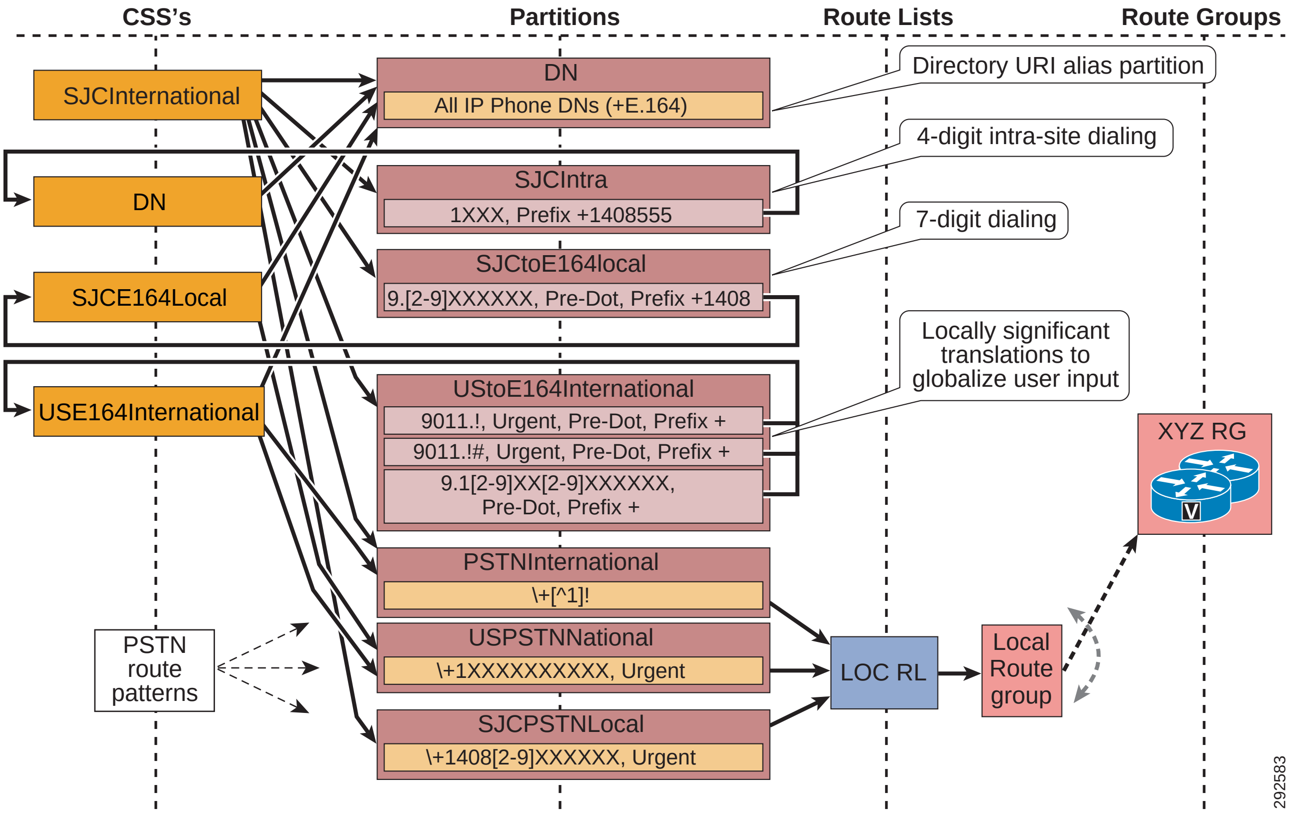
Aliasing the Directory URI Partition

In simple deployments where no differentiation between groups of users (as described in the section on Building Class of Service for Directory URIs, page 9-80) is required, the reachability of all automatically created directory URIs usually is equivalent to the reachability of end-user directory numbers. In a design based on Variable Length On-net Dialing (VLOD) with flat addressing, all end-user directory numbers reside in a single partition. In order to make all automatically created directory URIs reachable in an existing dial plan, the Directory URI partition must be added to the appropriate calling search spaces. Make sure to identify the correct places where to add the Directory URI partition. The calling search spaces that address the Directory URI partition need to be calling search spaces directly assigned to the calling line and/or device. Calling search spaces that are not directly assigned but get used only after hitting translation patterns are not a valid option because directory URI dialing will never match on translation patterns (see the section on Independent Call Routing, page 9-79).

Instead of changing a number of calling search spaces, Unified CM allows you to define an alias partition for the Directory URI partition. This is achieved by setting the enterprise parameter **Directory URI Alias Partition** to an existing partition that should be used as an alias. By selecting an existing partition as a directory URI alias partition, all calling search spaces that have access to the selected alias partition will automatically also have access to the Directory URI partition.

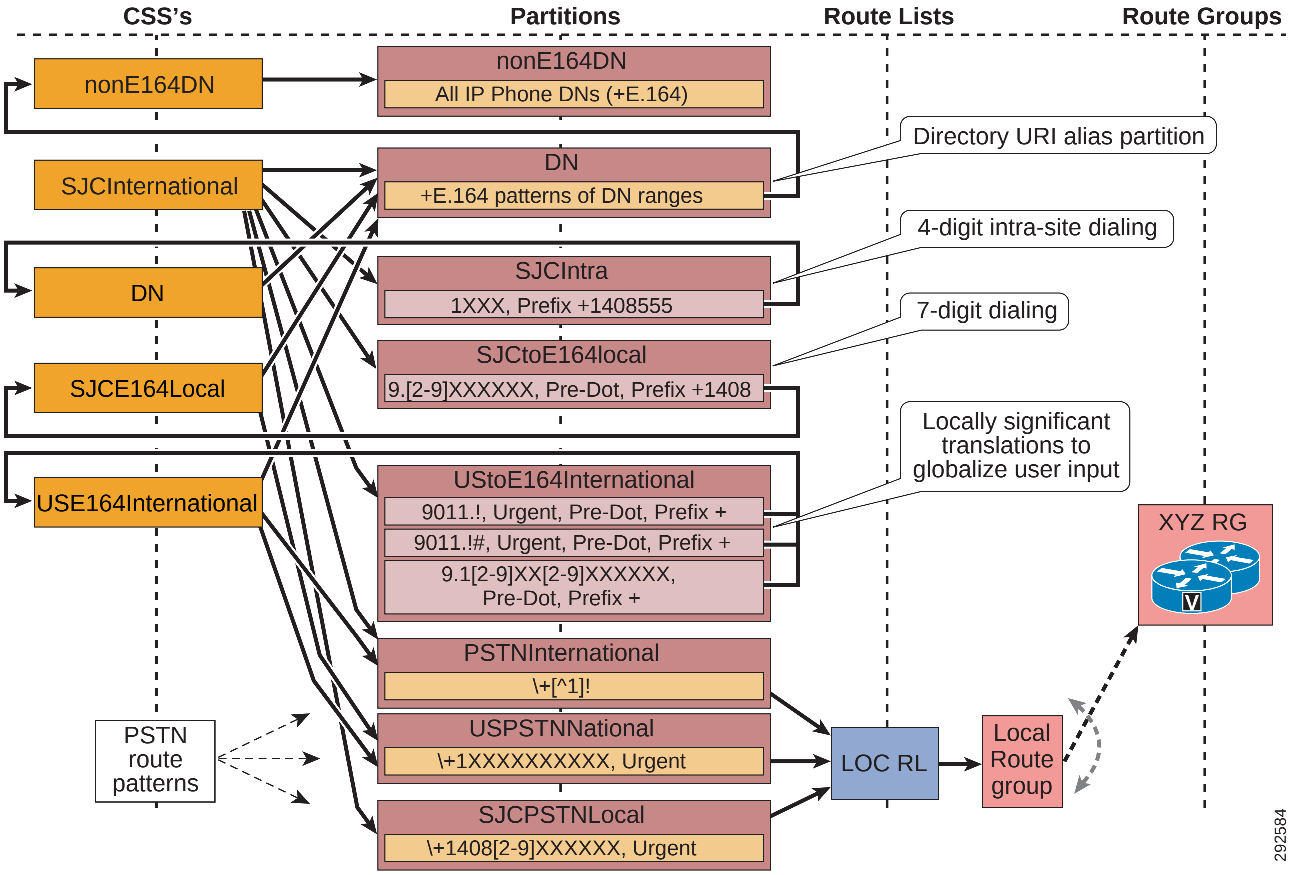
In Figure 9-29, partition DN would be a good choice for the directory URI alias partition because all devices having access to the DN partition that holds all directory numbers should also have access to the directory URIs associated with these directory numbers.

Figure 9-29 Class of Service "International" for +E.164 Dial Plans


CSS's
Partitions
Route Lists
Route Groups
SJCInternational
DN
SJCE164Local
USE164International
PSTN route patterns
DN
All IP Phone DNs (+E.164)
SJCIntra
1XXX, Prefix +1408555
SJCtoE164local
9.[2-9]XXXXXX, Pre-Dot, Prefix +1408
UStoE164International
9011.!, Urgent, Pre-Dot, Prefix +
9011.!#, Urgent, Pre-Dot, Prefix +
9.1[2-9]XX[2-9]XXXXXX, Pre-Dot, Prefix +
PSTNInternational
\+[^1]!
USPSTNNational
\+1XXXXXXXXXX, Urgent
SJCPSTNLocal
\+1408[2-9]XXXXXX, Urgent
Directory URI alias partition
4-digit intra-site dialing
7-digit dialing
Locally significant translations to globalize user input
XYZ RG
V
LOC RL
Local Route group
292583

Keep in mind that in Figure 9-30, partition nonE164DN is not a valid directory URI alias partition. Although partition nonE164DN holds all directory numbers, this partition is not accessible by any device directly, so dialing a directory URI would not work. In Figure 9-30, partition DN would be the better choice, although this partition in that example does not contain the actual directory numbers.

Figure 9-30 Indirection to Support Non +E.164 Directory Numbers


CSS's
Partitions
Route Lists
Route Groups
nonE164DN
SJCInternational
DN
SJCE164Local
USE164International
nonE164DN
All IP Phone DNs (+E.164)
DN
+E.164 patterns of DN ranges
SJCIntra
1XXX, Prefix +1408555
SJCtoE164local
9.[2-9]XXXXXX, Pre-Dot, Prefix +1408
UStoE164International
9011.!, Urgent, Pre-Dot, Prefix +
9011.!#, Urgent, Pre-Dot, Prefix +
9.1[2-9]XX[2-9]XXXXXX, Pre-Dot, Prefix +
PSTNInternational
\+[^1]!
USPSTNNational
\+1XXXXXXXXXX, Urgent
SJCPSTNLocal
\+1408[2-9]XXXXXX, Urgent
Directory URI alias partition
4-digit intra-site dialing
7-digit dialing
Locally significant translations to globalize user input
PSTN route patterns
LOC RL
Local Route group
XYZ RG
V
292584

Blended Identity

In Unified CM the primary identity of any line appearance always is the numeric (possibly with a leading +) directory number. Directory URIs are always configured as aliases of these numeric directory numbers. As soon as a directory URI is associated with a directory number, two different identities exist: the numeric directory number and the primary associated directory URI. The combination of these two pieces is called *blended identity*.

For every call involving a directory number with an associated directory URI, Unified CM has to decide which piece of the blended identity to use and to present to the endpoints or trunks involved in the call. This decision depends on the capabilities of the endpoints or trunks involved. Endpoints registering with Unified CM indicate during registration whether they are capable of handling directory URI-based caller IDs.

For endpoints indicating support of directory URIs during registration, Unified CM will always try to send directory URI-based caller IDs. Even if the call originated from a device not supporting directory URI dialing and caller ID, Unified CM can still use the primary directory URI of the calling directory number as a directory URI-based caller ID.

For SIP trunks, the format of calling and connected party information sent to that trunk is defined by the new **Calling and Connected Party Info Format** setting on the trunk. The default of this setting is to always send only numeric identification. This is to ensure backward compatibility with the behavior of Unified CM prior to release 9.0. Alternatively, the trunk can be configured so that identification is always sent only as a directory URI (if available), and the third option is to always send both pieces of the blended identity (directory URI and numeric identity). Cisco recommends using this third option (directory URI and numeric identity) when interconnecting multiple Unified CM clusters. This setting makes sure that the remote cluster always gets the full blended identity and can then decide which piece to present to the remote endpoint based on the capabilities of that endpoint.

Dial Plan Elements

This section provides design and configuration guidelines for the following dial plan elements within a Cisco Unified Communications system:

- User Input on SCCP Phones, page 9-86
- User Input on Type-A SIP Phones, page 9-87
- User Input on Type-B SIP Phones, page 9-89
- SIP Dial Rules, page 9-90
- Call Routing in Unified CM, page 9-93
- Calling Privileges in Unified CM, page 9-110
- Translation Patterns, page 9-116
- Automated Alternate Routing, page 9-117
- Device Mobility, page 9-122
- Extension Mobility, page 9-124
- Immediate Divert (iDivert), page 9-130
- Hunt Lists and Line Groups, page 9-131
- Time-of-Day Routing, page 9-134
- Call Routing in Cisco IOS with H.323 Dial Peers, page 9-138
- Call Routing in Cisco IOS with a Gatekeeper, page 9-141
- Calling Privileges in Cisco IOS with H.323 Dial Peers, page 9-150
- Digit Manipulation in Cisco IOS with H.323 Dial Peers, page 9-152

User Interface on IP Phones

Different types of IP telephones accept keypad input and present visual information in different ways. For purposes of this chapter only, we define the following phone types:

- Type-A phones — Include the Cisco Unified IP Phone 7905, 7912, 7940, and 7960.
- Type-B phones — Include the Cisco Unified IP Phone 6901, 6911, 6921, 6941, 6945, 6961, 7906, 7911, 7921, 7925, 7931, 7941, 7942, 7945, 7961, 7962, 7965, 7970, 7971, 7975, 8961, 9951, and 9971.

Calling Party Transformations on IP Phones

Calling Party Transformation Patterns allow the system to adapt the calling party numbers to different formats. The most typical use is to adapt from globalized to localized calling party numbers and vice versa.

The transformation pattern consists of a numerical representation of the calling party number to be matched. The syntax used is the same as that of other patterns such as route patterns, transformation patterns, directory numbers, and so forth.

The transformation operators include discard digit instructions (for example, pre-dot), a calling party transformation mask, prefix digits, and control over the calling party presentation (either Default, Allowed, or Restricted). Calling party transformation patterns can be configured to use the calling party's external phone number mask as the calling party number.

Partitions and calling search spaces control which calling party transformation patterns are applied to which phones. Phones can use either the device pool's calling party transformation calling search space (CSS) or the device's own calling party transformation CSS, in reverse order of precedence. Calls sent to phones are not processed through called party transformation patterns.

On IP phones, calling party transformations can be configured for calls originating from the phone and for calls terminated on the phone:

- For calls originating from phones where the configured directory numbers are not in a globalized (+E.164) form, the inbound call’s calling party transformation CSS can be used to define the appropriate globalization. Beginning with Unified CM 9.1, this CSS can be found on the phone configuration page in the Number Presentation Transformation section or in the Phone Settings section on the device pool configuration page under **Caller ID For Calls From This Phone**.
- For calls terminated on the phones, the outbound call’s calling party transformation CSS can be used to define the localization scheme to be applied to calling party numbers. Beginning with Unified CM 9.1, this CSS can be found on the phone configuration page in the Number Presentation Transformation section under **Remote Number**.

For phones, outbound or remote number calling party transformations affect the number displayed while the phone is ringing.

For Type-B phones, the corresponding entry in the missed and received calls directories holds both the transformed number and also the original pre-transformation number. The transformed number is displayed in the directories´ list, but the number used for callback is the pre-transformation number.

Starting with Unified CM release 9.1, the outbound call’s calling party transformation CSS (also referred to as localization or remote number calling party transformation CSS) can also be used to localize remote connected party information. To enable this, the advanced service parameter **Apply Transformations On Remote Number** must be enabled.

Being able to provide localized connected party information to phones enables consistent remote party information display on IP phones even if mid-call features are invoked.

Support for + Dialing on the Phones

On Type-A phones, it is not possible to dial a + sign using the keypad. On Type-B phones it is possible to dial a + sign by pressing and holding either the 0 key (Cisco Unified IP Phones 7921 and 7925) or the * key (all other phone models). On Cisco Unified Personal Communicator endpoints, the + sign may be typed in by the user using the computer's keyboard or entered as part of the input string when using a click-to-dial function of the endpoint.

On Type-A phones, there is no support to display the + sign.

On Type-B phones and on Cisco Unified Personal Communicator, incoming calls can present a calling party number including + as part of the number. When a call is offered to a phone, the number shown on the ringing phone is processed by any configured calling party number transformation patterns. The missed and received calls directories hold both the original pre-transformation number and the transformed number. The number displayed in the list will be the transformed number, and the pre-transformation number will be visible only when looking at the details of an entry. The number dialed from the directory is the original pre-transformation number, allowing the one-touch dialing of previously received calls featuring the + sign as part of the calling number.

Example 9-1 Calling Party Number with + Dialing

A Type-B phone in New York receives a call from +1 408 526 4000. Calling party transformation patterns are placed in the calling party transformation CSS in the phone's device pool. One of the patterns is configured as: \+1.!, strip pre-dot.

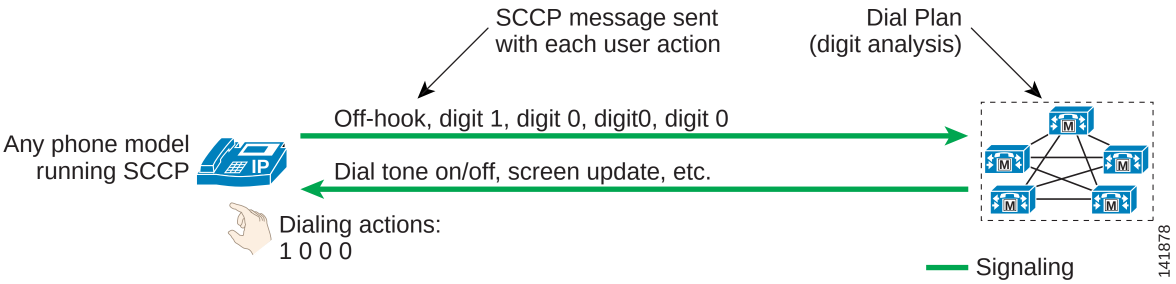
As the call rings in, the called phone displays an incoming calling party number of 4085264000. After the call is answered and released, the received-calls directory displays the last call as 408 526 4000, but the number called when the user initiated the callback from the directory entry is +1 408 526 4000.

User Input on SCCP Phones

IP phones using SCCP report every single user input event to Unified CM immediately. For instance, as soon as the user goes off-hook, a signaling message is sent from the phone to the Unified CM server with which it is registered. The phone can be considered to be a terminal, where all decisions resulting from the user input are made by the Unified CM server's configured dial plan.

As other user events are detected by the phone, they are relayed to Unified CM individually. A user who goes off-hook and then dials 1000 would trigger five individual signaling events from the phone to Unified CM. All the resulting feedback provided to the user, such as screen messages, playing dial tone, secondary dial tone, ring back, reorder, and so forth, are commands issued by Unified CM to the phone in response to the dial plan configuration. (See Figure 9-31.)

Figure 9-31 User Input and Feedback for SCCP Phones



It is neither required nor possible to configure dial plan information on IP phones running SCCP. All dial plan functionality is contained in the Unified CM cluster, including the recognition of dialing patterns as user input is collected.

If the user dials a pattern that is denied by Unified CM, reorder tone is played to the user as soon as that pattern becomes the best match in Unified CM's digit analysis. For instance, if all calls to the pay-per-minute Numbering Plan Area (or area code) 976 are denied, reorder tone would be sent to the user’s phone as soon as the user dials 91976.

User Input on Type-A SIP Phones

Type-A phones differ somewhat from Type-B phones in their behavior, and Type-A phones do not offer support for Key Press Markup Language (KPML) as do Type-B phones. (See User Input on Type-B SIP Phones, page 9-89.)

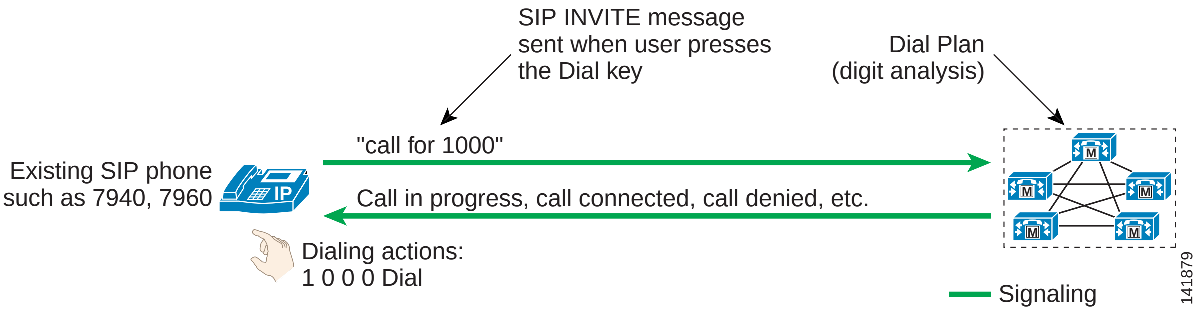
Type-A IP phones using SIP offer two distinct modes of operation:

- No SIP Dial Rules Configured on the Phone, page 9-87
- SIP Dial Rules Configured on the Phone, page 9-88

No SIP Dial Rules Configured on the Phone

Figure 9-32 illustrates the behavior of a SIP Type-A phone with no dial plan rules configured on the phone. In this mode of operation, the phone accumulates all user input events until the user presses either the # key or the Dial softkey. This function is similar to the "send" button used on many mobile phones. For example, a user making a call to extension 1000 would have to press 1, 0, 0, and 0 followed by the Dial softkey or the # key. The phone would then send a SIP INVITE message to Unified CM to indicate that a call to extension 1000 is requested. As the call reaches Unified CM, it is subjected to the dial plan configuration for this phone, including all the class-of-service and call-routing logic implemented in Unified CM's dial plan.

Figure 9-32 User Input and Feedback for Type-A SIP Phones with No Dial Rules Configured



If the user dials digits but then does not press the Dial softkey or the # key, the phone will wait for inter-digit timeout (15 seconds by default) before sending a SIP INVITE message to Unified CM. For the example in Figure 9-32, dialing 1, 0, 0, 0 and waiting for inter-digit timeout would result in the phone placing a call to extension 1000 after ten seconds.

Note If the user presses the Redial softkey, the action is immediate; the user does not have to press the Dial key or wait for inter-digit timeout.

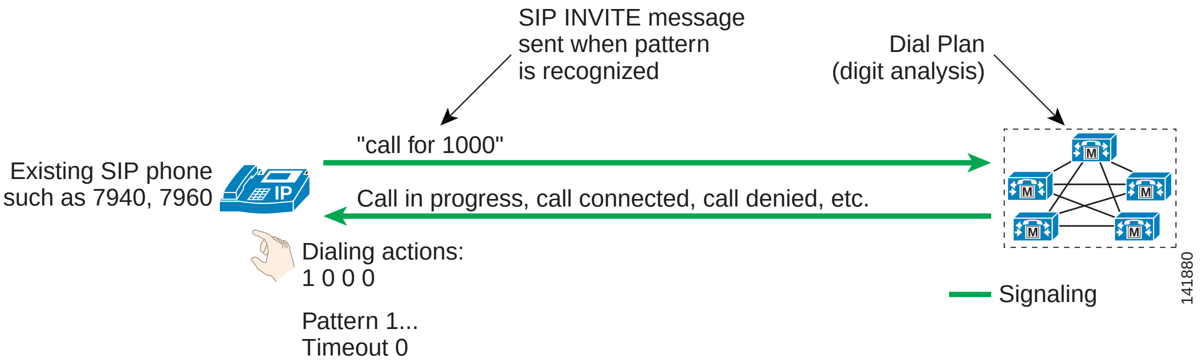
If the user dials a pattern that is denied by Unified CM, the user must enter the entire pattern and press the Dial key, and the INVITE message must be sent to Unified CM, before any indication that the call is rejected (reorder tone) is sent to the caller. For instance, if all calls to NPA 976 are denied, the user would have to dial 919765551234 and press Dial before the reorder tone would be played.

SIP Dial Rules Configured on the Phone

SIP dial rules enable the phone to recognize patterns dialed by users. Once the recognition has occurred, the sending of the SIP INVITE message to Unified CM is automated and does not require the user to press the Dial key or wait for the inter-digit timeout. (For more details, see SIP Dial Rules, page 9-90.)

For example, if a branch location of the enterprise requires that calls between phones within the same branch be dialed as four-digit extensions, the phone could be configured to recognize the four-digit patterns so that the user is not required to press the Dial key or wait for the inter-digit timeout. (See Figure 9-33.)

Figure 9-33 User Input and Feedback for Type-A SIP Phones with Dial Rules Configured



In Figure 9-33, the phone is configured to recognize all four-digit patterns beginning with 1 and has an associated timeout value of 0. All user input actions matching the pattern will trigger the sending of the SIP INVITE message to Unified CM immediately, without requiring the user to press the Dial key.

Type-A phones using SIP dial rules offer a way to dial patterns not explicitly configured on the phone. If a dialed pattern does not match a SIP dial rule, the user can press the Dial key or wait for inter-digit timeout.

If a particular pattern is recognized by the phone but blocked by Unified CM, the user must dial the entire dial string before receiving an indication that the call is rejected by the system. For instance, if a SIP dial rule is configured on the phone to recognize calls dialed in the form 919765551234 but such calls are blocked by the Unified CM dial plan, the user will receive reorder tone at the end of dialing (after pressing the final 4 key).

User Input on Type-B SIP Phones

Type-B phones differ somewhat from Type-A phones in their behavior, and Type-B phones offer support for Key Press Markup Language (KPML) but Type-A phones do not. (See User Input on Type-A SIP Phones, page 9-87.)

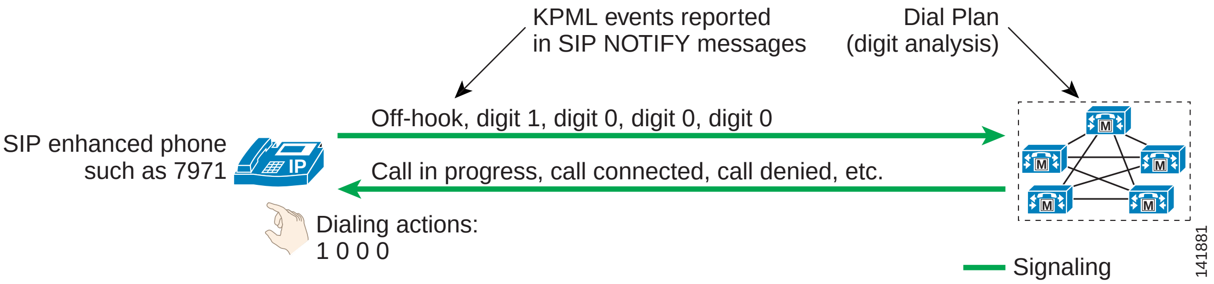
Type-B IP phones running SIP offer two distinct modes of operation:

- No SIP Dial Rules Configured on the Phone, page 9-89
- SIP Dial Rules Configured on the Phone, page 9-90

No SIP Dial Rules Configured on the Phone

Type-B IP telephones offer functionality based on the Key Press Markup Language (KPML) to report user key presses. Each one of the user input events will generate its own KPML-based message to Unified CM. From the standpoint of relaying each user action immediately to Unified CM, this mode of operation is very similar to that of phones running SCCP. (See Figure 9-34.)

Figure 9-34 User Input and Feedback for Type-B SIP Phones with No Dial Rules Configured



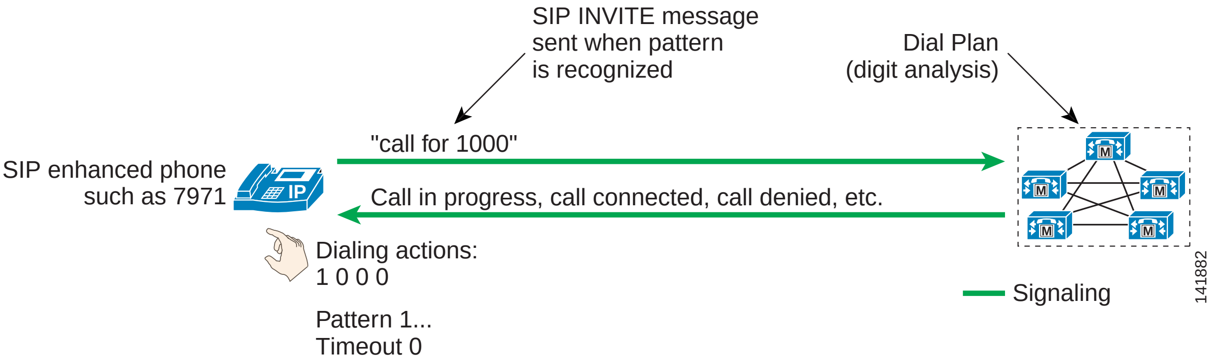
Every user key press triggers a SIP NOTIFY message to Unified CM to report a KPML event corresponding to the key pressed by the user. This messaging enables Unified CM's digit analysis to recognize partial patterns as they are composed by the user and to provide the appropriate feedback, such as immediate reorder tone if an invalid number is being dialed.

In contrast to Type-A IP phones running SIP without dial rules, Type-B SIP phones have no Dial key to indicate the end of user input. In Figure 9-34, a user dialing 1000 would be provided call progress indication (either ringback tone or reorder tone) after dialing the last 0 and without having to press the Dial key. This behavior is consistent with the user interface on phones running the SCCP protocol.

SIP Dial Rules Configured on the Phone

Type-B IP phones can be configured with SIP dial rules so that dialed pattern recognition is accomplished by the phone. (See Figure 9-35.)

Figure 9-35 User Input and Feedback for Type-B SIP Phones with Dial Rules Configured



In Figure 9-35, the phone is configured to recognize all four-digit patterns beginning with 1, and it has an associated timeout value of 0. All user input actions matching these criteria will trigger the sending of a SIP INVITE message to Unified CM.

Note As soon as SIP dial rules are implemented on Type-B IP phones, KPML-based dialing is disabled. If a user dials a string of digits that do not match a SIP dial rule, none of the individual digit events will be relayed to Unified CM. Instead, the entire dialed string will be sent en-bloc to Unified CM once the dialing is complete (that is, once inter-digit timeout has occurred).

Type-B phones using SIP dial rules offer only one way to dial patterns not explicitly configured on the phone. If a dialed pattern does not match a SIP dial rule, the user has to wait for inter-digit timeout before the SIP NOTIFY message is sent to Unified CM. Unlike Type-A IP phones, Type-B IP phones do not have a Dial key to indicate the end of dialing, except when on-hook dialing is used. In the latter case, the user can press the “dial” key at any time to trigger the sending of all dialed digits to Unified CM.

Note When a Type-B phone registers with the SRST router, the configured SIP dial rules are disabled.

If a particular pattern is recognized by the phone but blocked by Unified CM, the user must dial the entire dial string before receiving an indication that the call is rejected by the system. For instance, if a SIP dial rule is configured on the phone to recognize calls dialed in the form 919765551234 but such calls are blocked by the Unified CM dial plan, the user will receive reorder tone at the end of dialing (after pressing the 4 key).

SIP Dial Rules

Cisco Unified CM offers SIP dial rule functionality to allow phones to perform pattern recognition as user input is collected. For example, a phone can be configured to recognize the well established pattern 911 and to send a message to Unified CM to initiate an emergency call immediately, while at the same time allowing the user to enter patterns of variable length for international numbers.

It is important to note that pattern recognition configuration on the phone through the use of SIP dial rules does not supersede the Class of Service and Route Plan configurations of Unified CM. A phone might very well be configured to recognize long-distance patterns while Unified CM is configured to block such calls because the phone is assigned a class of service allowing only local calls.

There are two types of SIP dial rules, based on the phone model on which they will be deployed:

- 7905_7912 (Used for Cisco Unified IP Phones 7905 and 7912)
- 7940_7960_OTHER (Used for all other IP phone models))

There are four basic Dial Parameters that can be used as part of a dial rule:

- Pattern

  This parameter is the actual numerical representation of the pattern. It can contain digits, wildcards, or instructions to play secondary dial tone. The following table provides a list of values and their effect for the two types of dial rules.

	Dial Rule Type	
Pattern	**7905_7912**	**7940_7960_OTHER**
Digits 0 through 9	Corresponding digit	Corresponding digit
.	Matches any digit (0 through 9)	Matches any character (0 though 9, *, #)
-	Indication that more digits can be entered. Must be at the end of an individual rule.	n/a
#	Input termination key. Place the > character in the dial rule to indicate the character position after which the # key will be recognized as input termination. For instance, in 9>#..., the # character would be recognized any time after 9 has been pressed.	n/a
t*n*	Indicates a timeout value of *n* seconds. For example, 1…t3 would match 1000 and trigger the sending of an invite to Unified CM after 3 seconds.	n/a
r*n*	Repeats the last character *n* times. For example, 1.r3 is equivalent to 1….	n/a
S	When a pattern contains the modifier S, all other dial rules after this pattern are ignored. S effectively causes rule matching to cease.	n/a

Pattern	Dial Rule Type	
	7905_7912	7940_7960_OTHER
*	Input termination key. Place the > character in the dial rule to indicate the character position after which the * key will be recognized as input termination.	Matches one or more characters. For instance, pattern 1* would match 10, 112, 123456, and so forth.
,	n/a	Cause the phone to play secondary dial tone. For instance, 8,…. would cause the user to hear secondary dial tone after 8 is pressed.

- IButton

  This parameter specifies the button to which the dial pattern applies. If the user is initiating a call on line button 1, only the dial patterns specified for Button 1 apply. If this optional parameter is not configured, the dial pattern applies to all lines on the phone. This parameter applies only to the Cisco SIP IP Phones 7940, 7941, 7942, 7945, 7960, 7961, 7962, 7965, 7970, 7971, and 7975. The button number corresponds to the order of the buttons on the side of the screen, from top to bottom, with 1 being on top button.

- Timeout

  This parameter specifies the time, in seconds, before the system times out and dials the number as entered by the user. To have the number dialed immediately, specify 0. This parameter is available only for 7940_7960_OTHER dial rules. If this parameter is omitted, the phone's default inter-digit timeout value is used (default of 10 seconds).

- User

  This parameter represents the tag that automatically gets added to the dialed number. Valid values include **IP** (when SIP Call Agents other than Unified CM are deployed) and **Phone**. This parameter is available only for 7940_7960_OTHER dial rules. This parameter is optional, and it should be omitted in deployments where Unified CM is the only call agent. Keep in mind that a user=phone tag in a SIP request sent to Unified CM starting with release 9.0 will force Unified CM to treat the SIP URI as a numeric URI. A SIP URI of the form alice@cisco.com;user=phone will never be routed successfully because the user=phone tag forces numeric treatment and alice will not match any numeric pattern provisioned in Unified CM.

Note The Cisco Unified IP Phone 7905 and 7912 choose patterns in the order in which they have been created in the SIP dial rules, whereas all the other phone models choose the pattern offering the longest match. The following table shows which pattern would be chosen if a user dialed 95551212.

SIP Dial Rules	7905_7912	7940_7960_OTHER
…….. 9…….	Choose first matching pattern: ……..	Chooses longest matching pattern: 9…….

Call Routing in Unified CM

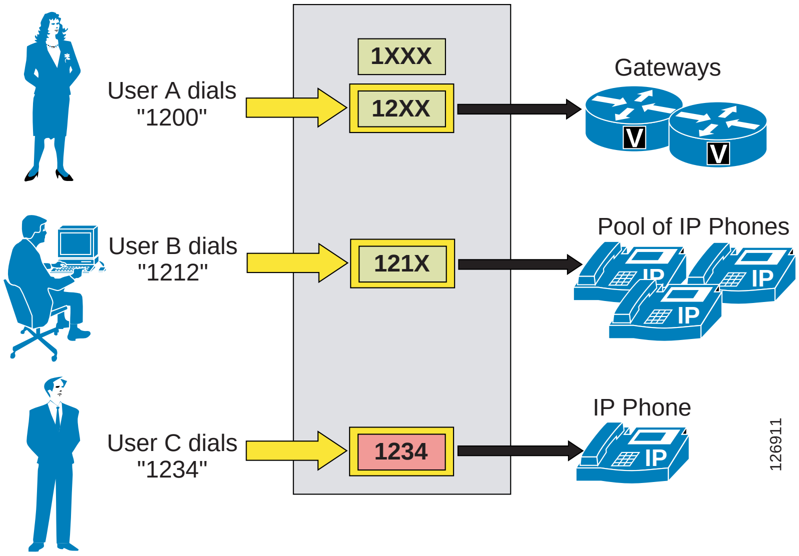
All numeric dialing destinations and directory URIs configured in Unified CM are added to its internal call routing table as patterns. These destinations include IP phone lines, voicemail ports, route patterns, translation patterns, and CTI route points. Unified CM uses two distinct routing tables for numeric dialing destinations and directory URIs.

When a directory URI is dialed, Unified CM uses full-match logic to find a case-sensitive match among the configured directory URIs in the directory URI routing table. When a number is dialed, Unified CM uses closest-match logic to select which pattern to match from among all the patterns in its numeric call routing table. In practice, when multiple potentially matching numeric patterns are present, the destination pattern is chosen based on the following criteria:

- It matches the dialed string, and
- Among all the potentially matching patterns, it matches the fewest strings other than the dialed string.

For example, consider the case shown in Figure 9-36, where the call routing table includes the patterns 1XXX, 12XX, and 1234.

Figure 9-36 Unified CM Call Routing Logic Example



When user A dials the string 1200, Unified CM compares it with the patterns in its call routing table. In this case, there are two potentially matching patterns, 1XXX and 12XX. Both of them match the dialed string, but 1XXX matches a total of 1000 strings (from 1000 to 1999) while 12XX matches only 100 strings (from 1200 to 1299). Therefore, 12XX is selected as the destination of this call.

When user B dials the string 1212, there are three potentially matching patterns, 1XXX, 12XX and 121X. As mentioned above, 1XXX matches 1000 strings and 12XX matches 100 strings. However, 121X matches only 10 strings; therefore it is selected as the destination of the call.

When user C dials the string 1234, there are three potentially matching patterns, 1XXX, 12XX, and 1234. As mentioned above, 1XXX matches 1000 strings and 12XX matches 100 strings. However, 1234 matches only a single string (the dialed string); therefore it is selected as the destination of this call.

Note Whenever a directory number (DN) is configured in Cisco Unified CM, it is placed in the call routing table, regardless of whether the respective device (for example, an IP phone) is registered or not. An implication of this behavior is that it is not possible to rely on secondary, identical patterns to provide failover capabilities to applications when the application (and hence the primary pattern) is unregistered. Because the primary pattern is permanently in the call routing table, the secondary pattern will never be matched.

Support for + Sign in Patterns

The + sign can be used in all patterns within Unified CM, including route patterns, translations patterns, and directory numbers. To use + in its literal sense, precede it with the escape character \ to differentiate it from the regular expression operator +, which means one or more instances of the preceding character. For example:

- \+14085264000 means +14085264000
- 2+ means 2, or 22, or 222, and so forth

Directory URIs

All endpoints registered with Unified CM are provisioned with one or more numeric (possibly including a leading +) directory numbers. Starting with Cisco Unified CM 9.0, up to five directory URIs can be associated with each directory number. This association can be created by explicitly associating directory URIs to directory numbers. If a directory URI is configured for an end user, this directory URI will be automatically associated with the primary extension of that end user as soon as the primary extension gets defined for that end user. All automatically associated directory URIs are created in the partition Directory URI, while manually configured directory URIs can be in any partition.

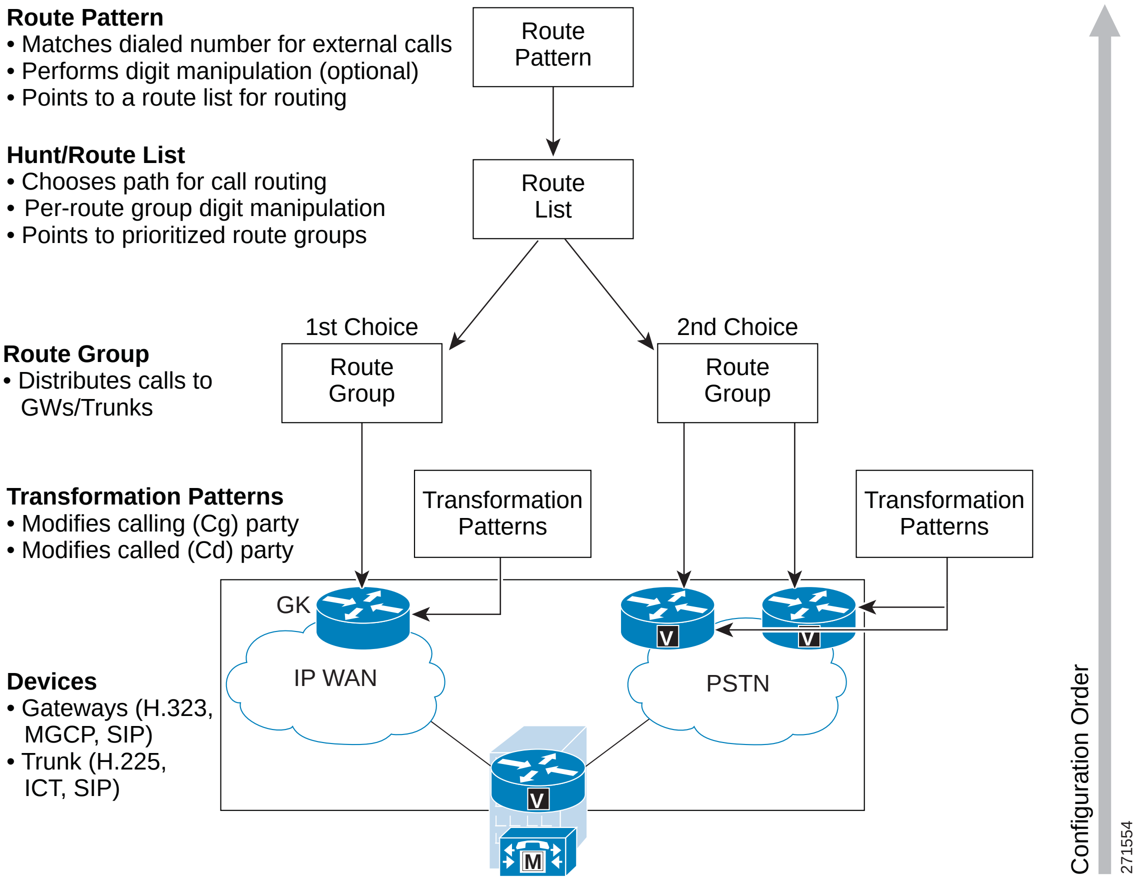
Exactly one of the directory URIs associated with a directory number has to be marked as the primary directory URI of that directory number. If a user directory URI gets associated automatically with the primary extension of that user, then this directory URI will also automatically be the primary directory URI for that directory number. If no directory URI is associated automatically, then one of the configured directory URIs has to be selected as the primary directory URI. The purpose of the primary directory URI is that this directory URI will be used as the calling identity directory URI for calls originating from the respective directory number.

The possible association of directory URIs with any directory number allows callers to reach any directory number by dialing the associated directory URI. The called directory number can be on any device registered to Unified CM using any protocol. Similarly, Unified CM can deliver a directory URI caller ID for any call from any directory number as longs as a directory URI is associated with the calling directory number.

External Routes in Unified CM

Unified CM automatically "knows" how to route calls to internal destinations within the same cluster. For external destinations such as PSTN gateways, H.323 gatekeepers, or other Unified CM clusters, you have to use the external route construct (described in the following sections) to configure explicit routing. This construct is based upon a three-tiered architecture that allows for multiple layers of call routing as well as digit manipulation. Unified CM searches for a configured route pattern that matches the external dialed string and uses it to select a corresponding route list, which is a prioritized list of the available paths for the call. These paths are known as route groups and are very similar to trunk groups in traditional PBX terminology. Figure 9-37 depicts the three-tiered architecture of the Unified CM external route construct.

Figure 9-37 External Route Pattern Architecture


Route Pattern
• Matches dialed number for external calls
• Performs digit manipulation (optional)
• Points to a route list for routing
Route Pattern
Hunt/Route List
• Chooses path for call routing
• Per-route group digit manipulation
• Points to prioritized route groups
Route List
1st Choice
2nd Choice
Route Group
• Distributes calls to GWs/Trunks
Route Group
Route Group
Transformation Patterns
• Modifies calling (Cg) party
• Modifies called (Cd) party
Transformation Patterns
Transformation Patterns
GK
IP WAN
PSTN
V
V
V
M
Devices
• Gateways (H.323, MGCP, SIP)
• Trunk (H.225, ICT, SIP)
Configuration Order
271554

The following sections describe the individual elements of the external route construct in Unified CM:

- Route Patterns, page 9-95
- Route Lists, page 9-99
- Route Groups, page 9-99
- Route Group Devices, page 9-102

Route Patterns

Route patterns are strings of digits and wildcards, such as 9.[2-9]XXXXXX, configured in Unified CM to route calls to external entities. The route pattern can point directly to a gateway for routing calls or point to a route list, which in turn points to a route group and finally to a gateway.

Cisco strongly recommends that you use the complete route pattern, route list, and route group construct because it provides the greatest flexibility for call routing, digit manipulation, route redundancy, and future dial plan growth.

The @ Wildcard

- The @ wildcard is a special macro function that expands into a series of patterns representing the entire national numbering plan for a certain country. For example, configuring a single unfiltered route pattern such as 9.@ with the North American Numbering Plan really adds 166 individual route patterns to the Unified CM internal dial plan database.
- It is possible to configure Unified CM to accept other national numbering plans. Once this is done, the @ wildcard can be used for different numbering plans even within the same Unified CM cluster, depending on the value selected in the Numbering Plan field on the Route Pattern configuration page. For more information, please refer to the *Cisco Unified CallManager Dial Plan Deployment Guide*, available at

  http://www.cisco.com/en/US/products/sw/voicesw/ps5629/prod_maintenance_guides_list.html
- The @ wildcard can be practical in several small and medium deployments, but it can become harder to manage and troubleshoot in large deployments because adopting the @ wildcard forces the administrator to use route filters to block certain patterns (see Route Filters, page 9-96).

Route Filters

- Use route filters only with the @ route pattern to reduce the number of route patterns created by the @ wildcard. A route filter applied to a pattern not containing the @ wildcard has no effect on the resulting dial plan.
- The logical expression you enter with the route filter can be up to 1024 characters, excluding the NOT-SELECTED fields.
- As you increase the number of logical clauses in a route filter, the refresh time of the configuration page also increases and can become unacceptably long.
- For large-scale deployments, use explicit route patterns rather than the @ wildcard and route filters. This practice also facilitates management and troubleshooting because all patterns configured in Unified CM are easily visible from the Route Pattern configuration page.

International and Variable-Length Route Patterns

- International destinations are usually configured using the ! wildcard, which represents any quantity of digits. For example, in North America the route pattern 9.011! is typically configured for international calls. In most European countries, the same result is accomplished with the 0.00! route pattern.
- The ! wildcard is also used for deployments in countries where the dialed numbers can be of varying lengths. In such cases, Unified CM does not know when the dialing is complete and will wait for 15 seconds (by default) before sending the call. You can reduce this delay in any of the following ways:
  - Reduce the T302 timer (Service Parameter TimerT302_msec) to indicate end of dialing, but do not set it lower than 4 seconds to prevent premature transmission of the call before the user is finished dialing.
  - Configure a second route pattern followed by the # wildcard (for example, 9.011!# for North America or 0.00!# for Europe), and instruct the users to dial # to indicate end of dialing. This action is analogous to hitting the "send" button on a cell phone.

Overlap Sending and Overlap Receiving

In countries whose national numbering plan is not easily defined with static route patterns, you can configure Unified CM for overlap sending and overlap receiving.

Overlap sending means that Unified CM keeps collecting digits as they are dialed by the end users, and passes them on to the PSTN as they are dialed. To enable overlap sending, check the Allow Overlap Sending box on the Route Pattern Configuration page. (In some early Unified CM releases, overlap sending is enabled by setting the SendingCompleteIndicator service parameter to False.) The route pattern needs only to include the PSTN access code (for example, "9." in North America or "0." in many European countries).

Overlap receiving means that Unified CM receives the dialed digits one-by-one from a PRI PSTN gateway, and it then waits for completion of the dialed string before attempting to route the call to an internal destination. To enable overlap receiving, set the OverlapReceivingFlagForPRI service parameter to True. (In some early Unified CM releases, the parameter name was OverlapReceivingForPriFlag.)

Digit Manipulation in Route Patterns

- Digit manipulations configured on a route pattern affect the calling and called party number, no matter what route group the call eventually takes. Digit manipulations configured in the route list's view of its member route groups have a route-specific effect: only the transformations configured on the route group used to place the call will be performed.
- Digit manipulation in the route list's view of its member route group overrides any digit manipulation done in the route pattern.
- The calling and called party numbers resulting from the digit transformations configured in the route pattern and/or route lists are then processed by any Transformation Patterns configured for the devices contained in the chosen Route Group.
- If you configure digit manipulation in the route pattern, the Call Detail Record (CDR) records the dialed number after the digit manipulation has occurred. If you configure digit manipulation only in the route group, the CDR records the actual dialed number prior to the digit manipulation.
- Similarly, if you configure digit manipulation in the route pattern, the IP phone display of the calling party will show the manipulated number. If you configure digit manipulation only in the route group, the manipulations will be transparent to the end user.

Calling Line ID

- The calling line ID presentation can be enabled or disabled on the gateway and also can be manipulated in the route pattern, based on site requirements.
- If you select the option Use Calling Party's External Phone Number Mask, then the external call uses the calling line ID specified for the IP phone placing the call. If you do not select this option, then the mask placed in the Calling Party Transform Mask field is used to generate the calling party ID.

Urgent Priority

- The Urgent Priority checkbox is often used to force immediate routing of certain calls as soon as a match is detected, without waiting for the T302 timer to expire. For example, in North America, if the patterns 9.911 and 9.[2-9]XXXXXX are configured and a user dials 9911, Unified CM has to wait for the T302 timer before routing the call because further digits may cause the 9.[2-9]XXXXXX to match. If Urgent Priority is enabled for the 9.911 route pattern, Unified CM makes its routing decision as soon as the user has finished dialing 9911, without waiting for the T302 timer.

- It is important to note that the Urgent Priority checkbox forces the T302 timer to expire as soon as the configured pattern is the best match for the dialed number. This does not mean that the urgent pattern has a higher priority than other patterns; the closest-match logic described in the section on Call Routing in Unified CM, page 9-93, still applies.

  For example, assume the route pattern 1XX is configured as urgent and the pattern 12! is configured as a regular route pattern. If a user dials 123, Unified CM will not make its routing decision as soon as it receives the third digit because even though 1XX is an urgent pattern, it is not the best match (10 total patterns matched by 12! versus 100 patterns matched by 1XX). Unified CM will have to wait for inter-digit timeout before routing the call because the pattern 12! allows for more digits to be input by the user.

  Consider another example, where pattern 12[2-5] is marked as urgent and 12! is configured as a regular pattern. If the user dials 123, the pattern 12[2-5] is the best match (4 total patterns matched by 12[2-5] versus 10 patterns matched by 12!). Because the T302 timer is aborted and because the urgent-priority pattern is the best match, no further user input is expected. Unified CM routes the call using pattern 12[2-5].

Call Classification

- Calls using this route pattern can be classified as on-net or off-net calls. This route pattern can be used to prevent toll fraud by prohibiting off-net to off-net call transfers or by tearing down a conference bridge when no on-net parties are present. (Both of these features are controlled by Service Parameters within Unified CM Administration.)
- When the "Allow device override" box is enabled, the calls are classified based on the call classification settings on the associated gateway or trunk.

Forced Authorization Codes (FAC)

- The Forced Authorization Codes (FAC) checkbox is used to restrict the outgoing calls when using a particular route pattern. If you enable FAC through route patterns, users must enter an authorization code to reach the intended recipient of the call.
- When a user dials a number that is routed through a FAC-enabled route pattern, the system plays a tone that prompts for the authorization code. To complete the call, the user authorization code must meet or exceed the level of authorization that is specified to route the dialed number.
- Only the authorization name appears in the call detail records (CDR); the authorization code does not appear in the CDR.
- The FAC feature is not available if the "Allow overlap sending" checkbox is enabled.

Client Matter Codes (CMC)

- The Client Matter Code (CMC) checkbox is used to track calls to certain numbers when using a particular route pattern. (For example, a company can use it to track calls to certain clients.)
- If you enable CMC for a route pattern, users must enter a code to reach the intended destination.
- When a user dials a number that is routed through a CMC-enabled route pattern, the system plays a tone that prompts for the code. The user must enter the correct code in order to complete the call.
- Client Matter Codes appear in the call detail records so that they can be used by the CDR analysis and reporting tool to generate reports for client billing and accounting.
- The CMC feature is not available if the "Allow overlap sending" checkbox is enabled.
- If both CMC and FAC are enabled, the user dials a number, enters the FAC when prompted to do so, and then enters the CMC at the next prompt.

SIP Route Pattern

SIP route patterns are configured in Unified CM to route or block calls to external entities based on the host portion (right-hand side) of SIP URIs. A SIP route pattern can point directly to a SIP trunk or (starting with Unified CM 9.0) point to a route list that then refers to one or more route groups and finally to SIP trunks. The use of the full SIP route pattern, route list, route group construct is highly recommended because it offers more flexibility.

SIP route patterns matching on the host piece of a SIP URI can match on a domain name or an IP address, both of which are possible as the right-hand side of a SIP URI. Wildcards can be used in domain name SIP route patterns to match on multiple domains (for example, *.cisco.com and ccm[1-4].uc.cisco.com). In IP address SIP route patterns, a subnet notation can be used (for example, 192.168.10.0/24).

Route Lists

A route list is a prioritized list of eligible paths (route groups) for an outbound call. A typical use of a route list is to specify two paths for a remote destination, where the first-choice path is across the IP WAN and the second-choice path is through a PSTN gateway.

Route lists have the following characteristics:

- Multiple route patterns may point to the same route list.
- A route list is a prioritized list of route groups that function as alternate paths to a given destination. For example, you can use a route list to provide least-cost routing, where the primary route group in the list offers a lower cost per call and the secondary route group is used only if the primary is unavailable due to an "all trunks busy" condition or insufficient IP WAN resources.
- Each route group in the route list can have its own digit manipulation. For example, if the route pattern is 9.@ and a user dials 9 1 408 555 4000, the IP WAN route group can strip off the 9 1 while the PSTN route group may strip off just the 9.
- Multiple route lists can contain the same route group. The digit manipulation for the route group is associated with the specific route list that points to the route group.
- If you are performing several digit manipulations in a route pattern or a route group, the order in which the transformations are performed can impact the resulting calling and called party numbers used for the call. Unified CM performs the following major types of digit manipulations in the order indicated:
    1. Discarding digits
    2. Called/calling party transformations
    3. Prefixing digits
    4. Called/calling party transformation patterns

Route Groups

Route groups control and point to specific devices, which are typically gateways (MGCP, SIP, or H.323), H.323 or SIP trunks to a gatekeeper, remote Unified CM cluster, or Cisco Unified Border Element. Unified CM sends calls to the devices according to the distribution algorithm assigned. Unified CM supports top-down and circular algorithms.

Calling and Called Party Transformation Patterns

A calling party transformation pattern allows the system to adapt the global form of the calling party's number into the local form required by off-cluster networks connected to the route group devices, such as gateways or trunks.

A called party transformation pattern allows the system to adapt the global form of the called party's number into the local form required by off-cluster networks connected to the route group devices.

Note Called party transformation patterns do not have any effect on phones. The called party transformation pattern CSS of the device pool does not impart any effects on the phones to which it is assigned.

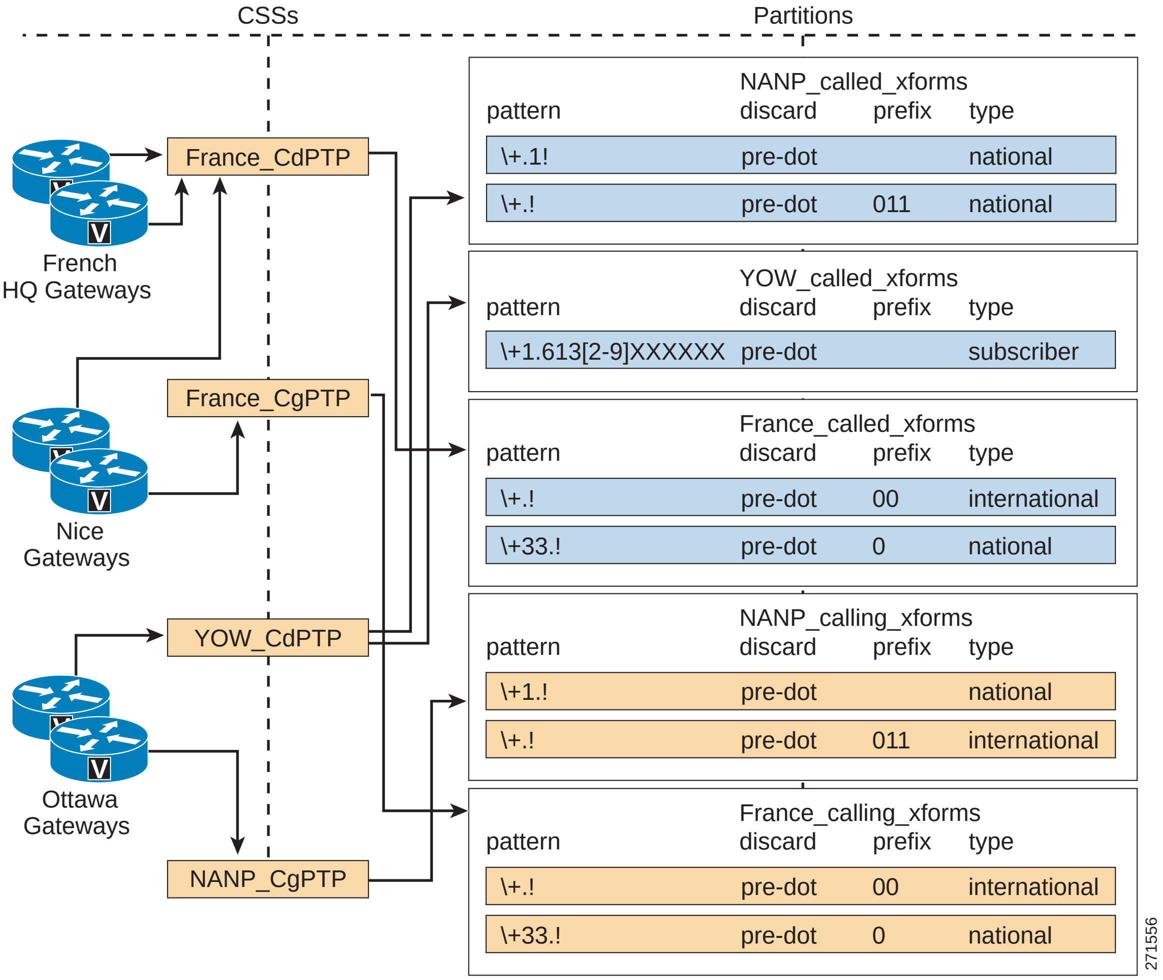
Both pattern types consist of a numerical representation of the calling or called party number to be matched. The syntax used is the same as that of other patterns such as route patterns, transformation patterns, directory numbers, and so forth. (See Figure 9-38.)

The transformation operators include discard digit instructions (for example, pre-dot), a calling party transformation mask, prefix digits, and control over the calling party presentation (either Default, Allowed, or Restricted). Calling party transformation patterns can be configured to use the calling party's external phone number mask as the calling party number.

Partitions and calling search spaces control which calling party transformation patterns are applied to which gateways or trunks. Gateways or trunks can use either their associated device pool's calling party transformation CSS or the device's own calling party transformation CSS, in reverse order of precedence. The same mechanism is used to control the applicability of called party transformation patterns.

Calling and called party transformation patterns configured on a Gateway Configuration page under **Call Routing Information - Outbound Calls** affect the calling or called party number sent to the gateway, as well as the calling or called party's numbering type and numbering plan. Calling party transformation patterns applied under **Incoming Calling Party Settings** apply to calls coming from the gateway.

Figure 9-38 Calling and Called Party Transformation Patterns



NANP_called_xforms

pattern	discard	prefix	type
\+.1!	pre-dot		national
\+.!	pre-dot	011	national

YOW_called_xforms

pattern	discard	prefix	type
\+1.613[2-9]XXXXXX	pre-dot		subscriber

France_called_xforms

pattern	discard	prefix	type
\+.!	pre-dot	00	international
\+33.!	pre-dot	0	national

NANP_calling_xforms

pattern	discard	prefix	type
\+1.!	pre-dot		national
\+.!	pre-dot	011	international

France_calling_xforms

pattern	discard	prefix	type
\+.!	pre-dot	00	international
\+33.!	pre-dot	0	national

Figure 9-38 illustrates how calling and called party transformation patterns would be applied to different groups of gateways connected to the PSTN in different parts of the PSTN.

Within the North American Numbering Plan (NANP), gateways located in Ottawa, Canada (airport code YOW) are assigned to the Calling Party Transformation CSS NANP_CgPTP, which contains partition NANP_calling_xforms. Any call with a calling party number beginning with +1 (that is, originating from within the NANP) would match both patterns configured within partition NANP_calling_xforms. Following the best-match logic, the first pattern will be chosen, and the calling party number will be stripped of the + sign and NANP country code 1. The remaining part of the calling party number will be used as the calling party number sent to the PSTN, with numbering type set to National.

For example, if a call from +1 613 555 1234 were sent out the YOW gateways, the calling party number would be transformed to 613 555 1234 with a numbering type set to National.

If a call from the same caller were to be sent to a French gateway, a different set of calling party transformation patterns would apply. For example, if a call from +1 613 555 1234 were sent out a gateway located in Nice, France (airport code NCE), the calling party transformation patterns contained in partition France_calling_xforms would be applied. In this case, the calling party number would be transformed to 001 613 555 1234 with the numbering type set to International.

Note Calling party number transformations may be overridden once the call is sent out the gateway. Many service providers will not permit calling party numbers outside a given range, as determined by local service agreements or regulations.

The same process applies to the called party number transformation patterns. For Ottawa gateways, the assigned called party transformation CSS is YOW_CdPTP, which contains partitions NANP_Called_xforms and YOW_Called_xforms. A call placed to a destination number within the Numbering Plan Area 613 would match all patterns contained in these two partitions. However, the best match process would select pattern \+1.613[2-9]XXXXXX.

For example, on a call placed to +1 613 555 9999 through the Ottawa gateways, the called party number would be transformed to 516 555 9999 with a numbering type set to Subscriber.

Incoming Calling Party Settings (per Gateway)

Incoming calling party settings can be configured on individual gateways, at the device pool level, or at the service parameter level, in order of precedence. For each numbering type (Subscriber, National, International, or Unknown), Unified CM allows for the appropriate prefix digits to be configured. Digits can be stripped from and prefixed to the string provided as the incoming party number. The notation takes the form PP:SS, where PP represents the digits to be prefixed and SS represents a quantity of digits to be stripped. The digit stripping operation is performed first on the incoming calling party number, and then the prefix digits are added to the resulting string. For example, if the prefix digits field is configured as +33:1 and the incoming calling party number is 01 58 40 58 58, the resulting string will be +33 1 58 40 58 58.

In Cisco Unified CM 7.1, each numbering type can be configured with a Calling Search Space used to apply Calling Party Transformation Patterns to the calls. The calling search space should contain partitions containing calling party transformation patterns exclusively. This allows the modifications applied to the calling party number to be based on the structure of the calling party number rather than strictly on its numbering type. The calling party transformation patterns use regular expressions to match the calling party number. The best-match process is used to choose between multiple matches, and the selected pattern's Calling Party Transformations are applied to the call.

Route Group Devices

The route group devices are the endpoints accessed by route groups, and they typically consist of gateways or trunks to a gatekeeper or to remote Unified CMs. You can configure the following types of devices in Unified CM:

- Media Gateway Control Protocol (MGCP) gateways
- SIP gateways
- H.323 gateways
- H.225 trunk, gatekeeper controlled — trunk to standard H.323 gateways, via a gatekeeper
- Intercluster trunk, not gatekeeper controlled — direct trunk to another Unified CM cluster
- Intercluster trunk, gatekeeper controlled — trunk to other Unified CM clusters and/or H.323 gateways, via a gatekeeper
- SIP trunk — trunk to another Unified CM cluster, a Cisco Unified Border Element, a Session Border Controller, or a SIP proxy

Note Both the H.225 and intercluster trunk (gatekeeper controlled) will automatically discover if the other endpoint is a standard H.323 gateway or a Unified CM and will select H.225 or Intercluster Trunk protocol accordingly.

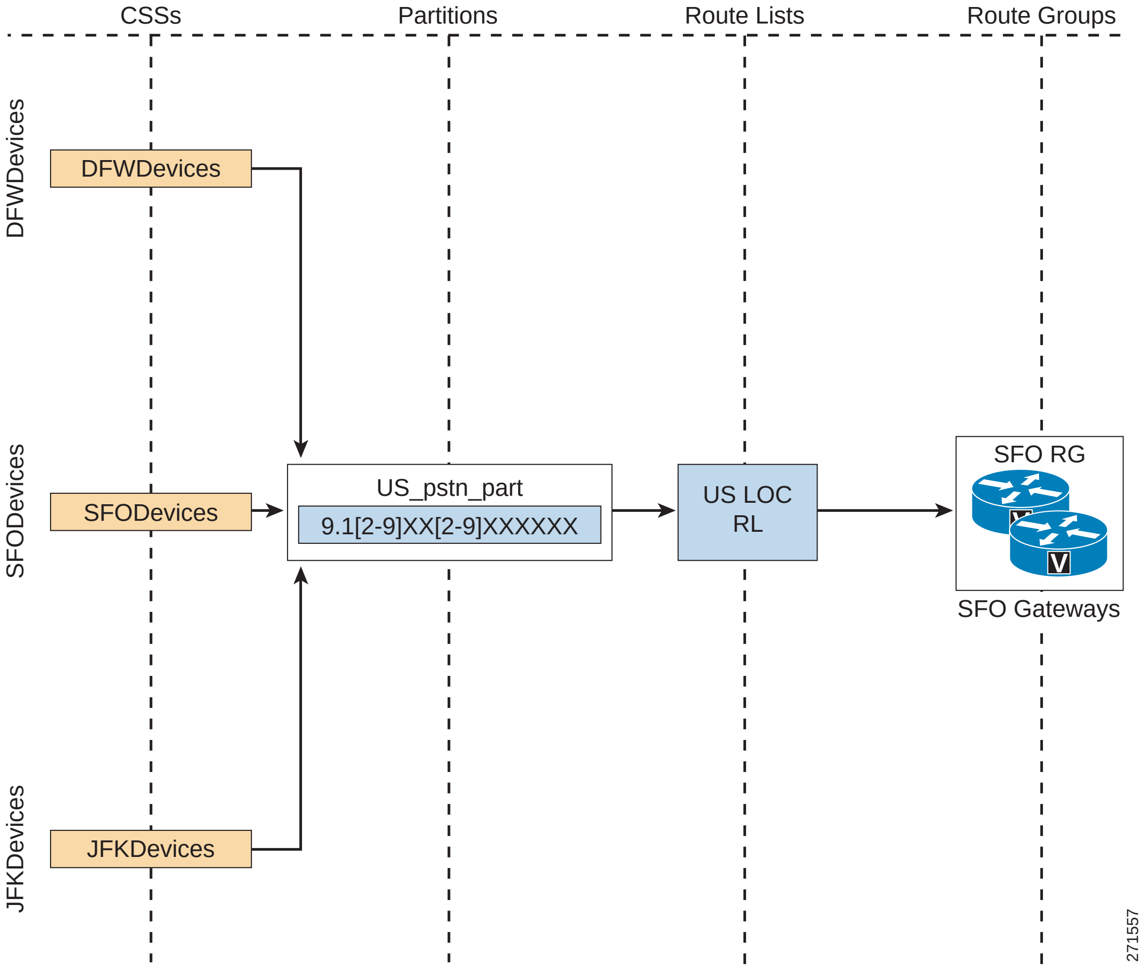
Local Route Group

Device pools can be associated with a local route group. Route patterns using the local route group offer a unique characteristic: they allow for dynamic selection of the egress gateway, based on the device originating the call. By contrast, calls routed by route patterns using static route groups will route the call to the same gateway, no matter what device originated the call.

Example 9-2 Comparison of Local and Non-Local Route Groups

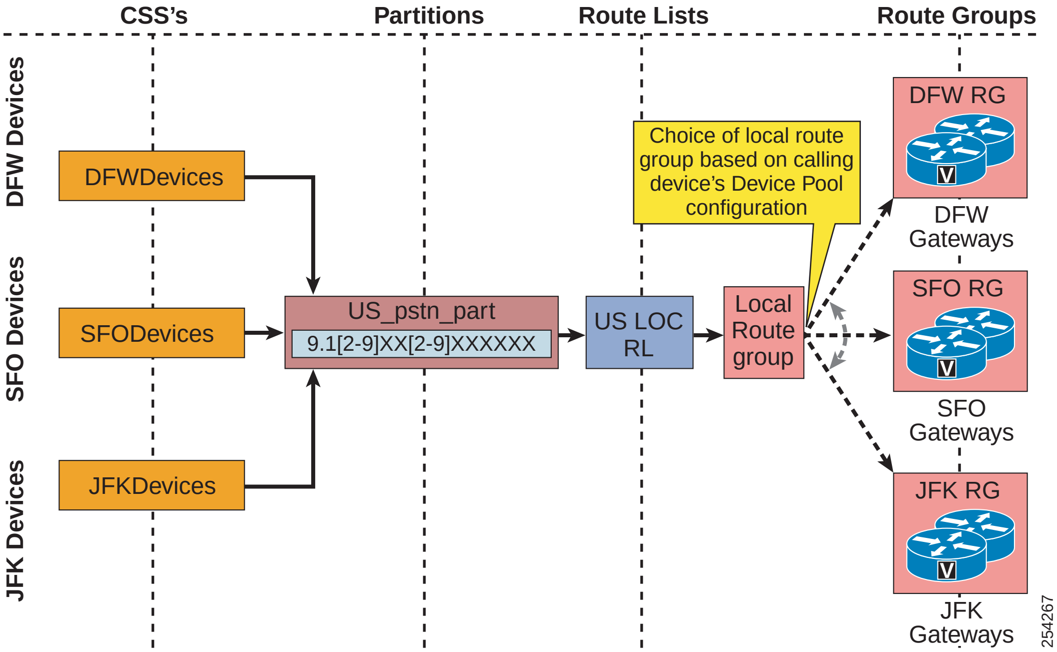
In Figure 9-39, a route pattern defined as 9.1[2-9]XX[2-9]XXXXXX points to a route list referencing a non-local route group containing San Francisco gateways. If this route pattern is placed in a partition contained in the calling search spaces of phones in Dallas, San Francisco, and New York, national calls from devices in those three cities will egress to the PSTN in San Francisco.

Figure 9-39 Non-Local Route Group Behavior



By contrast, if this same route pattern is modified to point to a route list containing the Standard Local Route Group as in Figure 9-40, then calls made from the Dallas site would egress to the PSTN through the Dallas gateway, calls made from the New York site would egress to the PSTN through the New York gateway, and calls made from the San Francisco site would egress to the PSTN through the San Francisco gateway.

Figure 9-40 Local Route Group Behavior



The Device Mobility feature allows the device pool to be assigned to an endpoint based on the current subnet to which it has roamed. This permits assignment of the local route group to be based on the site where the phone is currently located.

Example 9-3 Device Mobility

A phone is moved from the San Francisco site to the New York site. Based on the phone's new IP address (part of the IP subnet associated with the New York site), a New York device pool is assigned to the phone. If the next call placed by the roaming phone matches a route pattern using a route list containing the Standard Local Route Group, it will be routed through the New York gateway.

If a local route group is used in forwarded call scenarios where, for example, phone A calls phone B and B is forwarded to a destination in the PSTN, then the route pattern in the call forward calling search space of phone B determines the class of service for calls forwarded by phone B, whereas by default the local route group associated with phone A's device pool is used to determine the egress gateways when hitting Standard Local Route group in the route list selected by the route pattern found using phone B's call forward calling search space. As a result, typically a gateway local to phone A is used for the forwarded call. This makes sure that the caller ID of the initial caller (phone A) can be sent to the PSTN and that this caller ID will not be screened by the provider. Starting with Unified CM 9.0, there is a service parameter that allows you to configure the local route group selection policy for forwarded calls. The service parameter can be set to:

- **Calling Party's Local Route Group** — Backward compatible default. The local route group associated with the initial caller's device pool is selected (phone A in above example).
- **Original Called Party** — The local route group associated with the called phone's device pool is selected (phone B in above example).

- **Last Redirecting Party** — The local route group associated with the phone's device pool that is forwarding the call to the PSTN is selected (phone B in above example). These last two options differ only in cases where the call is forwarded through multiple hops before it finally gets forwarded out to the PSTN.

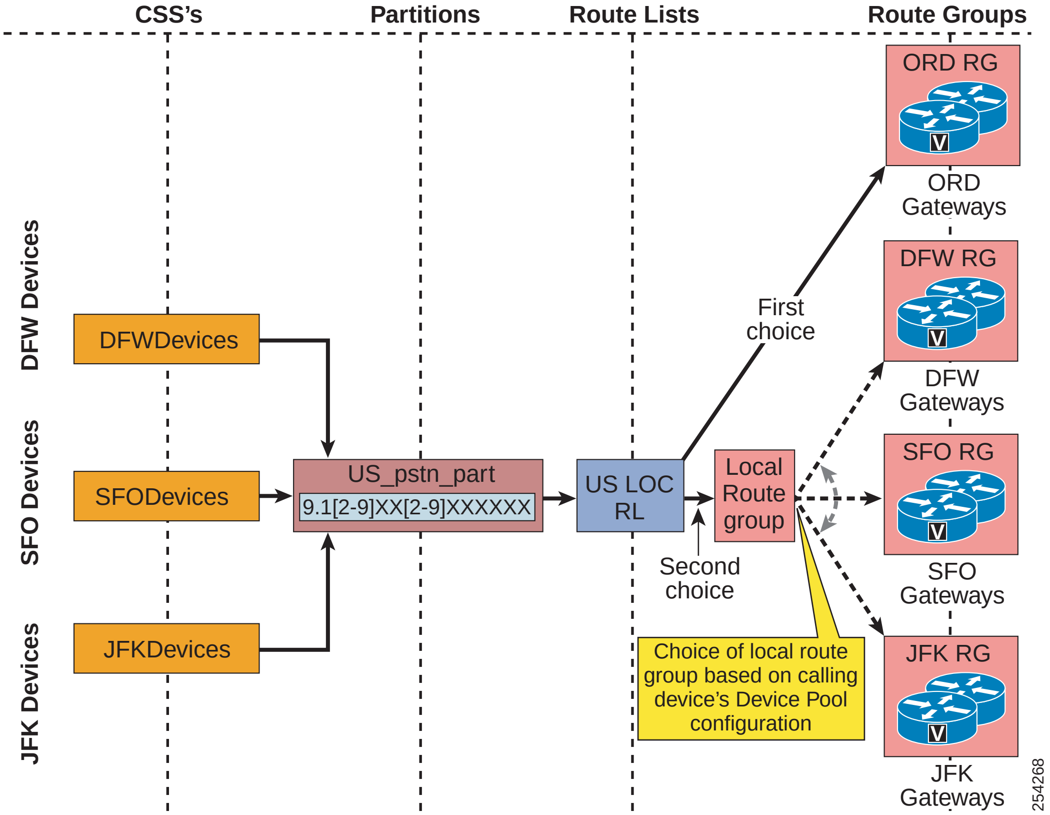
Centralized Gateway with Local Failover to the PSTN

Local route groups simplify the local failover to the PSTN for systems where a centralized gateway is configured. A single route list can be used to route PSTN calls for multiple sites while allowing local failover to the gateway at the site of origin.

Example 9-4 Centralized Gateways and Local Failover

A company negotiates a favorable PSTN interconnection rate for a group of trunks located in Chicago. If a route list includes a route group containing gateways in Chicago as its first entry and the Standard Local Route Group as the second choice, then any call it processes will first be sent to the preferred lower-cost gateways in Chicago. If a Chicago gateway is not available, if no ports are free, or if there is not enough bandwidth to allow the call between the calling phone and the Chicago gateway, then the next choice will be to attempt to route the call through the gateway co-located with the calling phone, as determined by the local route group in the calling phone's device pool configuration. (See Figure 9-41.)

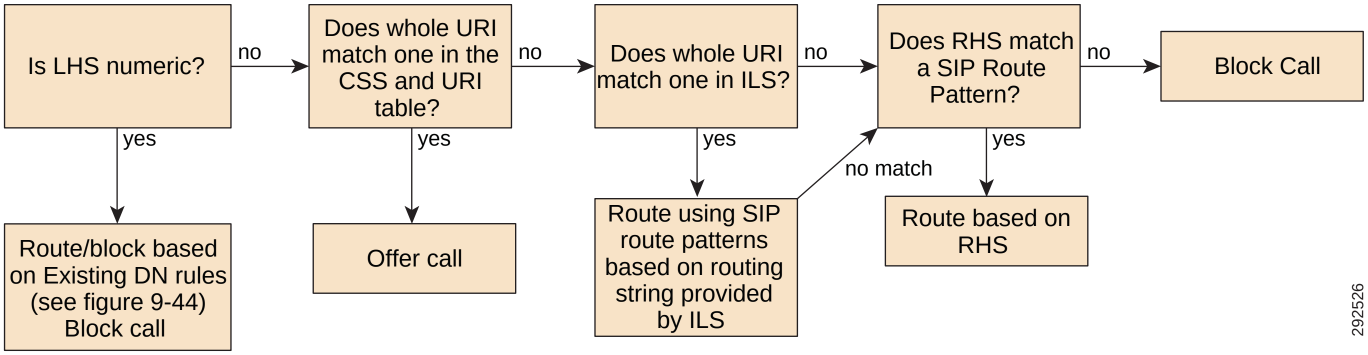
Figure 9-41 Centralized Gateway with Local Failover to the PSTN



Routing of SIP Requests in Unified CM

Routing of SIP requests received from SIP trunks or SIP endpoints follows certain rules to make sure that both local and intercluster routing requirements are met. Figure 9-42 shows a flowchart of routing decisions made by Unified CM. The first step is to check whether the left-hand side (user portion) of the URI is a directory number or a directory URI.

Figure 9-42 Call Routing Logic for SIP Request


Is LHS numeric?
no
Does whole URI match one in the CSS and URI table?
no
Does whole URI match one in ILS?
no
Does RHS match a SIP Route Pattern?
no
Block Call
yes
Route/block based on Existing DN rules (see figure 9-44) Block call
yes
Offer call
yes
Route using SIP route patterns based on routing string provided by ILS
no match
yes
Route based on RHS
292526

Numeric URI Versus Directory URI

If the SIP request carries a user=phone tag, the SIP URI will always be interpreted as a numeric SIP URI and Unified CM assumes that the user portion of the SIP URI is a directory number. If no user=phone is present, the decision is based on the dial string interpretation setting in the calling device's (endpoint or trunk) SIP profile. This setting either defines a set of characters that Unified CM will accept as part of numeric SIP URIs (0-9, *, #, +, and optionally A-D) or it enforces the interpretation as a directory URI.

For routing purposes prior to Unified CM 9.0, whether a URI is numeric or alpha, all SIP URI routing follows the numeric routing logic described in the section on Routing Numeric URIs, page 9-109.

Routing Directory URIs

The next step after identifying a SIP URI as being non-numeric is to try to route the SIP request based on the calling search space of the calling device. Unified CM searches for a full match of the SIP URI against all directory URIs configured in the partitions addressed by the calling device's calling search space. If a match is found, the call is extended to the directory number associated with the matched local directory URI.

In case no matching local directory URI is found, Unified CM tries to locate the SIP URI in imported directory URI catalogs or directory URI catalogs learned through ILS from remote systems, again by searching for a full match. In case of a match, the SIP request is routed by matching the SIP route string associated with the found directory URI against configured SIP route patterns. (See Figure 9-43.)

In case the SIP URI does not match a local directory URI and also does not match any directory URI in any directory URI catalog, Unified CM then routes the SIP request based only on matching the right-hand side of the SIP URI against configured SIP route patterns. This routing of last resort can be used to create a default route for all SIP URIs not known locally or on any other call control connected through ILS.

Figure 9-43 Example for Routing a Directory URI

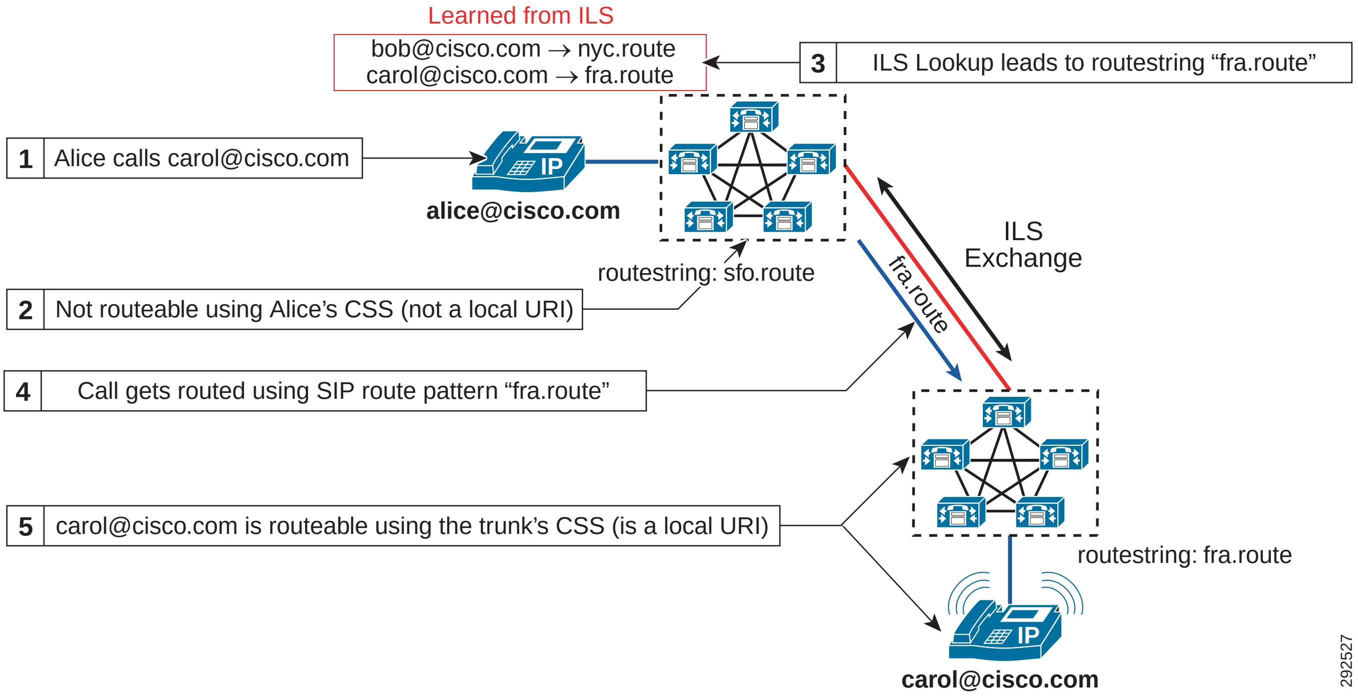

Learned from ILS
bob@cisco.com → nyc.route
carol@cisco.com → fra.route
3 ILS Lookup leads to routestring "fra.route"
1 Alice calls carol@cisco.com
alice@cisco.com
routestring: sfo.route
2 Not routeable using Alice's CSS (not a local URI)
ILS
Exchange
fra.route
4 Call gets routed using SIP route pattern "fra.route"
5 carol@cisco.com is routeable using the trunk's CSS (is a local URI)
routestring: fra.route
carol@cisco.com
292527

Figure 9-43 shows an example of how a dialed directory URI might be routed by Unified CM. In this example the bottom Unified CM cluster advertises the local directory URI carol@cisco.com. All local directory URIs of this Unified CM cluster are advertised under the SIP routestring fra.route. As part of this information exchange over ILS, the Unified CM cluster at the top populated its local directory URI catalog with the association of carol@cisco.com to the SIP routestring fra.route. If someone then places a call from the phone registered in the top cluster to directory URI carol@cisco.com, the local lookup of directory URI carol@cisco.com will fail because carol@cisco.com is not a local directory URI. The next step in the routing process is to search for carol@cisco.com in the table of directory URIs learned through ILS. This search will find the information learned from the bottom cluster, and the originating cluster at the top then takes the learned SIP routestring fra.route and tries to find a route by matching this SIP routestring fra.route against the configured SIP route patterns. A SIP route pattern fra.route is configured and points to a route list that ultimately leads to the SIP trunk pointing to the target Unified CM cluster. The originating Unified CM cluster thus routes the call down to the destination Unified CM cluster. The destination in the sent SIP request will be carol@cisco.com. On the destination cluster, the same routing logic as shown in Figure 9-42 then tries to match carol@cisco.com against all local directory URIs on the destination cluster, which leads to a full match and the target device rings.

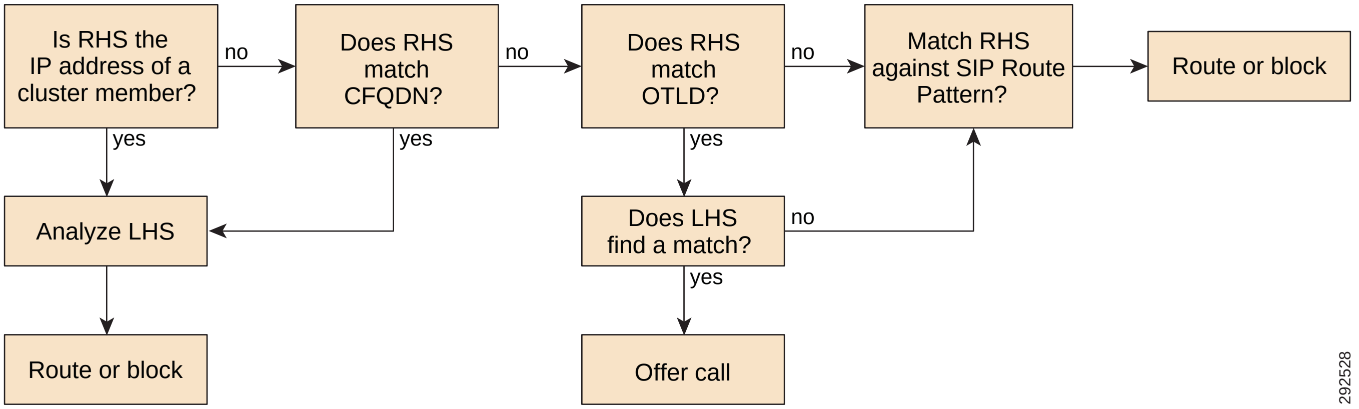
The above example shows that the SIP route string namespace is completely independent of the directory URI namespace. There is no requirement to use SIP route strings that are related in any way to the structure of the namespace used for the host portion of directory URIs. This allows to optimize the SIP route string namespace based on the desired routing topology. To disambiguate between SIP route patterns used to directly match on the URI host portion and SIP route patterns used to route directory URIs based on SIP route strings, Cisco highly recommends using an independent namespace for SIP route string route patterns (for example, ".route" or ".ils").

In the above example, the SIP route strings chosen basically identify the individual call controls (fra.route, nyc.route), and the SIP route pattern grid used to route directory URI SIP requests based on learned SIP route strings uses explicit patterns (fra.route, nyc.route) to create the desired reachability. In a hierarchical topology, hierarchical SIP route strings (for example, sjc.us.route, nyc.us.route, fra.de.route, and muc.de.route) might be used together with wildcard SIP route patterns (*.de.route, *.us.route) routing to the respective aggregating Cisco Unified Communications Manager Session Management Edition (SME) clusters responsible for the addressed set of Unified CM clusters.

Routing Numeric URIs

If a SIP URI is considered to be a numeric URI either because the request included a user=phone tag or based on the originating device's SIP profile dial string interpretation settings, the call is handled according to the flowchart shown in Figure 9-44. For Unified CM prior to release 9.0, this is the standard routing procedure for routing of SIP requests.

Figure 9-44 Call Routing Logic for numeric SIP Request



The first step is to check whether the right-hand side of the SIP URI is an IP address of any server that is a member of the Unified CM cluster or matches the cluster fully qualified domain name configured in Unified CM enterprise parameters. In this case the left-hand side of the URI is considered to be a local directory number and will be routed as a number using the calling device's calling search space.

The next step is to check whether the right-hand side of the SIP URI matches the organization’s top-level domain configured in Unified CM enterprise parameters. If this is the case, again Unified CM will try to route the call using the calling device's calling search space. But if no match can be found, then routing will fall back to route the call by matching the right-hand side of the SIP URI against the configured SIP route patterns.

Assuming a Unified CM cluster with cluster members having IP addresses 192.168.10.10, 192.168.10.11, 192.168.20.10, and 192.168.20.11, cluster fully qualified domain name configured as ucm1.cisco.com, and organization top-level domain configured as cisco.com, then all of the following SIP URIs would be routed to local directory number 1234:

- 1234@192.168.10.10
- 1234@192.168.10.11
- 1234@192.168.20.10
- 1234@192.168.20.11
- 1234@ucm1.cisco.com
- 1234@cisco.com

Assuming that no local directory number 1234 exists, the first five calls would fail immediately while Unified CM would try to route the sixth call by matching cisco.com against the configured SIP route patterns.

Calling Privileges in Unified CM

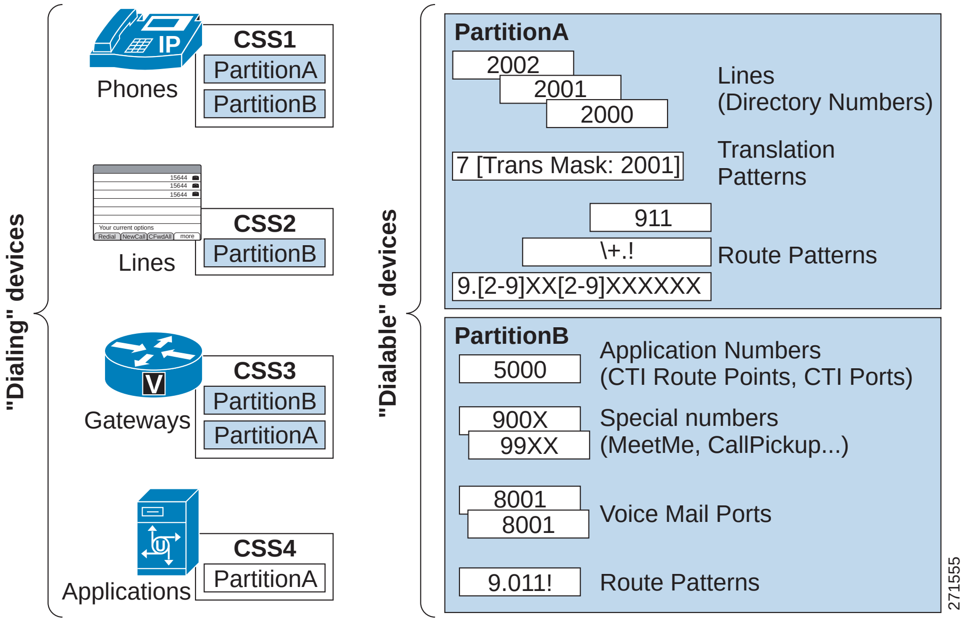
Dialing privileges are configured in order to control which types of calls are allowed (or prevented) for a particular endpoint (such as phones, gateways, or CTI applications). All calls handled by Unified CM are subjected to the dialing privileges implemented through the configuration of the following elements:

- Partitions, page 9-110
- Calling Search Spaces, page 9-112

A *partition* is a group of directory numbers (DNs) or directory URIs with similar accessibility, and a *calling search space* defines which partitions are accessible to a particular device. A device can call only those DNs and directory URIs located in the partitions that are part of its calling search space.

As illustrated in Figure 9-45, items that can be placed in partitions all have a dialable pattern, and they include phone lines, route patterns, translation patterns, CTI route group lines, CTI port lines, voicemail ports, and Meet-Me conference numbers. Conversely, items that have a calling search space are all devices capable of dialing a call, such as phones, phone lines, gateways, and applications (via their CTI route groups or voicemail ports).

Figure 9-45 Partitions and Calling Search Spaces

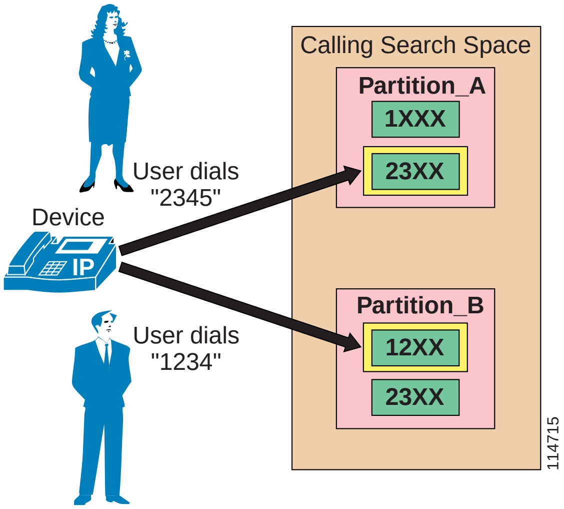


Partitions

The dial plan entries that you may place in a partition include IP phone directory numbers, directory URIs, translation patterns, route patterns, CTI route points, and voicemail ports. As described in the section on Call Routing in Unified CM, page 9-93, if two or more numeric dial plan entries (directory numbers, route patterns, or so forth) overlap, Unified CM selects the entry with the closest match (most specific match) to the dialed number. In cases where two dial plan entries match the dialed pattern equally, Unified CM selects the dial plan entry that appears first in the calling search space of the device making the call. Directory URIs always have to match completely; there is no concept of partial matches for directory URIs.

For example, consider Figure 9-46, where route patterns 1XXX and 23XX are part of Partition_A and route patterns 12XX and 23XX are part of Partition_B. The calling search space of the calling device lists the partitions in the order Partition_A:Partition_B. If the user of this device dials 2345, Unified CM selects route pattern 23XX in Partition_A as the matching entry because it appears first in the calling device's calling search space. However, if the user dials 1234, Unified CM selects route pattern 12XX in Partition_B as the matching entry because it is a closer match than 1XXX in Partition_A. Remember that the partition order in a calling search space is used exclusively as a tie-breaker in case of equal matches based on the closest-match logic.

Figure 9-46 Impact of Partition Order on the Matching Logic



Note When multiple equal-precision matches occur in the same partition, Unified CM selects the entry that is listed first in its local dial plan database. Since you cannot configure the order in which the dial plan database lists dial plan entries, Cisco strongly recommends that you avoid any possibility of equal-precision matches coexisting within the same partition because the resulting dial plan logic is not predictable in such cases.

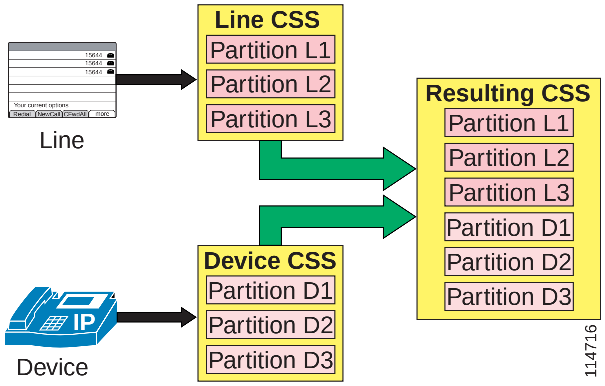
Partitions can be activated or deactivated based on the time and date. You can activate or deactivate partitions by first configuring time periods and schedules within Unified CM Administration and then assigning a specific time schedule to each partition. Outside of the times and days specified by the schedule, the partition is inactive, and all patterns contained within it are ignored by the Unified CM call routing engine. For more information on this feature, see Time-of-Day Routing, page 9-134.

Calling Search Spaces

A calling search space defines which partitions are accessible to a particular device. Devices that are assigned a certain calling search space can access only the partitions listed in that calling search space. Attempts to dial a DN or directory URI in a partition outside that calling search space will fail, and the caller will hear a busy signal.

If you configure a calling search space both on an IP phone line and on the device (phone) itself, Unified CM concatenates the two calling search spaces and places the line's calling search space in front of the device's calling search space, as shown in Figure 9-47.

Figure 9-47 Concatenation of Line and Device Calling Search Spaces for IP Phones



Note When device mobility is not used, the device calling search space is static and remains the same even as the device is moved to different parts of the network. When device mobility is enabled, the device calling search space can be determined dynamically based on where in the network the phone is physically located, as determined by the phone's IP address. See Device Mobility, page 9-122, for more details.

If the same route pattern appears in two partitions, one contained in the line's calling search space and one contained in the device's calling search space, then according to the rules described in the section on Partitions, page 9-110, Unified CM selects the route pattern listed first in the concatenated list of partitions (in this case, the route pattern associated with the line's calling search space).

For recommendations on how to set the line and device calling search spaces, refer to the sections on Building Classes of Service for Unified CM with the Traditional Approach, page 9-54, and Building Classes of Service for Unified CM with the Line/Device Approach, page 9-57.

The maximum length of the combined calling search space (device plus line) is 1024 characters, including separator characters between each partition name. (For example, the string “partition_1:partition_2:partition_3” contains 35 characters.) Thus, the maximum number of partitions in a calling search space varies, depending on the length of the partition names. Also, because the calling search space clause combines the calling search space of the device and that of the line, the maximum character limit for an individual calling search space is 512 (half of the combined calling search space clause limit of 1024 characters).

Therefore, when you are creating partitions and calling search spaces, keep the names of partitions short relative to the number of partitions that you plan to include in a calling search space. For more details on configuring calling search spaces, refer to the *Cisco Unified Communications Manager Administration Guide*, available online at http://www.cisco.com

Before you configure any partitions or calling search spaces, all DNs reside in a special partition named <None>, and all devices are assigned a calling search space also named <None>. When you create custom partitions and calling search spaces, any calling search space you create also contains the <None> partition, while the <None> calling search space contains *only* the <None> partition.

Note Any dial plan entry left in the <None> partition is implicitly reachable by *any* device making a call. Therefore, to avoid unexpected results, Cisco strongly recommends that you do not leave dial plan entries in the <None> partition.

Note Cisco strongly recommends that you do not leave any calling search space defined as <None>. Doing so can introduce dial plan behavior that is difficult to predict.

Special Considerations for Transformation Patterns

Calling and called transformation patterns are also placed in partitions, and those partitions are included in calling search spaces (CSSs) but not in order to control calling privileges. The partitioning of transformation patterns serves to choose which transformations are applied to which gateways, trunks, or phones. Partitions contained in calling party transformation pattern CSSs should contain only calling party transformation patterns. Likewise, partitions contained in called party transformation pattern CSSs should contain only called party transformation patterns.

Call-Forward Calling Search Spaces

Note Call Forward All actions are different than any other call-forward action in that the destination number is entered by each individual user when the feature is activated from a phone.

The system allows you to decide how call-forward calling search spaces take effect. There are three possible options, as selected by the Calling Search Space Activation policy:

- Use System Default

  If you configure the Calling Search Space Activation Policy to Use System Default, then the CFA CSS Activation Policy cluster-wide service parameter determines which Forward All Calling Search Space will be used. The CFA CSS Activation Policy service parameter can be set to With Configured CSS or to With Activating Device/Line CSS (see below). By default, the CFA CSS Activation Policy service parameter is set to With Configured CSS.

- With Configured CSS

  If you select the With Configured CSS option, the Forward All Calling Search Space and Secondary Calling Search Space for Forward All explicitly configured in the Directory Number Configuration window control the forward-all activation and call forwarding. If the Forward All Calling Search Space is set to None, no CSS gets configured for Forward All. A forward-all activation attempt to any directory number with a partition will fail. No change in the Forward All Calling Search Space and Secondary Calling Search Space for Forward All occurs during the forward-all activation.

- With Activating Device/Line CSS

  If you prefer to use the combination of the Directory Number Calling Search Space and Device Calling Search Space without explicitly configuring a Forward All Calling Search Space, select With Activating Device/Line CSS for the Calling Search Space Activation Policy. With this option, when Forward All is activated from the phone, the Forward All Calling Search Space and Secondary Calling Search Space for Forward All are automatically populated with the Directory Number Calling Search Space and Device Calling Search Space for the activating device. When you set the Forward All Destination from Unified CM Administration, the Forward All Calling Search Space and Secondary Calling Search Space are not automatically populated and have to be configured explicitly. The two calling search spaces are concatenated, and the resulting calling search space is used to validate the number entered as a Call Forward All destination. For further details, see Building Classes of Service for Unified CM with the Line/Device Approach, page 9-57.

With this configuration (Calling Search Space Activation Policy set to With Activating Device/Line), if the Forward All Calling Search Space is set to None when forward-all is activated through the phone, the combination of Directory Number Calling Search Space and activating Device Calling Search Space is used to verify the forward-all attempt.

Note Call Forward All configuration typically has to satisfy two requirements: controlling the destinations to which the device is allowed to forward calls, and ensuring that optimum call routing is achieved when calls originating from various points of origin are forwarded to the Call Forward All destination. To achieve both requirements, Cisco recommends using the With Configured CSS activation policy, which allows for controlling the destinations through the Line-Device dial plan approach; the Call Forward All CSS is used to implement a set of restrictions through the use of blocked patterns, and it can be set to the same calling search space configured on the line if the regular class of service is to be used for Call Forward All. The Secondary Calling Search Space for Call Forward All should then be configured to route calls to the Standard Local Route Group; the actual route group used to route the call will be determined at the time of the call, based on the calling (forwarded) device's local route group as configured on its device pool.

On Type-A IP phones running SIP, if Call Forward All is invoked from the phone itself, the device's Rerouting Calling Search Space is used for forwarded calls. If Forward All actions are invoked from the Unified CM User page or the Unified CM Administrative page, then any Forward All action initiated from the phone is irrelevant.

For example, assume an Type-A IP phone running SIP is configured with Forward All to extension 3000 from the Unified CM User page. At the same time, the phone itself is configured to Forward All to extension 2000. All calls made to that phone will be forwarded to extension 3000.

Note On Type-A IP phones running SIP, invoking Forward All from the Unified CM User or Administrative pages will not be reflected on the phone. The phone does not display any visual confirmation that calls are forwarded.

When Forward All is initiated from an IP phone running SCCP or from an Type-B IP phone running SIP, user input is simultaneously compared to the patterns allowed in the configured Forward All calling search space(s). If an invalid destination pattern is configured, the user will be presented with reorder tone. When Forward All is invoked from an Type-A IP phone running SIP, Forward All user input is stored locally on the phone and is not verified against any calling search space in Unified CM. If user input corresponds to an invalid destination, no notification is offered to the user. Calls made to that phone will be presented with reorder tone as the phone tries to initiate a SIP re-route action to an invalid destination number.

Other Call Forward Types

The calling search spaces configured for the various other types of call forward (Forward Busy, Forward No Answer, Forward No Coverage, forward on CTI failure, and Forward Unregistered) are standalone values not concatenated with any other calling search space.

Call Forward settings (except Forward All) can be configured separately for internal or external call types. For example, a user might want to have their phone Call Forward No Answer to voicemail for external callers but forward to a cell phone number if the caller is a co-worker calling from another IP phone on the network. This is possible by using different configurations for the Internal and External Call Forward settings.

When the Forward All calling search space is left as <None>, the results are difficult to predict and depend on the Unified CM release. Therefore, Cisco recommends the following best practices when configuring call-forward calling search spaces:

- Always provision the call-forward calling search spaces with a value other than <None>. This practice avoids confusion and facilitates troubleshooting because it enables the network administrator to know exactly which calling search space is being used for forwarded calls.
- Configure the Call Forward Busy and Call Forward No Answer calling search spaces with values that allow them to reach the DNs for the voicemail pilot and voicemail ports but not external PSTN numbers.
- Configure both the Call Forward All calling search space and the Secondary Calling Search Space for Forward All, according to your company's policy. Many companies choose to restrict forwarded calls to internal numbers only, to prevent users from forwarding their IP phone lines to a long-distance number and dialing their local IP phone number from the PSTN to bypass long-distance toll charges on personal calls.

The Call Forward Unregistered (CFUR) feature is a way to reroute calls placed to a temporarily unregistered destination phone. The configuration of CFUR consists of two main elements:

- Destination selection

  When the DN is unregistered, calls can be rerouted to either of the following destinations:

  - Voicemail

    Calls can be sent to voicemail by selecting the voicemail checkbox and configuring the CFUR calling search space to contain the partition of the voicemail pilot number.

  - A directory number used to reach the phone through the PSTN

    This approach is preferred when a phone is located within a site whose WAN link is down. If the site is equipped with Survivable Remote Site Telephony (SRST), the phone (and its co-located PSTN gateway) will re-register with the co-located SRST router. The phone is then able to receive calls placed to its PSTN DID number.

    In this case, the appropriate CFUR destination is the corresponding PSTN DID number of the original destination DN. Configure this PSTN DID in the destination field, preferably in E.164 format, including the + sign (for example, +1 415 555 1234). This allows the CFUR destination to be processed by the calling phone's local route group, whether or not it uses the same off-net access code and PSTN prefixes as the unregistered phone.

- Calling search space

  Unified CM attempts to route the call to the configured destination number by using the called DN's CFUR calling search space. The CFUR calling search space is configured on the target phone and is used by all devices calling the unregistered phone. This means that all calling devices will use the same combination of route pattern, route list, and route group to place the call. Cisco recommends that you configure the CFUR calling search space to route calls to the CFUR destination using patterns pointing to route lists referencing the Standard Local Route Group. This will ensure that the egress gateway to the PSTN is chosen based on the calling device.

The Call Forward Unregistered functionality can result in telephony routing loops if a phone is unregistered while the gateway associated with the phone's DID number is still under control of Unified CM, as is the case if a phone is simply disconnected from the network. In such a case, the initial call to the phone would prompt the system to attempt a first CFUR call to the phone's DID through the PSTN. The resulting incoming PSTN call would in turn trigger another CFUR attempt to reach the same phone's DN, triggering yet another CFUR call from the central PSTN gateway through the PSTN. This cycle could repeat itself until system resources are exhausted.

The service parameter MaximumForwardUnRegisteredHopsToDn controls the maximum number of CFUR calls that are allowed for a DN at the same time. The default value of 0 means the counter is disabled. If any DNs are configured to reroute CFUR calls through the PSTN, loop prevention is required. Configuring this service parameter to a value of 1 would stop CFUR attempts as soon as a single call is placed through the CFUR mechanism. This setting would also allow only one call to be forwarded to voicemail, if CFUR is so configured. Configuring this service parameter to a value of 2 would allow for up to two simultaneous callers to reach the voicemail of a DN whose CFUR setting is configured for voicemail, while also limiting potential loops to two for DNs whose CFUR configuration sends calls through the PSTN.

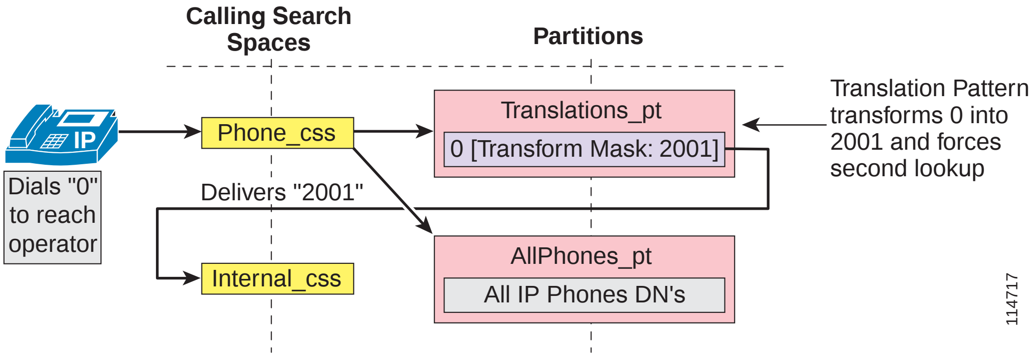
Note Extension Mobility DNs should not be configured to send Call Forward Unregistered calls to the PSTN DID associated with the DN. The DNs of Extension Mobility profiles in the logged-out state are deemed to be unregistered, therefore any calls to the PSTN DID number of a logged-out DN would trigger a routing loop. To ensure that calls made to Extension Mobility DNs in the logged-out state are sent to voicemail, ensure that their corresponding Call Forward Unregistered parameters are configured to send calls to voicemail.

Translation Patterns

Translation patterns are one of the most powerful tools in Unified CM to manipulate digits for any type of call. They follow the same general rules and use the same wildcards as route patterns. As with route patterns, you assign a translation pattern to a partition. However, when the dialed digits match the translation pattern, Unified CM does not route the call to an outside entity such as a gateway; instead, it performs the translation first and then routes the call again, this time using the calling search space configured within the translation pattern.

Translation patterns can be used for a variety of applications, as shown by the example in Figure 9-48.

Figure 9-48 Application Example for Translation Patterns



In this example, the administrator wishes to provide users with an operator service that is reached by dialing 0, while also maintaining a fixed-length internal numbering plan. The IP phones are configured with the Phone_css calling search space, which contains the Translations_pt partition (among others). A translation pattern 0 is defined in this partition, and the configured Called Party Transform Mask instructs Unified CM to replace the dialed string (0) with the new string 2001, which corresponds to the DN of the operator phone. A second lookup (of 2001 this time) is forced through the call routing engine, using the Internal_css calling search space, and the call can now be extended to the real operator DN of 2001, which resides in the AllPhones_pt partition.

Note When a dialed number is manipulated using a translation pattern, the translated number is recorded in the call detail record (CDR). However, when the digit manipulation occurs within a route list, the CDR will show the originally dialed number, not the translated one. The Placed Calls directory on the IP phone always shows the string as it was dialed by the user.

Automated Alternate Routing

The automated alternate routing (AAR) feature enables Unified CM to establish an alternate path for the voice media when the preferred path between two endpoints within the same cluster runs out of available bandwidth, as determined by the locations mechanism for call admission control.

The AAR feature applies primarily to deployments with sites connected via a WAN. For instance, if a phone in branch A calls a phone in branch B and the available bandwidth for the WAN link between the branches is insufficient (as computed by the Locations mechanism), AAR can reroute the call through the PSTN. The audio path of the call would be IP-based from the calling phone to its local (branch A) PSTN gateway, TDM-based from that gateway through the PSTN to the branch B gateway, and IP-based from the branch B gateway to the destination IP phone.

AAR can be transparent to the users. You can configure AAR so that users dial only the on-net (for example, four-digit) directory number of the called phone and no additional user input is required to reach the destination through the alternate network (such as the PSTN).

Note AAR does not support CTI route points as the origin or the destination of calls. Also, AAR is incompatible with the Extension Mobility feature when users roam across different sites. Refer to Extension Mobility, page 9-124, for more details.

You must provide the following main elements for AAR to function properly:

- Establish the PSTN Number of the Destination, page 9-118
- Prefix the Required Access Codes, page 9-118
- Select the Proper Dial Plan and Route, page 9-120

Establish the PSTN Number of the Destination

The rerouting of calls requires using a destination number that can be routed through the alternate network (for example, the PSTN). AAR uses the dialed digits to establish the on-cluster destination of the call and then combines them with the called party's AAR Destination Mask; if it is not configured, the External Phone Number Mask is used instead. The combination of the dialed digits and the applicable mask must yield a fully qualified number that can be routed by the alternate network.

Alternatively, by selecting the voicemail checkbox in the AAR configuration, you can allow calls to be directed to the voicemail pilot number. This choice does not rely on the numbers originally dialed by the caller, but routes the call according to the voicemail profile configuration.

Note By default, the directory number configuration retains the AAR leg of the call in the call history, which ensures that the AAR forward to the voice messaging system will select the proper voice mailbox. If you choose "Remove this destination from the call forwarding history," the AAR leg of the call is not present in the call history, which would prevent the automated voice mailbox selection and would offer the caller the generic voicemail greeting.

The AAR Destination Mask is used to allow the destination phone number to be determined independently of the External Phone Number mask. For example, if Caller ID policy for a company required a phone's external phone number mask to be the main directory number of an office (such as 415 555 1000), the AAR destination mask could be set to+1 415 555 1234, to provide AAR with the phone's specific PSTN number.

For example, assume phone A in San Francisco (DN = 2345) dials an on-net DN (1234) configured on phone B located in New York. If locations-based call admission control denies the call, AAR retrieves the AAR Destination Mask of the New York phone (+1212555XXXX) and uses it to derive a number (+12125551234) that can be used to route the call on the PSTN.

It is best to configure the AAR destination mask to yield a fully qualified E.164 number, including the + sign, because this will greatly simplify the overall configuration of AAR. For example, a phone in Paris is configured with an AAR destination mask of +33 1 58 04 58 58. Because this number is a fully qualified E.164 number, it contains all the information required for the Cisco Unified Communications system to derive a routable PSTN number as required by the calling phone's gateway to the PSTN, regardless of whether it is located in France, in Canada, or anywhere else in the world. The following sections elaborate on this approach.

Prefix the Required Access Codes

If the AAR Destination Yields a Fully Qualified E.164 Number Including the + Sign

This is the simplest case; the AAR destination contains + as a wildcard to be replaced by the appropriate access codes require at each gateway. The destination number is ready to be routed to an appropriate route pattern and then transformed at the point of egress to the PSTN by the appropriate called party transformation patterns.

Example 1: A phone in Ottawa, Canada calls a phone in Paris, which triggers AAR due to a lack of bandwidth on the WAN. The AAR destination is +33 1 58 04 58 58. The AAR calling search space of the calling phone contains a route pattern \+!, which routes the call to the Standard Local Route Group. The call is routed to the local gateway in Ottawa, where called party transformation patterns will replace the + with the applicable international access code 011. The resulting call is placed to 011 33 1 58 04 58 58.

Example 2: A phone in Nice, France calls a phone in Paris, which triggers AAR due to a lack of bandwidth on the WAN. The AAR destination is +33 1 58 04 58 58. The AAR calling search space of the calling phone contains a route pattern \+!, which routes the call to the Standard Local Route Group. The call is routed to the local gateway in Nice, where called party transformation patterns will replace the + 33 with the applicable national access code 0. The resulting call is placed to 01 58 04 58 58.

If the AAR Destination Mask Yields a Number Including the Country Code

The destination number (assumed to include the country code) might require a prefix to be routed properly by the origination branch's dial plan. Furthermore, if the point of origin is located in a different area code or even a different country, then other prefixes such as international dialing access codes (for example, 00 or 011) might be required as part of the dialed string.

When configuring AAR, you place the DNs in AAR groups. For each pair of AAR groups, you can then configure prefix digits to add to the DNs for calls between the two groups, including prefix digits for calls originating and terminating within the same AAR group.

As a general rule, place DNs in the same AAR group if they share the same inter-country dialing structure. For example, all phones in the UK dial numbers outside the UK with 9 as a PSTN access code, followed by 00 for international access; all phones in France and Belgium use 0 as a PSTN access code, followed by 00 for international access; all phones in the NANP use 9 as a PSTN access code, followed by 011 for international access.

This yields the following AAR group configuration:

AAR Group	NANP	Cent_EU	UK
NANP	9	9011	9011
Cent_EU	000	000	000
UK	900	900	9

Example 3: A phone in Ottawa, Canada calls a phone in Paris, which triggers AAR due to a lack of bandwidth on the WAN. The AAR destination is 33 1 58 04 58 58. The AAR group of the calling phone is NANP and that of the destination phone is Cent-EU, thus yielding a prefix of 9011. The AAR calling search space of the calling phone contains a site-specific route pattern 9011!, which routes the call to a route list in Ottawa, stripping the 9. The call is routed to the local gateway in Ottawa. The resulting call is placed to 011 33 1 58 04 58 58.

Example 4: A phone in Brussels, Belgium calls a phone in Paris, which triggers AAR due to a lack of bandwidth on the WAN. The AAR destination is 33 1 58 04 58 58. The AAR group of the calling phone and that of the destination phone is Cent-EU, thus yielding a prefix of 000. The AAR calling search space of the calling phone contains a site-specific route pattern 000!, which routes the call to a route list in Brussels, stripping the leading 0. The call is routed to the local gateway in Brussels. The resulting call is placed to 00 33 1 58 04 58 58.

Voicemail Considerations

AAR can direct calls to voicemail. The voicemail pilot number is usually dialed without the need for an off-net access code (if the voicemail pilot number is a fully qualified on-net number, such as 8 555 1000). When AAR is configured to send calls to voicemail, the AAR group mechanism will still prefix the configured access code(s). This configuration requires the creation of an AAR group to be used by all DNs whose desired AAR destination is voicemail (for example, vmail_aar_grp). Ensure that the configuration for this voicemail AAR group uses no prefix numbers when receiving calls from other AAR group DNs.

For example: Assume that DNs located in sites San Francisco and New York are configured with AAR group NANP, which prefixes 9 to calls made between any two DNs in the group. If a DN in San Francisco is configured to send AAR calls to voicemail (for example, 8 555 1000), a call would be placed to 985551000, which would result in a failed call. Instead, the San Francisco DN should be configured with AAR group vmail. The prefix digits for calls from AAR group NANP to AAR group vmail are <none>, as shown in the following table. The call will be placed successfully to 85551000.

AAR Group	NANP	Cent_EU	UK	vmail
NANP	9	9011	9011	<none>
Cent_EU	000	000	000	<none>
UK	900	900	9	<none>

Note When Device Mobility is not used, the AAR group configuration of a DN remains the same even as the device is moved to different parts of the network. With Device Mobility, the AAR group can be determined dynamically based on where in the network the phone is physically located, as determined by the phone's IP address. See Device Mobility, page 9-122, for more details.

Select the Proper Dial Plan and Route

AAR calls should egress through a gateway within the same location as the calling phone, thus causing the completed dial string to be sent through the origination site's dial plan. To ensure that this is the case, select the appropriate AAR calling search space on the device configuration page in Unified CM Administration. Configure the off-net dial plan entries (for example, route patterns) in the AAR calling search space to point to co-located gateways and to remove the access code before presenting the call to the PSTN.

For example, phones at the San Francisco site can be configured with an AAR calling search space that permits long distance calls dialed as 91-NPA-NXX-XXXX but that delivers them to the San Francisco gateway with the access code (9) stripped.

The AAR calling search space configuration can be greatly simplified if the local route group is used in conjunction with using a fully qualified E.164 address (including the + sign) as the AAR destination mask. A single calling search space configured with a single partition, containing a single route pattern \+!, pointing to a single route list featuring the Standard Local Route Group, can be used to route the calls of all phones at all sites in an entire cluster. This relies on the pre-configuration of the appropriate gateway-specific called party transformation patterns to adapt the universal form of the destination number to the localized form required by the service provider networks to which the call is delivered at each site.

Note If you have configured additional route patterns to force on-net internal calls dialed as PSTN calls, ensure that these patterns are not matched by the AAR feature. Refer to Design Guidelines for Multisite Deployments, page 9-35, for more details.

Note To avoid denial of re-routed calls due to call admission control, AAR functionality requires the use of a LAN as the IP path between each endpoint and its associated gateway to the PSTN. Therefore, AAR dial plans cannot rely on centralized gateways for PSTN access.

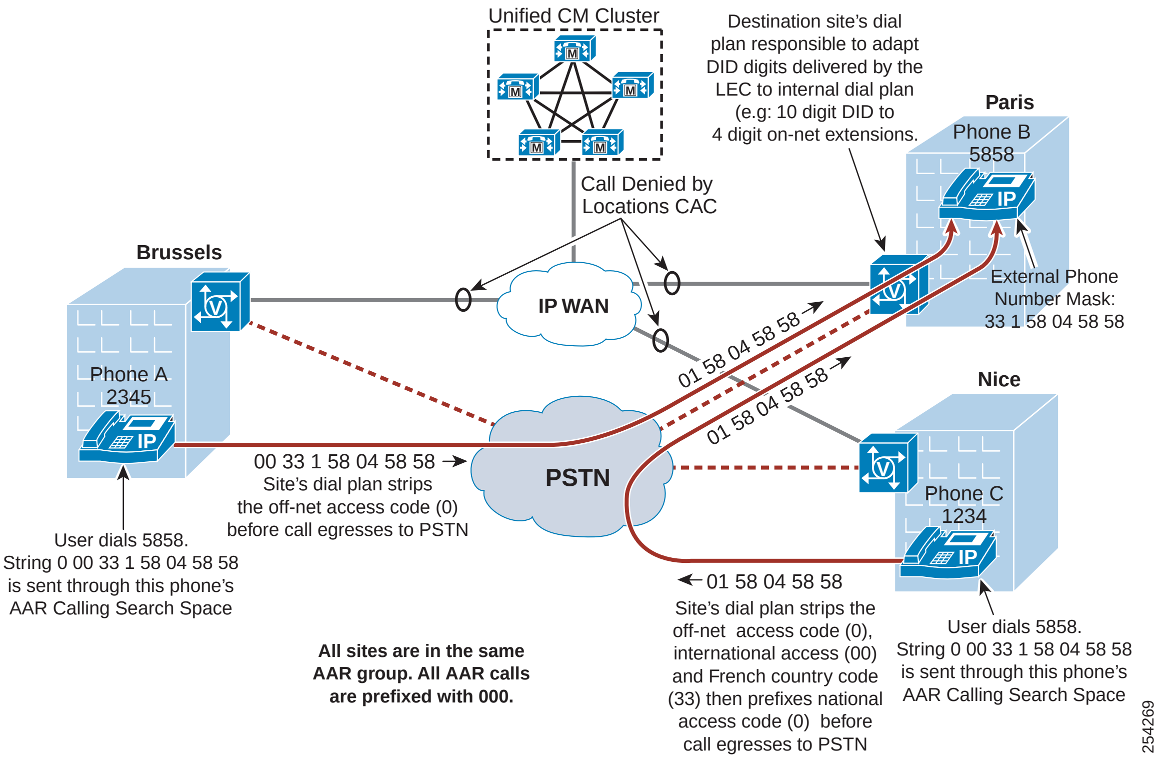
Note When Device Mobility is configured, the AAR calling search space can be determined dynamically based on where in the network the phone is physically located, as determined by the phone's IP address. See Device Mobility, page 9-122, for more details.

Special Considerations for Sites Located Within the Same Local Dialing Area

In some instances, the AAR dial string might have to be modified locally to allow for dialing between sites whose phones belong to the same AAR group. For example, assume two separate sites located in France share the same country code of 33. (See Figure 9-49.) In this case, a number dialed as 0 00 33 1 58 04 58 58 would have to be transformed to 01 58 04 58 58. Note that this is required only if the AAR configuration does not rely on called party transformation patterns.

This transformation is best done with a site-specific translation pattern of 00033.[1-6]XXXXXXXX to strip the pre-dot digits and prepend 0. This translation pattern should be placed in a member partition of the AAR calling search space for the French sites only; the Belgian site still needs to reach this same destination as 0 00 33 1 58 04 58 58.

Figure 9-49 Dialed Number Transformations for AAR Calls Between Sites


Unified CM Cluster
M
Destination site's dial plan responsible to adapt DID digits delivered by the LEC to internal dial plan (e.g: 10 digit DID to 4 digit on-net extensions.
Paris
Phone B
5858
IP
Call Denied by Locations CAC
Brussels
IP WAN
External Phone Number Mask: 33 1 58 04 58 58
Phone A
2345
01 58 04 58 58
01 58 04 58 58
Nice
00 33 1 58 04 58 58
Site's dial plan strips the off-net access code (0) before call egresses to PSTN
PSTN
Phone C
1234
User dials 5858. String 0 00 33 1 58 04 58 58 is sent through this phone's AAR Calling Search Space
01 58 04 58 58
Site's dial plan strips the off-net access code (0), international access (00) and French country code (33) then prefixes national access code (0) before call egresses to PSTN
User dials 5858. String 0 00 33 1 58 04 58 58 is sent through this phone's AAR Calling Search Space
All sites are in the same AAR group. All AAR calls are prefixed with 000.
254269

Note The AAR functionality is not triggered upon detection that the destination phone is unreachable. Therefore, WAN failures do not trigger AAR functionality.

In order to understand this better, consider an example where a Unified CM cluster has a site in London (United Kingdom), one in Paris (France), and one in Nice (France). The E.164 address of the DID range in Paris is +33145678XXX, but these extensions are usually reached as 0145678XXX when calling from within the French PSTN. When somebody in the London office wishes to dial the Paris office via the PSTN, the dialed string is 90033145678XXX. However, when somebody in the Nice office wishes to dial the Paris office via the PSTN, the dialed string is 00145678XXX.

To allow all three cases above with a single, simple AAR configuration, it is best to configure the AAR Destination Mask with the E.164 notation (including the + sign); this creates a destination number which can be interpreted by the system differently for each calling phone.

Device Mobility

Device Mobility offers functionality designed to enhance the mobility of devices within an IP network. (For example, a phone initially configured for use in San Francisco is physically moved to New York.) Although the device still registers with the same Unified CM cluster, it now will adapt some of its behavior based on the new site where it is located. Those changes are triggered by the IP subnet in which the phone is located.

When roaming, a phone will inherit the parameters associated with the device pool associated with the device's current subnet. From a dial-plan perspective, the functionality of the following five main configuration parameters can be modified due to the physical location of the phone. For these parameters to be modified, the device must be deemed as roaming outside its home physical location but within its home device mobility group.

- Local route group

  the roaming device pool's Local Route Group is used. For example, if a device is roaming from San Francisco to New York, the local route group of the New York device pool is used to route calls to the PSTN whenever a pattern points to a route list invoking the Standard Local Route Group.

- Calling party transformation CSS

  The roaming device pool's calling party transformation CSS is used. This allows a phone to inherit the calling party presentation mode that is customary for the phones of the visited location.

- Device calling search space

  The roaming device pool's Device Mobility calling search space is used instead of the device calling search space configured on the device's configuration page. For example, if a device is roaming from San Francisco to New York, the Device Mobility calling search space of the New York device pool is used as the roaming phone's device calling search space. If you use the line/device approach to classes of service, this approach will establish the path taken for PSTN calls, routing them to the local New York gateway.

- AAR calling search space

  The roaming device pool's AAR calling search space is used instead of the AAR calling search space configured on the device's configuration page. For example, if a device is roaming from San Francisco to New York, the AAR calling search space of the New York device pool is used as the roaming phone's AAR calling search space. This calling search space will establish the path taken for outgoing AAR PSTN calls, routing them to the local New York gateway.

- DN's AAR group

  For incoming AAR calls, the AAR group assigned to a DN is retained, whether or not the DN's host phone is roaming. This ensures that the reachability characteristics established for the AAR destination number are retained.

  For outgoing AAR calls, the calling DN's AAR group uses the roaming device pool's AAR group instead of the AAR group selected on the DN's configuration page. Note that this AAR group will be applied to all DNs on the roaming device. For example, all DNs on a device roaming from New York to Paris (assuming both locations are in the same Device Mobility group) would inherit the AAR group configured for outgoing calls in the Paris device pool. This AAR group would be applied to all DNs on the roaming device and would allow for the appropriate prepending of prefix digits to AAR calls made from DNs on the roaming phone.

Call Forward All When Roaming

When a device is roaming in the same device mobility group, Unified CM uses the Device Mobility CSS to reach the local gateway. If a user sets Call Forward All at the phone, if the CFA CSS is set to None, and if the CFA CSS Activation Policy is set to With Activating Device/Line CSS, then:

- The Device CSS and Line CSS get used as the CFA CSS when the device is in its home location.
- If the device is roaming within the same device mobility group, the Device Mobility CSS from the Roaming Device Pool and the Line CSS get used as the CFA CSS.
- If the device is roaming within a different device mobility group, the Device CSS and Line CSS get used as the CFA CSS.

The section on Device Mobility, page 25-14, explains the details of this feature.

Extension Mobility

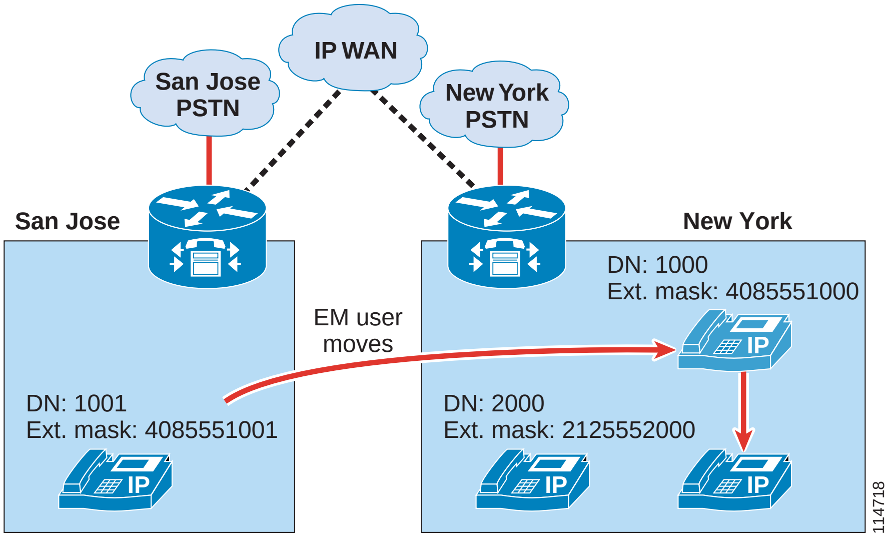
The Extension Mobility feature enables a user to log in to an IP phone and automatically apply his or her profile to that phone, including extension number, speed dials, message waiting indicator (MWI) status, and calling privileges. This mechanism relies on the creation of a device profile associated with each Extension Mobility user. The device profile is effectively a virtual IP phone on which you can configure one or more lines and define calling privileges, speed dials, and so on.

When an IP phone is in the logged-out state, (that is, no Extension Mobility user has logged into it), the phone characteristics are determined by the device configuration page and the line configuration page(s). When a user logs in to an IP phone, the device configuration does not change, but the existing line configuration is saved in the Unified CM database and is replaced by the line configuration of the user's device profile.

One of the key benefits of Extension Mobility is that users can be reached at their own extensions regardless of where they are located, provided that they can log in to an IP phone controlled by the same Unified CM cluster. When Extension Mobility is applied to multisite deployments with centralized call processing, this capability is extended to multiple sites geographically separated from each other.

However, if you combine the Extension Mobility feature with the AAR feature described in the section on Automated Alternate Routing, page 9-117, some limitations exist. Consider the example shown in Figure 9-50, where Extension Mobility and AAR are deployed in a centralized call processing Unified CM cluster with one site in San Jose and one in New York.

Figure 9-50 Extension Mobility and AAR



In this example, assume that an Extension Mobility user who is normally based in San Jose has a DN of 1000 and a DID number of (408) 555-1000. That user’s external phone number mask (or AAR mask, if used) is therefore configured as 4085551000. The user now moves to the New York site and logs in. Also, assume that the IP WAN bandwidth between San Jose and New York has been entirely utilized.

When the user in San Jose with extension 1001 tries to call 1000, AAR is triggered and, based on the AAR calling search space of the calling party and the AAR groups of both parties, a new call to 914085551000 is attempted by the San Jose phone. This call uses the San Jose gateway to access the

PSTN, but because the DID (408) 555-1000 is owned by that same gateway, the PSTN sends the call back to it. The San Jose gateway tries to complete the call to the phone with extension 1000, which is now in New York. Because no bandwidth is available to New York, the AAR feature is invoked again, and one of the following two scenarios will occur:

- If the gateway's AAR calling search space contains external PSTN route patterns, this is the beginning of a loop that eventually uses all the PSTN trunks at the San Jose site.
- If, on the other hand, the gateway's AAR calling search space contains only internal numbers, the call fails and the caller hears a fast-busy tone. In this case, one PSTN call is placed and one is received, so two PSTN trunks are utilized on the San Jose gateway for the duration of the call setup.

Tip To prevent routing loops such as the one described here, always configure all calling search spaces on the gateway configuration pages to include only internal destinations and no route patterns pointing to route lists or route groups containing that same gateway.

This example highlights the fact that Extension Mobility leverages the dynamic aspect of Cisco IP Communications and, therefore, requires that the call routing between sites use the IP network. Because the E.164 numbers defined in the PSTN are static and the PSTN network is unaware of the movements of the Extension Mobility users, the AAR feature, which relies on the PSTN for call routing, cannot be used to reach Extension Mobility users who move to a site other than their home site.

Note However, if the Extension Mobility user moves to a remote site that belongs to the same AAR group as his or her home site, he or she can use the AAR feature to place calls to other sites when the available IP WAN bandwidth is not sufficient. This is because the path of such a call is determined by the AAR calling search space of the phone from which the call originates. This AAR calling search space does not change when users log in or out of Extension Mobility, and it should be configured to use the visited remote site's gateway.

Tip Configure unregistered Extension Mobility profile DNs to send calls to voicemail. See Call-Forward Calling Search Spaces, page 9-113, for details.

Special Considerations for Cisco Unified Mobility

Cisco Unified Mobility (see the section on Cisco Unified Mobility, page 25-32) relies on functionality that has a direct impact on call routing. To understand the effects of the Cisco Unified Mobility parameters related to dial plans, consider the following example:

Note Only those parameters required in the discussion are mentioned here.

User Paul has an IP phone configured as follows:

DN: 8 555 1234

DID number: +1 408 555 1234

External Phone Number Mask: 408 555 1234

Line Calling Search Space: P_L_CSS

Device Calling Search Space: P_D_CSS

Paul's DN is associated with a Remote Destination Profile configured as follows:

Calling Search Space: P_RDP_CSS

Rerouting Calling Search Space: P_RDP_Rerouting_CSS

Calling Party Transformation CSS: P_CPT_CSS

Paul's RDP is associated with a Remote Destination configured as follows:

Destination Number: +1 514 000 9876 (This is Paul's mobile phone number, on either a single-mode or dual-mode phone.)

Calls from the PSTN placed to Paul or Ringo's DID number are handled by a gateway configured as follows:

Calling Search Space: GW_CSS

Significant digits: 7

Prefix DN: 8

User Ringo has an IP phone configured as follows:

DN: 8 555 0001

DID number: 408 555 0001

External Phone Number Mask: 408 555 0000 (This is the enterprise's main business number.)

Line Calling Search Space: R_L_CSS

Device Calling Search Space: R_D_CSS

The following sections explain the effects of the above mobility parameters on call routing.

Remote Destination Profile

Remote destination profiles (RDPs) are associated with directory numbers (for example, the DN of a user's IP phone) and with remote destinations (for example, the mobile phone number of a user). The RDP controls the interaction between the IP phone and the external numbers (for example, a mobile phone) configured as remote destinations.

Note Remote destinations cannot be configured with on-cluster DNs as destination numbers.

Remote Destination Profile's Rerouting Calling Search Space

When a call is placed to a DN associated with a remote destination profile, the call has the effect of ringing both the DN and the number(s) configured as remote destination(s).

The ability of the caller to reach the destination IP phone is controlled by the caller's Calling Search Space settings. However, the ability for the call to be forked toward the remote destination (for example, a mobile phone) is controlled by the called mobility user's Rerouting Calling Search Space.

For example:
Ringo calls Paul from his IP phone by dialing 8 555 1234. Paul's IP phone rings, as well as his mobile phone.

Here, the ability for Ringo to reach Paul's DN is controlled by the Line and Device calling search spaces on Ringo's IP phone. The dialed destination (8 555 1234) must be in a partition found in the concatenated calling search spaces R_L_CSS and R_D_CSS.

For this same call to be forked to ring Paul's mobile phone, the configured remote destination (+1 514 000 9876) must match a pattern found in the calling search space P_RDP_Rerouting_CSS.

Note Even if the dialing privileges assigned to Ringo's phone do not allow for external calls, the call to the remote destination is handled by the rerouting calling search space associated with Paul's remote destination profile.

Remote Destination Profile's Calling Search Space

In Cisco Unified CM 6.0, the RDP's calling search space is used to route calls originating from numbers defined as remote destinations. It is concatenated with the associated DN's Line CSS. The order of concatenation is Line CSS followed by the RDP's CSS.

When the calling party number of an external call made into the cluster matches a number defined as a remote destination, the calling party number is replaced with the DN of the line associated with the matching remote destination. Also, the calling search space used to route the call is the concatenation of:

- The Line calling search space of the DN associated with the matched remote destination number
- The calling search space of the RDP associated with the matched remote destination

In Unified CM 6.1 and later releases, a new service parameter (Inbound Calling Search Space for Remote Destination) controls which calling search space is used to route calls originating from one of the cluster's remote destinations. Its default setting is Trunk or Gateway Inbound Calling Search Space, which routes all incoming calls using the trunk’s or gateway's configured CSS. If the service parameter is set to Remote Destination Profile + Line Calling Search Space, then the behavior is identical for all Unified CM 6.*x* releases. This new service parameter has no effect on the calling party number replacement.

Note The default behavior of Unified CM 6.1 and later releases is different than the behavior of Unified CM 6.0 with regard to the routing of incoming calls originating from numbers defined as remote destinations. Cisco recommends using the default setting of Unified CM 6.1 because it simplifies call routing.

All the numbers defined as remote destinations within the same cluster will be searched to find a match for any external call coming into the cluster.

The following examples assume Unified CM 6.1 and later releases with the service parameter Inbound Calling Search Space for Remote Destination set to Trunk or Gateway Inbound Calling Search Space.

For example:
Paul uses his mobile phone to call Ringo at his desk. The call comes into the gateway from the PSTN, with a calling party number of 514 000 9876 and a called party number of 408 555 0001. The call is routed to Ringo's phone. The number displayed as the calling party number on Ringo's phone is Paul's desk phone number, 8 555 1234. This allows Paul's mobile phone number to remain confidential and allows Ringo's calls placed from the missed and received calls lists to ring into Paul's IP phone, thus making the full set of enterprise mobility features available.

When the call comes into the gateway, the PSTN offers a calling party number of 514 000 9876 and a called party of 408 555 0001. The gateway’s configuration will retain the last seven significant digits of the called number and prefix 8, yielding 8 555 0001 as the destination number.

The system detects that the calling party number matches Paul's remote destination number. Upon detecting this match, the system will:

1. Change the calling party number to Paul's DN, 8 555 1234.

2. Route the call to the called number using the incoming gateway's calling search space. Specifically, the routing is done through the GW_CSS calling search space.

The destination (called) number presented by the gateway should be the DN of the phone, and the calling party substitution illustrated in step 1 above renders possible the use of one-touch dialing from the missed/received calls lists.

Note There is no way to partition remote destination numbers. This is worth noting in case multiple user groups (such as different companies, sub-contractors, and so forth) are using the same cluster. In Unified CM 6.1 and later releases, when the service parameter Inbound Calling Search Space for Remote Destination is set to Trunk or Gateway Inbound Calling Search Space, the call routing is based on the incoming trunk’s or gateway's CSS, regardless of whether or not the calling number matches a remote destination. However, the calling party number substitution still occurs if the calling party matches any remote destination. This means that calls from one tenant's remote destination numbers to another tenant's DID numbers will be presented with a transformed calling party number that matches the caller's on-net extension DN.

Note Any incoming external call where Calling Party Number is not available will be routed according to the incoming gateway's CSS. This also applies to incoming calls from IP trunks, such as SIP or H.323 trunks.

Remote Destination Profile's Calling Party Transformation CSS and Transformation Patterns

Calls originating from an enterprise IP phone to a mobility-enabled DN are forked to both the enterprise destination IP phone's DN and one (or multiple) external destinations. One challenge this creates is to deliver calling party numbers adapted to each destination phone's dial plan. This is to allow for redialing of calls from missed calls and received calls lists. For an enterprise phone, the calling party numbers should be redialable enterprise phone numbers. For a remote destination on the PSTN (such as a home phone or a mobile phone), the calling party number should be transformed from the enterprise number associated with the calling IP phone to a number redialable from the PSTN (generally, the DID number of the calling phone).

When a call is placed to a mobility-enabled enterprise DN, the associated remote destination profile's calling party transformation calling search space is used to find a match to the caller's calling party number. It contains partitions which themselves contain transformation patterns.

Transformation patterns control the adaptation of calling party numbers from enterprise format to PSTN format. They differ from all other patterns in Unified CM in that they match on the calling number, not the called number. The matching process is done through a regular expression (for example, 8 555 XXXX), and the transformation process allows for the optional use of the calling DN's external phone number mask as well as transformation patterns and digit prefixing.

Once matched, they perform all configured transformations, and the resulting calling party number is used to reach all remote destinations associated with the Remote Destination Profile for which the match occurred.

For example:
When Ringo calls Paul, we want Paul's IP phone to display the calling party number as 8 555 0001 and Paul's mobile phone to display 408 555 0001.

For this case, we create a transformation pattern with the following parameters:

Pattern: 8 555 XXXX

Partition: SJ_Calling_Transform

Use calling party's external phone number mask: un-checked

Calling Party Transformation mask: 555 XXXX

Prefix Digits (outgoing calls): 408

We also have to ensure that partition SJ_Calling_Transform is placed in calling search space P_CPT_CSS.

When the call from Ringo is anchored on Paul's phone, two separate call legs are attempted. The first rings Paul's IP phone and offers the caller's DN as Calling Party Number (that is, 8 555 0001). The second call leg is attempted through Paul's Remote Destination Profile. The RDP's calling party transformation CSS, P_CPT_CSS, is used to find a match for 8 555 0001 in all the referenced partition's transformation patterns. Pattern 8 555 XXXX is matched in partition SJ_Calling_Transform. The transformation mask is applied to the calling party number and yields 555 0001. The prefix digits are added, and the resulting calling party number 408 555 0001 is used when placing the call to the remote destinations.

Note that, in this example, we chose not to use the external phone number mask because it is set to a number different than that of Ringo's DID. This offers flexibility in situations where the calling party number offered to off-net destinations is required to be different based on the relationship of the caller to the called party. The call from Ringo to Paul is between co-workers, thus the disclosure of Ringo's DID number is deemed acceptable. Ringo's next call could be to a customer, in which case the main enterprise number 408 555 0000 is the desired Calling Party Number to be offered to the destination.

Note Calling Party Transformation calling search spaces do not implicitly include the <none> partition; therefore, transformation patterns left in the <none> partition do not apply to any Calling Party Transformation calling search space. This is different from all other patterns in Unified CM, where all patterns left in the <none> partition are implicitly part of every calling search space.

Application Dial Rules

Numbers defined as remote destinations are also used to identify and anchor incoming calls as enterprise mobility calls. Often, the form in which the PSTN identifies calls differs from the form in which an enterprise dial plan requires that calls to external numbers be dialed. Application dial rules can be used to adapt the form in which remote destinations are configured to the form required when forking a call to the remote destination. They allow for the removal from, and prefixing of digits to, the numbers configured as remote destinations.

For example:
Assume the number 514 000 9876 is configured as Paul's remote destination number. This corresponds to the form used by the PSTN to identify calls coming into the enterprise. But it differs from the form used by the enterprise dial plan for outgoing calls, which requires that 91 be prefixed. In this case, we need to create an application dial rule to adapt the remote destination form to the enterprise dial plan's form:

Application Dial Rule:

Name: 514000_ten

Description: Used to prefix 91 to ten-digit numbers beginning with 514000

Number begins with: 514000

Number of Digits: 10

Total digits to be removed: 0

Prefix with Pattern: 91

In this example, calls made from Paul's mobile phone into the enterprise are identified as coming from 514 000 9876. This matches the form in which his number is configured as a remote destination, thus allowing the match to be made and triggering the anchoring of the call on Paul's desk phone as well as adapting the Calling Party Number offered to the on-net destination. (For example, when a call is placed to Ringo's DID number, he sees the call as coming from 8 555 1234.)

When a call is placed to Paul's enterprise DN number, the call leg forked to his remote destination number will be processed by the application dial rule above. The string 514 000 matches the beginning of Paul's remote destination number, and it is ten digits long, so no digits are removed and 91 is prefixed. This yields 91 514 000 9876 as a number to be routed through Paul's Remote Destination Profile calling search space (P_RDP_CSS in this case).

Note This approach offers the ability to reuse calling search spaces already defined to route calls made from IP phones. Creating new calling search spaces not requiring prefixes for outbound calls (that is, ones able to route calls to 514 000 9876 directly) is less preferable because it can create situations where external patterns overlap with on-net patterns.

Immediate Divert (iDivert)

The Immediate Divert (iDivert) function is used to send calls directly to voicemail. It can be invoked when the call is ringing (incoming), when a call is on hold, or when a call is connected. The iDivert function allows incoming calls to be diverted to either the invoking phone's voicemail box or the voicemail box of the originally called party. The enhanced functionality is applicable only to diverted calls such as forwarded calls or calls redirected by an application.

iDivert Enhancements in Cisco Unified CM 5.1

The iDivert function has been augmented in Cisco Unified CM 5.1 to allow incoming calls to be diverted to either the invoking phone's voicemail box (legacy behavior) or the voicemail box of the originally called party (enhanced behavior). The enhanced functionality is applicable only to diverted calls such as forwarded calls or calls redirected by an application.

For example:

> Assume that phone A calls phone B, whose calls are forwarded to phone C. As phone C is ringing, the user at phone C activates the iDivert softkey, which offers two choices. The first choice results in the call being sent to the original called party's voicemail (in this case, phone B's voicemail box), while the second choice results in the call being sent to the iDivert invoker's voicemail (in this case, phone C's voicemail box). The same choices are available whether phone C invokes the feature while the call is ringing, connected, or placed on hold.

If a call is handled by Auto Call Pickup, Call Transfer, Call Park, Call Park Reversion, Conference, or a MeetMe Conference prior to the invocation of the iDivert function, the call is no longer considered to be a "diverted" call, and the only iDivert functionality available in this case is the legacy iDivert behavior (that is, sending the call to the invoker’s voicemail). For example, assume phone A calls phone B, whose calls are forwarded to phone C, and then phone C transfers the call to phone D. This is *not* a diverted call because the last action applied to the call was the transfer to phone D. If phone D invokes the iDivert function, the call will be sent to phone D's voicemail box.

To enable the full iDivert functionality described above, set the Unified CM service parameter **Use Legacy Immediate Divert** to **False**. When enabled, enhanced iDivert automatically allows the use of the feature over QSIG trunks, thus allowing an invoker's voicemail box to be hosted in a telephony system connected via QSIG.

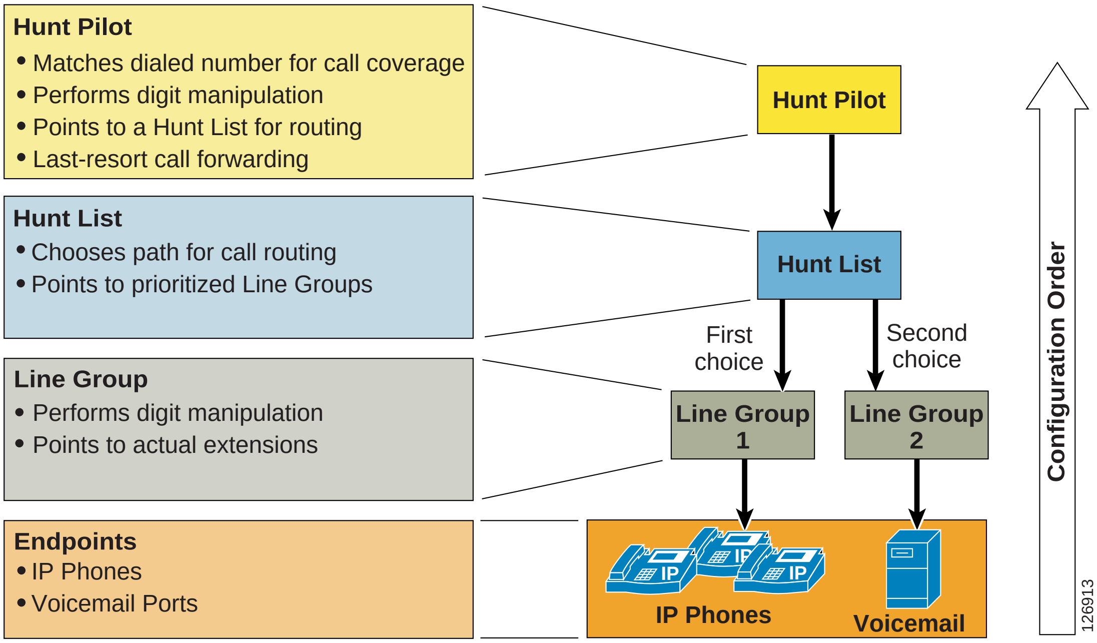
In cases where iDivert is used in a cluster connected to other telephone systems using QSIG, there might be situations in which only the legacy iDivert functionality (where the only available choice is to send the call to the invoker's voicemail) is offered to a phone when receiving a call. For instance, assume phones A and B are in cluster 1, and phone X is another QSIG-connected telephony system. Phone A calls phone X, which is call-forwarded to phone C. After the call is connected to phone C, iDivert will offer both the legacy (invoker's voicemail) and enhanced (original called party's voicemail) destinations only if QSIG path replacement has not occurred. If phone C invokes iDivert after QSIG path replacement, the only destination available is phone C's voicemail box.

Hunt Lists and Line Groups

The *hunt pilot* is typically used for call coverage, or distributing a call through a list of endpoints. For call distribution, you can use a hunt construct. This hunt construct is based upon a three-tiered architecture, similar to that used to route external calls, that allows for multiple layers of call routing as well as digit manipulation.

Unified CM searches for a configured hunt pilot that matches an incoming called number and uses it to select a corresponding hunt list, which is a prioritized list of the available paths for the call. These paths are known as *line groups*. Figure 9-51 depicts the three-tiered architecture of the hunt construct in Unified CM.

Figure 9-51 Three-Tiered Architecture for the Hunt Construct in Unified CM


Hunt Pilot
• Matches dialed number for call coverage
• Performs digit manipulation
• Points to a Hunt List for routing
• Last-resort call forwarding
Hunt List
• Chooses path for call routing
• Points to prioritized Line Groups
Line Group
• Performs digit manipulation
• Points to actual extensions
Endpoints
• IP Phones
• Voicemail Ports
Hunt Pilot
Hunt List
First choice
Second choice
Line Group 1
Line Group 2
IP
IP Phones
Voicemail
Configuration Order
126913

Hunt Pilot

Hunt pilots are strings of digits and wildcards similar to route patterns, such as 9.[2-9]XXXXXX, configured in Unified CM to route calls to directory numbers. The hunt pilot points directly to a hunt list. Hunt lists point to line groups, which finally point to SCCP endpoints.

Calls can be redirected to a final destination when the hunting fails because of one or both of the following reasons:

- All hunting options have been exhausted and the call still is not answered.
- A time-out period has expired.

This call redirection is configured in the Hunt Forward Settings section of the Hunt Pilot configuration page, and the destination for this redirect can be either of the following options:

- A specific pattern in the internal call routing table of Unified CM
- Personal preferences, which point to the Call Forward No Coverage settings for the originally called number when hunting on behalf of that number fails

For example, you can implement the personal preferences option by configuring a user's phone so that the Forward No Answer field redirects the call to a hunt pilot, in order to search for someone else who can answer the call. If the call hunting fails, either because all the hunting options were exhausted or because a time-out period expired, the call can be sent to a destination personalized for the person who was originally called. For example, if you set the Forward No Coverage field within the person's DN configuration page to the voicemail number, the call will be sent to that person's voicemail box if hunting fails.

The hunt pilot can distribute calls to any of its line group members, even if the members and the hunt pilot reside in different partitions. A call distributed by the hunt pilot overrides all the partitions and calling search space restrictions.

Hunt List

A hunt list is a prioritized list of eligible paths (line groups) for call coverage. Hunt lists have the following characteristics:

- Multiple hunt pilots may point to the same hunt list.
- A hunt list is a prioritized list of line groups that function as alternate sets of phones which are offered a call placed to the hunt pilot number. For example, you can use a hunt list to attempt to find a taker for the call within a set of phones at a particular site. If the call is not taken, then the hunt list attempts to offer the call through a second line group pointing to phones at a second site.
- Hunt lists do not do any digit manipulation.
- Multiple hunt lists can contain the same line group.

Line Group

Line group members are user extension numbers that are controlled by Unified CM. Thus, when the call is being distributed through the line group members, Unified CM is in control of the call. Hunt options can be applied to the call when it is not answered or if the extension is busy or unregistered.

Line groups control the order in which the call is distributed, and they have the following characteristics:

- Line groups point to specific extensions, which are typically IP phone extensions or voicemail ports.
- A single extension may be present in multiple line groups.
- Computer Telephony Integration (CTI) ports and CTI route points cannot be added within a line group. Therefore, calls cannot be distributed to endpoints controlled through CTI applications such as Cisco Customer Response Solution (CRS), IP Interactive Voice Response (IP IVR), and so forth

- Unified CM distributes calls to the devices according to the distribution algorithm assigned. Unified CM supports the following algorithms:
  - Top-down
  - Circular
  - Longest idle time
  - Broadcast
- In the event of No-Answer, Busy, or Not-Available, line groups redirect a call distributed to an extension based on the hunt options. Unified CM supports the following hunt options:
  - Try next member, then try next group in hunt list.
  - Try next member, but do not go to next group.
  - Skip remaining members and go directly to next group.
  - Stop hunting.

Hunt Group Logout

A user can log out of a hunt group by activating the HLog softkey. Once activated, this function effectively makes all lines configured on the phone act as though they are not part of any hunt group. The phone displays "Logged out of Hunt Group." If a line group contains a shared line, all instances of the shared line that are on devices in the logged-out state will not ring; conversely, all instances of the shared line on devices in the logged-in state will ring.

Lines that are not part of any hunt groups will continue to ring normally, no matter the state of the HLog function.

The HLog function can be activated from Unified CM Administration. By default, the HLog softkey is not configured on the softkey templates. Once added to a phone's softkey template, the HLog button appears in the display when the phone is in the connected, off-hook, or on-hook state.

The Hunt Group Logoff Notification service parameter provides the option of audible ring tones when calls that come in from a line group arrive at the phone in the logged-off state. The Hunt Group Logoff Notification service parameter is in the Clusterwide Parameters (Device - Phone) section of the Service Parameters Configuration page. For enabling the function, ensure that a valid ring tone file is present on the TFTP server. If an invalid file name is provided, no tone will be played

For more information about hunt algorithms and hunt options, refer to the *Cisco Unified Communications Manager Administration Guide*, available at

http://www.cisco.com

Line Group Devices

The line group devices are the endpoints accessed by line groups, and they can be of any of the following types:

- Any Skinny Client Control Protocol (SCCP) endpoints, such as Cisco Unified IP Phones
- SIP endpoints
- Voicemail ports (for Cisco Unity)
- H.323 clients
- FXS extensions attached to an MGCP gateway

Time-of-Day Routing

To use this feature, configure the following elements:

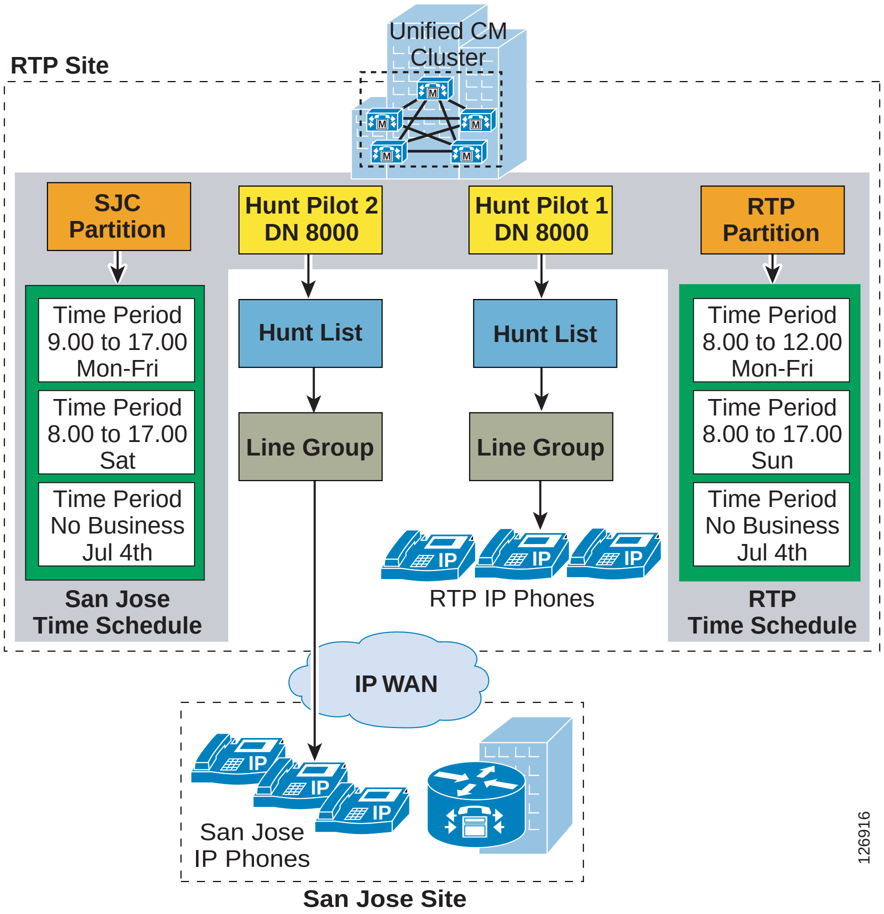
- Time period
- Time schedule

The time period allows you to configure start and end times for business hours. The start and end times indicate the times during which the calls can be routed. In addition to these times, you can set the event to repeat itself on a weekly or yearly basis. Moreover, you can also configure non-business hours by selecting "No business hours" from the Start Time and End Time options. All incoming calls will be blocked when this option is selected.

A time schedule is a group of specific time periods assigned to the partition. It determines whether the partition is active or inactive during the specified time periods. A matching/dialing pattern can be reached only if the partition in which the dialing pattern resides is active.

As illustrated in Figure 9-52, two hunt pilots with the same calling pattern (8000) are configured in two partitions (namely, RTP_Partition and SJC_Partition). Each of these partitions is assigned a time schedule, which contains a list of defined time periods. For example, RTP phones can be reached using Hunt Pilot 1 from 8:00 AM to 12:00 PM EST (GMT - 5.00) Monday through Friday as well as 8:00 AM to 5:00 PM on Sundays. In the same way, SJC phones can be reached using Hunt Pilot 2 from 8:00 AM to 5:00 PM PST (GMT - 8.00) Monday through Friday and 8:00 AM to 5:00 PM on Saturdays. Both of the hunt pilots in this example are inactive on July 4th.

Figure 9-52 Time-of-Day Routing



For the example in Figure 9-52, an incoming call to the hunt pilot (8000) on Wednesday at 3:00 PM will be forwarded to the SJC phones, while a person calling the hunt pilot on July 4th will get a fast busy tone unless there is another pattern that matches 8000.

Logical Partitioning

The elements of logical partitioning include:

- Device types, where phones are classified as *interior*, and gateways and trunks are defined as *border*. Table 9-9 lists the endpoint types for different devices.
- Geolocations, where endpoints are assigned a civic address to be used in policy decisions.
- Geolocation filters, where policy decisions can be made on a subset of the geolocation objects.
- Policies, where communications between endpoints are either allowed or denied based on their comparative (filtered) geolocations and device types.

Note Policies are not applied if all participants in a call (or call attempt) are classified as *interior*. This means that calls between phones on the same cluster are never subjected to logical partitioning policies.

Note Geolocations are not to be confused with locations configured in Unified CM, which are used for call admission control, or with physical locations used for Device Mobility.

Table 9-9 Device Types

Logical Partitioning Device Types	Cisco Unified Communications Manager Device
Border	• Gateway (for example, H.323 gateway) • Inter-luster trunk (ICT), both gatekeeper-controlled and non-gatekeeper-controlled • H.225 trunk • SIP trunk • MGCP port (E1, T1, PRI, BRI, FXO)
Interior	• Phones (SCCP, SIP, or third-party) • CTI route points • VG224 analog phones • MGCP port (FXS) • Cisco Unity voicemail (SCCP)

Logical Partitioning Device Types

Unified CM classifies endpoints as either *interior* or *border*. This classification is fixed and cannot be modified by the system administrator.

Geolocation Creation

The (RFC) 4119 standard provides the basis for geolocations. Geolocations use the civic location format specified by the following objects:

- Name
- Description
- Country using the two-letter abbreviation
- State, Region, or Province (A1)
- County or Parish (A2)
- City or Township (A3)
- Borough or City District (A4)
- Neighborhood (A5)
- Street (A6)
- Leading Street Direction, such as N or W (PRD)
- Trailing Street Suffix, such as SW (POD)
- Address Suffix, such as Avenue, Platz (STS)
- Numeric house number (HNO)

- House Number Suffix, such as A, 1/2 (HNS)
- Landmark (LMK)
- Additional Location Information, such as Room Number (LOC)
- Floor (FLR)
- Name of Business or Resident (NAM)
- Zip or Postal Code (PC)

Note In Unified CM, you must define geolocations manually.

Geolocation Assignment

Devices are assigned a geolocation from either the device page, the device pool, or the default Geolocation as configured under Enterprise Parameters, in that order of precedence.

Geolocation Filter Creation

Geolocation filters define which of the geolocation objects should be used when comparing the geolocations of different endpoints. For example, a group of phones may be assigned identical geolocations, except for the room and floor in which they are located. Policies may want to consider endpoints located within the same building as being within the same Closed User Group, and thus allowed to communicate. Even though the actual geolocations of each phone differ, the filtered geolocation is the same. This is useful when policies need to be applied to only the top-level fields of geolocation. For instance, a policy that denies communications between phones and gateways in different cities but allows communications between phones and gateways in the same city, could be based on the comparative filtered geolocations where objects more granular than the City are ignored.

Geolocation Filter Assignment

Phones inherit the filter assignment of their device pool. Gateways and trunks can be configured with a geolocation filter at the device or device pool level, in that order of precedence.

Logical Partitioning Policy Configuration

Logical partitioning policies are configured between geolocation identifiers. A geolocation identifier is the combination of a filtered geolocation and a device type. The filtered geolocation is obtained by taking a device's geolocation and applying the device's associated geolocation filter.

A policy is created as the combination of a set of geolocation objects and a device type (a source geolocation identifier) in relationship with another such combination (the target geolocation identifier). When the relationship is matched, the configured action of "allow" or "deny" is applied to the call leg.

Note Each set of geolocation objects configured in a policy is considered in association with a single device type. For example, a set of geolocation objects such as Country=India, State=Karnataka, City=Bangalore needs to be associated with device type Interior for actions pertaining to Bangalore phones, and separately associated with device type Border for actions pertaining to Bangalore gateways.

Logical Partitioning Policy Application

When user action results in the creation of a new call leg (for example, when a user conferences a third caller into a preexisting call), Unified CM will match the geolocation identifiers of each participant pairs to those of preconfigured policies.

Note When the geolocation identifiers of two devices are being evaluated by logical partitioning, no policy is applied if both devices are of device type Interior. This means that no call, conference, transfer, or so forth, between IP phones within the same cluster will ever be denied due to logical partitioning policies.

For example, consider phones A and B located in Bangalore, India, and gateway C located in Ottawa, Canada. Phone A calls phone B. Because both devices are of type Interior, no policy is invoked. The call is established, and then the user at phone A invokes a conference, which would bring in gateway C. Before the action is allowed, Unified CM will check the geolocation identifiers of A and C, as well as those of B and C, for a match with the preconfigured policies. If any of the matching policies results in a deny action, the new call leg cannot be established.

Note The default policy in Unified CM is deny; in other words, if no policy is configured explicitly to permit a call leg, the call leg will be denied.

In the example above, unless an explicit policy is configured to allow Bangalore Interior devices to connect to Ottawa Border devices, the call leg will be denied.

Call Routing in Cisco IOS with H.323 Dial Peers

The call routing logic on Cisco IOS routers using the H.323 protocol relies on the dial peer construct. Dial peers are similar to static routes; they define where calls originate and terminate and what path the calls take through the network. Dial peers are used to identify call source and destination endpoints and to define the characteristics applied to each call leg in the call connection. Attributes within the dial peer determine which dialed digits the router collects and forwards to telephony devices.

For a detailed description of dial peers and their configuration, refer to *Configuring Dial Plans, Dial Peers, and Digit Manipulation,* part of the *Cisco IOS Voice, Video, and Fax Configuration Guide, Release 12.2*, available at

http://www.cisco.com

One of the keys to understanding call routing with dial peers is the concept of incoming versus outgoing call legs and, consequently, of incoming versus outgoing dial peers. Each call passing through the Cisco IOS router is considered to have two call legs, one entering the router and one exiting the router. The call leg entering the router is the *incoming call leg*, while the call leg exiting the router is the *outgoing call leg*.

Call legs can be of two main types:

- Traditional TDM telephony call legs, connecting the router to the PSTN, analog phones, or PBXs
- IP call legs, connecting the router to other gateways, gatekeepers, or Unified CMs

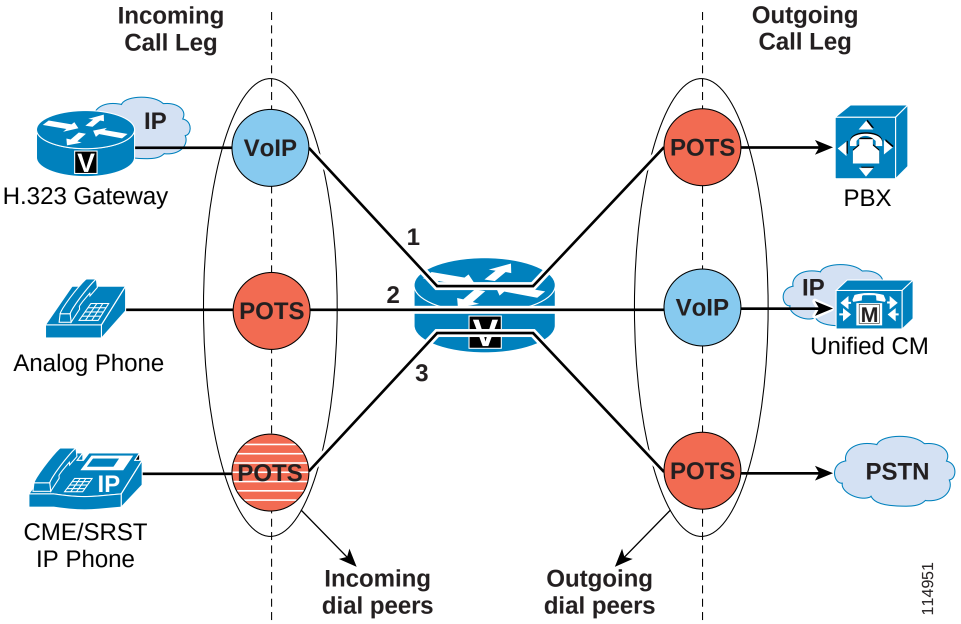
For all calls going through the router, Cisco IOS associates one dial peer to each call leg. Dial peers are also of two main types, according to the type of call leg with which they are associated:

- POTS dial peers, associated with traditional TDM telephony call legs
- VoIP dial peers, associated with IP call legs

Figure 9-53 shows the following examples of different types of calls going through a Cisco IOS router:

- Call 1 is from another H.323 gateway across the IP network to a traditional PBX connected to the router (for example, via a PRI interface). For this call, an incoming VoIP dial peer and an outgoing POTS dial peer are selected.
- Call 2 is from an analog phone connected to an FXS port on the router to a Unified CM cluster across the IP network. For this call, an incoming POTS dial peer and an outgoing VoIP dial peer are selected by the router.
- Call 3 is from an IP phone controlled by Cisco Unified Communications Manager Express (Unified CME) or SRST to a PSTN interface on the router (for example, a PRI interface). For this call, an automatically generated POTS dial peer (corresponding to the **ephone** configured on the router) and an outgoing POTS dial peer are selected.

Figure 9-53 Incoming and Outgoing Dial Peers



To match incoming call legs to incoming dial peers, the router selects a dial peer by matching the information elements in the setup message (called number/DNIS and calling number/ANI) with four configurable dial peer attributes. The router attempts to match these items in the following order:

1. Called number with **incoming called-number**
2. Calling number with **answer-address**
3. Called number with **destination-pattern**
4. Incoming voice port with configured voice port

The router must match only one of these conditions. It is not necessary for all the attributes to be configured in the dial peer or that every attribute match the call setup information; only one condition must be met for the router to select a dial peer. The router stops searching as soon as one dial peer is matched, and the call is routed according to the configured dial peer attributes. Even if there are other dial peers that would match, only the first match is used.

How the router selects an outbound dial peer depends on whether **direct-inward-dial** (DID) is configured in the inbound POTS dial peer:

- If DID is not configured in the inbound POTS dial peer, the router performs two-stage dialing and collects the incoming dialed string digit-by-digit. As soon as one dial peer matches the destination pattern, the router immediately places the call using the configured attributes in the matching dial peer.
- If DID is configured in the inbound POTS dial peer, the router uses the full incoming dial string to match the destination pattern in the outbound dial peer. With DID, the setup message contains all the digits necessary to route the call, so no additional digit collection is required. If more than one dial peer matches the dial string, all of the matching dial peers are used to form a *hunt group*. The router attempts to place the outbound call leg using all of the dial peers in the hunt group until one is successful.

By default, dial peers in a hunt group are selected according to the following criteria, in the order listed:

1. Longest match in phone number

   This method selects the destination pattern that matches the greatest number of dialed digits. For example, if one dial peer is configured with a dial string of 345.... and a second dial peer is configured with 3456789, the router would first select 3456789 because it has the longest explicit match of the two dial peers.

2. Explicit preference

   This method uses the priority configured with the **preference** dial peer command. The lower the preference number, the higher the priority. The highest priority is given to the dial peer with preference order 0. If the same preference is defined in multiple dial peers with the same destination pattern, a dial peer is selected randomly.

3. Random selection

   In this method, all destination patterns are weighted equally.

You can change this default selection order or choose different methods for hunting dial peers by using the **dial-peer hunt** global configuration command. An additional selection criterion is *least recent use*, which selects the destination pattern that has waited the longest since being selected (equivalent to *longest idle* for Unified CM line groups).

Observe the following best practices when configuring H.323 dial peers on a Cisco IOS router:

- To ensure that incoming PSTN calls are directly routed to their destination based on the DNIS information, create a default POTS dial peer with the **direct-inward-dial** attribute, as follows:

```
dial-peer voice 999 pots
  incoming called-number .
  direct-inward-dial
  port 1/0:23
```

- When using the router as an H.323 gateway connected to a Unified CM cluster, provide redundancy by configuring at least two VoIP dial peers with the same destination pattern pointing to two different Unified CM servers. Use the **preference** attribute to select the priority order between primary and secondary Unified CM servers. The following example shows the use of the **preference** attribute:

```
dial-peer voice 100 voip
 preference 1

!--- Make this the first choice dial peer.

 ip precedence 5
 destination-pattern 1...
 session target ipv4:10.10.10.2

!--- This is the address of the primary Unified CM.

 dtmf-relay h245-alpha


dial-peer voice 101 voip
 preference 2

!--- This  is the second choice.

 ip precedence 5
 destination-pattern 1...
 session target ipv4:10.10.10.3

!--- This is the address of the secondary Unified CM.

 dtmf-relay h245-alpha
```

Call Routing in Cisco IOS with a Gatekeeper

An H.323 gatekeeper is an optional node that manages endpoints (such as H.323 terminals, gateways, and Multipoint Control Units (MCUs), as well as Cisco Unified Communications Manager Express (Unified CME) and Unified CM clusters) in an H.323 network, providing them with call routing and call admission control functions. The endpoints communicate with the gatekeeper using the H.323 Registration Admission Status (RAS) protocol.

Endpoints attempt to register with a gatekeeper on startup. When they want to communicate with another endpoint, they request admission to initiate a call using a symbolic alias for the endpoint, such as an E.164 address or an email address. If the gatekeeper decides that the call can proceed, it returns a destination IP address to the originating endpoint. This IP address might not be the actual address of the destination endpoint, but instead it might be an intermediate address, such as the address of a Cisco Unified Border Element or a gatekeeper that routes call signaling.

For more details about the H.323 protocol and the message exchange between H.323 endpoints and gatekeepers, refer to the *Cisco IOS H.323 Configuration Guide*, available at

http://www.cisco.com

The gatekeeper feature is supported on a number of router platforms. For the full list of supported platforms, consult the gatekeeper product data sheet. You can configure Cisco IOS gatekeepers in a number of different ways for redundancy, load balancing, and hierarchical call routing. This section focuses on the call routing capabilities of the gatekeeper feature. For redundancy and scalability considerations, refer to Gatekeeper Redundancy, page 8-38; for call admission control considerations, refer to Cisco IOS Gatekeeper Zones, page 11-40.

Call routing in Cisco IOS gatekeeper is based on the following types of information:

- Statically configured information, such as zone prefixes and default technology prefixes
- Dynamic information, such as E.164 addresses and technology prefixes provided by H.323 devices during the registration phase
- Per-call information, such as called number and technology prefix

A zone is a collection of H.323 devices (such as endpoints, gateways, or MCUs) that register with a gatekeeper. There can be only one active gatekeeper per zone, and you can define up to 100 local zones on a single gatekeeper.

When an H.323 endpoint registers with the gatekeeper, it is assigned to a zone and it can optionally register one or more E.164 addresses for which it is responsible, as well as a technology prefix that specifies which kinds of calls it can handle (for example, voice, video, fax, and so on).

For each zone, you can configure one or more *zone prefixes* on the gatekeeper. Zone prefixes are strings that contain digits and wildcards and are used by the gatekeeper to facilitate call routing decisions. The following characters are allowed in a zone prefix string:

- All numbers between 0 and 9, which match a single specific digit
- The dot (.) wildcard, which matches one digit between 0 and 9
- The * wildcard, which matches one or more digits between 0 and 9

To understand the gatekeeper call routing behavior, it is helpful to consider the message parsing logic. Figure 9-54 illustrates the parsing logic for an Admission Request (ARQ). To initiate a call, an endpoint sends an Admission Request (ARQ) to the gatekeeper. The ARQ contains either an H.323 ID or the E.164 address of the destination, or called party, as well as the E.164 address or H.323 ID of the source, or calling party.

If the ARQ contains the E.164 address (with Unified CM, the ARQ always contains an E.164 address), the ARQ may or may not contain a technology prefix. If the ARQ contains a technology prefix, the gatekeeper strips it from the called number. If the ARQ does not contain a technology prefix, the gatekeeper uses the default technology prefix if one is configured (see the **gw-type-prefix** command in the section on Centralized Gatekeeper Configuration, page 9-145). The technology prefix thus obtained is stored in memory, and the gatekeeper continues with the call routing algorithm.

Next, the gatekeeper tries to match the called number with one of the configured zone prefixes. Longest-match is used if multiple potential matches exist. If no zone prefix can be matched, and if the gatekeeper is configured to accept calls with an unknown prefix, the gatekeeper then assumes that the destination zone is equal to the source zone.

At this point, the gatekeeper searches in the chosen destination zone for a registered E.164 address that matches the called number. If there is a match, the gatekeeper can send an Admission Confirm (ACF), provided that the requested bandwidth for the call is available and that the called endpoint is registered with the gatekeeper. The ACF will contain the IP address of the destination endpoint. If the bandwidth is unavailable or the called endpoint is not registered, the gatekeeper returns an Admission Reject (ARJ) to the calling endpoint.

If there is no matching E.164 address registered in the destination zone, the gatekeeper will use the previously stored technology prefix to choose a gateway registered in that zone as the destination for the call. The same considerations regarding bandwidth availability and endpoint registration dictate whether the gatekeeper will send an ACF or an ARJ to the calling endpoint.

Upon receipt of an ACF from the gatekeeper, the source endpoint can send a setup message directly to the destination endpoint by using the IP address returned in the ACF.

Figure 9-54 Gatekeeper Address Resolution for an ARQ

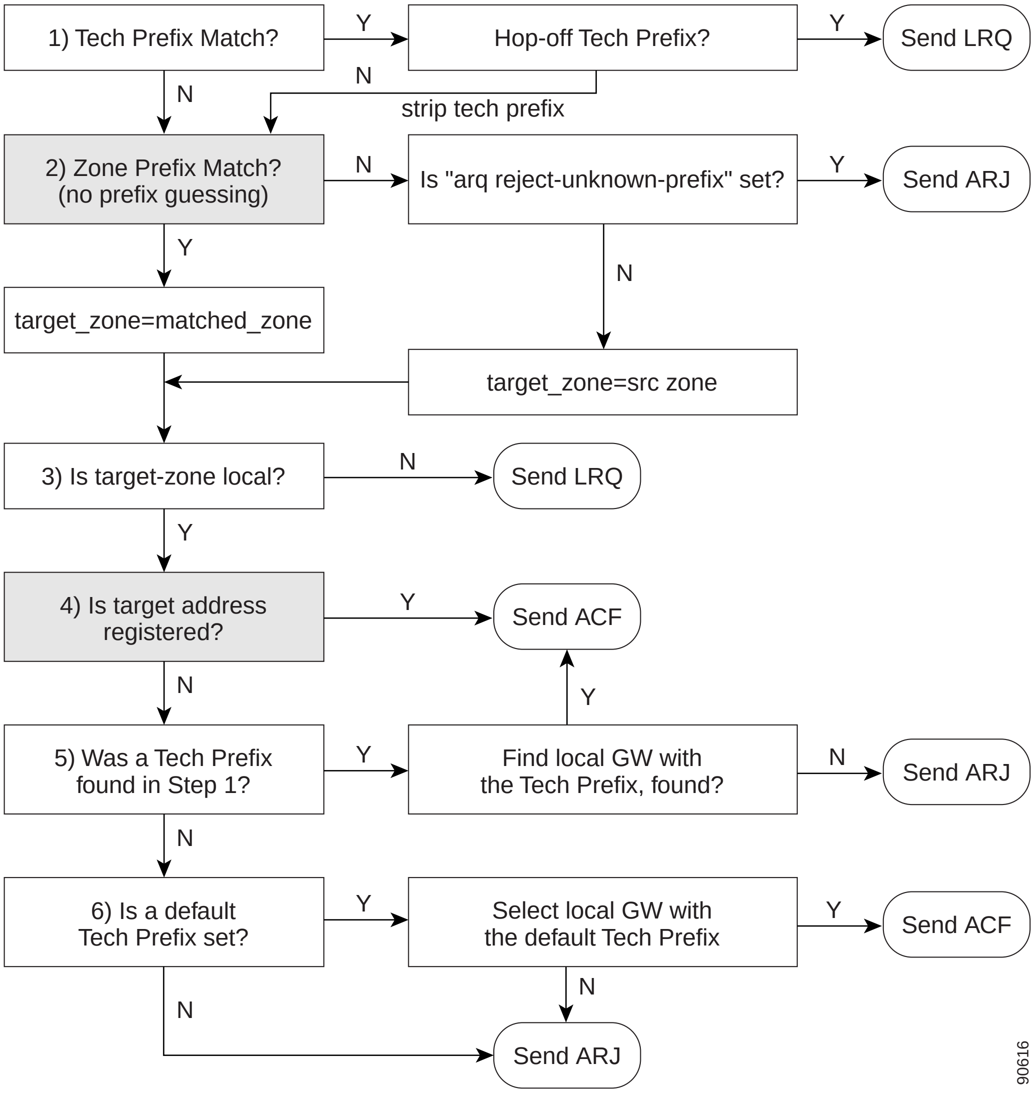
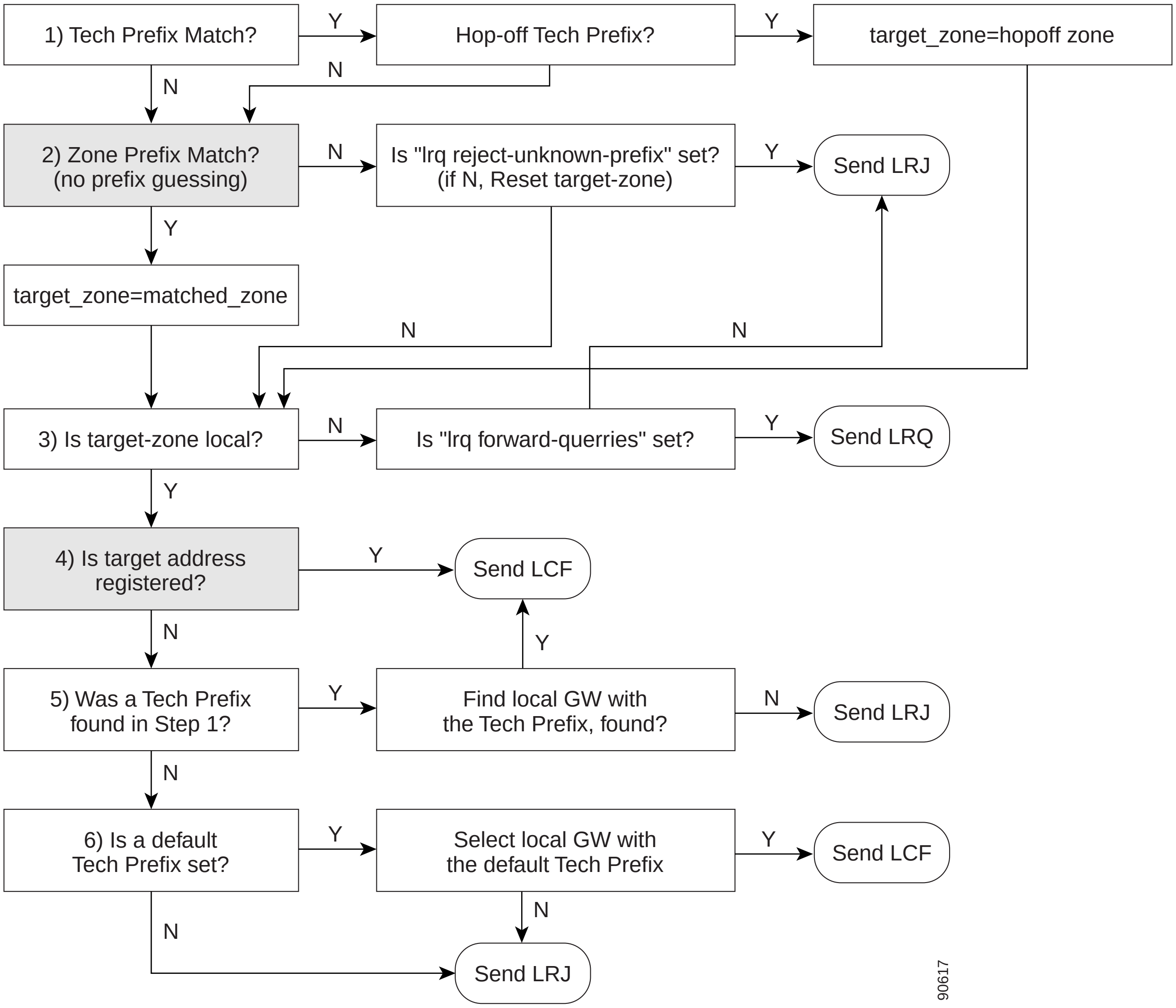


Figure 9-55 illustrates the parsing logic for a Location Request (LRQ). LRQ messages are exchanged between gatekeepers and are used for inter-zone (remote zone) calls. For example, gatekeeper A receives an ARQ from a local zone gateway requesting call admission for a remote zone device. Gatekeeper A then sends an LRQ message to gatekeeper B. Gatekeeper B replies to the LRQ message with either a Location Confirm (LCF) or Location Reject (LRJ) message, depending on whether it is configured to admit or reject the inter-zone call request and whether the requested resource is registered.

Figure 9-55 Gatekeeper Address Resolution for an LRQ



Traditional Cisco IOS gatekeeper functionality has been extended to accommodate for Cisco Unified Border Elements through the concept of a *via-zone gatekeeper*.

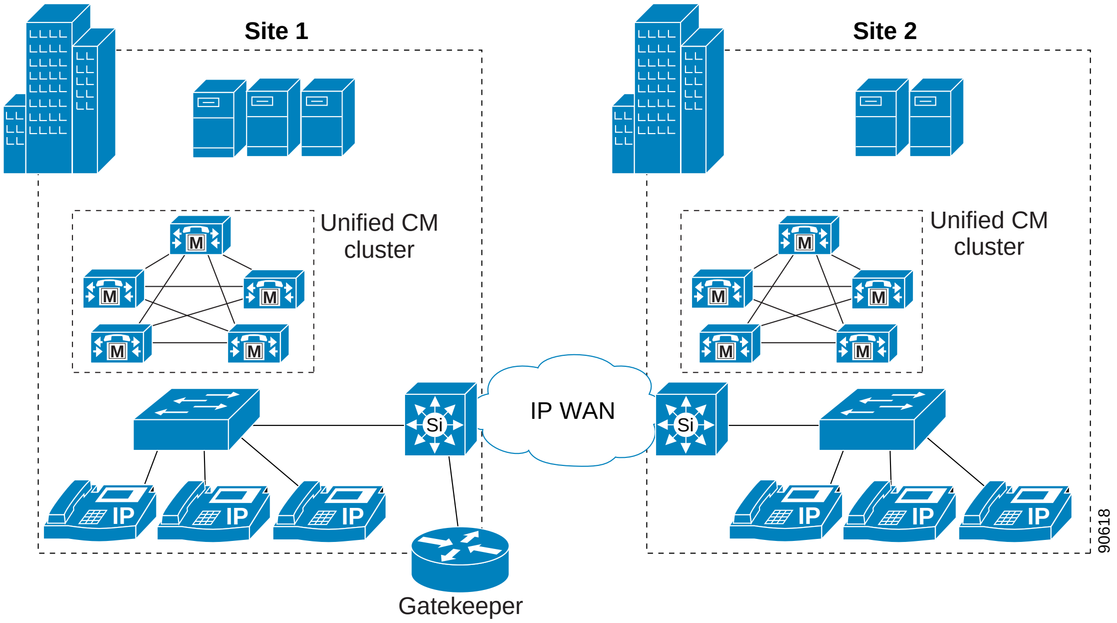
A via-zone gatekeeper differs from legacy gatekeepers in how it uses LRQ and ARQ messages for call routing. Using via-zone gatekeepers will maintain normal clusters and functionality. Legacy gatekeepers examine incoming LRQs based on the called number, and more specifically the dialedDigits field in the destinationInfo portion of the LRQ. Via-zone gatekeepers look at the origination point of the LRQ before looking at the called number. If an LRQ comes from a gatekeeper listed in the via-zone gatekeeper's remote zone configurations, the gatekeeper checks to see that the zone remote configuration contains an **invia** or **outvia** keyword. If the configuration contains either of these keywords, the gatekeeper uses the new via-zone behavior; if not, it uses legacy behavior.

For ARQ messages, the gatekeeper determines if an **outvia** keyword is configured on the destination zone. If the **outvia** keyword is configured and the zone named with the **outvia** keyword is local to the gatekeeper, the call is directed to a Cisco Unified Border Element in that zone by returning an ACF pointing to the Cisco Unified Border Element. If the zone named with the **outvia** keyword is remote, the gatekeeper sends a location request to the **outvia** gatekeeper rather than the remote zone gatekeeper. The **invia** keyword is not used in processing the ARQ.

Centralized Gatekeeper Configuration

A single gatekeeper can support call routing between clusters and call admission control for up to 100 Unified CM clusters. Figure 9-56 illustrates a distributed call processing environment with two Unified CM clusters and a single, centralized gatekeeper.

Figure 9-56 Centralized Gatekeeper Supporting Two Clusters



Example 9-5 shows the gatekeeper configuration for the example in Figure 9-56.

Example 9-5 Configuration for Centralized Gatekeeper

```
gatekeeper
 zone local GK-Site1 customer.com 10.1.10.100
 zone local GK-Site2 customer.com
 zone prefix GK-Site1 408.......
 zone prefix GK-Site2 212.......
 bandwidth interzone GK-Site1 160
 bandwidth interzone GK-Site2 160
 gw-type-prefix 1#* default-technology
 arq reject-unknown-prefix
 no shutdown
```

The following notes also apply to Figure 9-56:

- Each Unified CM cluster has a local zone configured to support Unified CM trunk registrations.
- A zone prefix is configured for each zone to allow inter-zone and inter-cluster call routing.

- Bandwidth statements are configured for each site. Cisco recommends that you use the **bandwidth interzone** command because using the **bandwidth total** command can cause issues in some configurations. Bandwidth is measured in kilobits per second (kbps).
- The **gw-type-prefix 1# default-technology** command allows all locally unresolved calls to be forwarded to a device registered with a technology prefix of 1# in the local zone. In this example, all Unified CM trunks have been configured to register with a 1# prefix.

  Technology prefixes indicate the type of call being made. The specific values used as technology prefixes are arbitrary and are defined by the network administrator. The same values should be used consistently throughout the entire deployment.

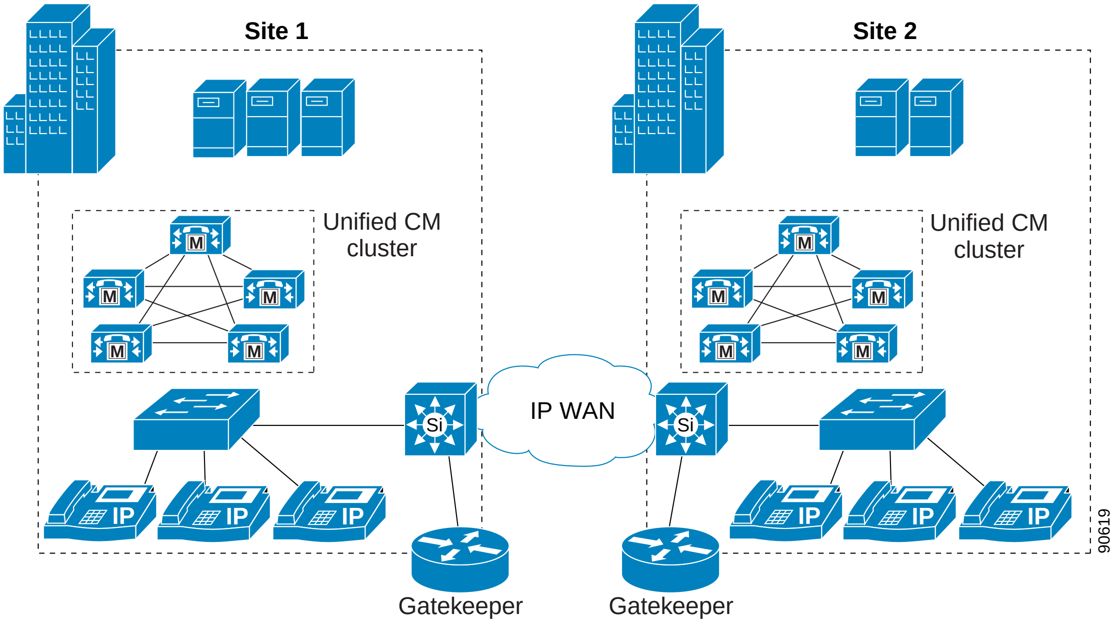
  Technology prefixes are sent as a prefix to the E.164 address (phone number) to indicate whether the call is voice, video, or some other type. The # symbol is generally used to separate the prefix from the E.164 number. If a prefix is not included, the default technology prefix is used to route the call. There can be only one default technology prefix for the entire deployment.

  Cisco IOS gateways automatically add a technology prefix to outbound calls if the gateway has a prefix configured. The gateway also automatically strips the prefix from incoming H.323 calls. Unified CM can register with the gatekeeper using a technology prefix, as specified on the configuration page for gatekeeper-controlled H.323 trunks. However, this technology prefix is not automatically added to outgoing calls to the gatekeeper, and is not automatically stripped from inbound calls to Unified CM. You can use translation patterns and significant digits to manipulate the called number so as to add or strip the technology prefix as needed.
- The **arq reject-unknown-prefix** command guards against potential call routing loops across redundant Unified CM trunks.

Distributed Gatekeeper Configuration

Gatekeepers can be distributed to conserve bandwidth or to provide local call routing for H.323 gateways in case of a WAN failure. Figure 9-57 illustrates a distributed call processing environment with two clusters and two gatekeepers.

Figure 9-57 Distributed Gatekeepers Supporting Two Clusters



Example 9-6 shows the gatekeeper configuration for Site 1 in Figure 9-57.

Example 9-6 Gatekeeper Configuration for Site 1

```
gatekeeper
 zone local GK-Site1 customer.com 10.1.10.100
 zone remote GK-Site2 customer.com 10.1.11.100
 zone prefix GK-Site1 408.......
 zone prefix GK-Site2 212.......
 bandwidth remote 160
 gw-type-prefix 1#* default-technology
 arq reject-unknown-prefix
 no shutdown
```

The following notes apply to Example 9-6:

- A local zone is configured for registration of local Unified CM cluster trunks.
- A remote zone is configured for routing calls to the Site 2 gatekeeper.
- Zone prefixes are configured for both zones for inter-zone call routing.
- The **bandwidth remote** command is used to limit bandwidth between the local zone and any other remote zone.
- The **gw-type-prefix 1# default-technology** command allows all locally unresolved calls to be forwarded to a device registered with a technology prefix of 1# in the local zone. In this example, all Unified CM trunks have been configured to register with a 1# prefix.

- The **arq reject-unknown-prefix** command guards against potential call routing loops across redundant Unified CM trunks.

Example 9-7 shows the gatekeeper configuration for Site 2 in Figure 9-57.

Example 9-7 Gatekeeper Configuration for Site 2

```
gatekeeper
 zone local GK-Site2 customer.com 10.1.11.100
 zone remote GK-Site1 customer.com 10.1.10.100
 zone prefix GK-Site2 212.......
 zone prefix GK-Site1 408.......
 bandwidth remote 160
 gw-type-prefix 1#* default-technology
 arq reject-unknown-prefix
 no shutdown
```

The following notes apply to Example 9-7:

- A local zone is configured for registration of local Unified CM cluster trunks.
- A remote zone is configured for routing calls to the Site 1 gatekeeper.
- Zone prefixes are configured for both zones for inter-zone call routing.
- The **bandwidth remote** command is used to limit bandwidth between the local zone and any other remote zone.
- The **gw-type-prefix 1# default-technology** command allows all locally unresolved calls to be forwarded to a device registered with a technology prefix of 1# in the local zone. In this example, all Unified CM trunks have been configured to register with a 1# prefix.
- The **arq reject-unknown-prefix** command guards against potential call routing loops across redundant Unified CM trunks.

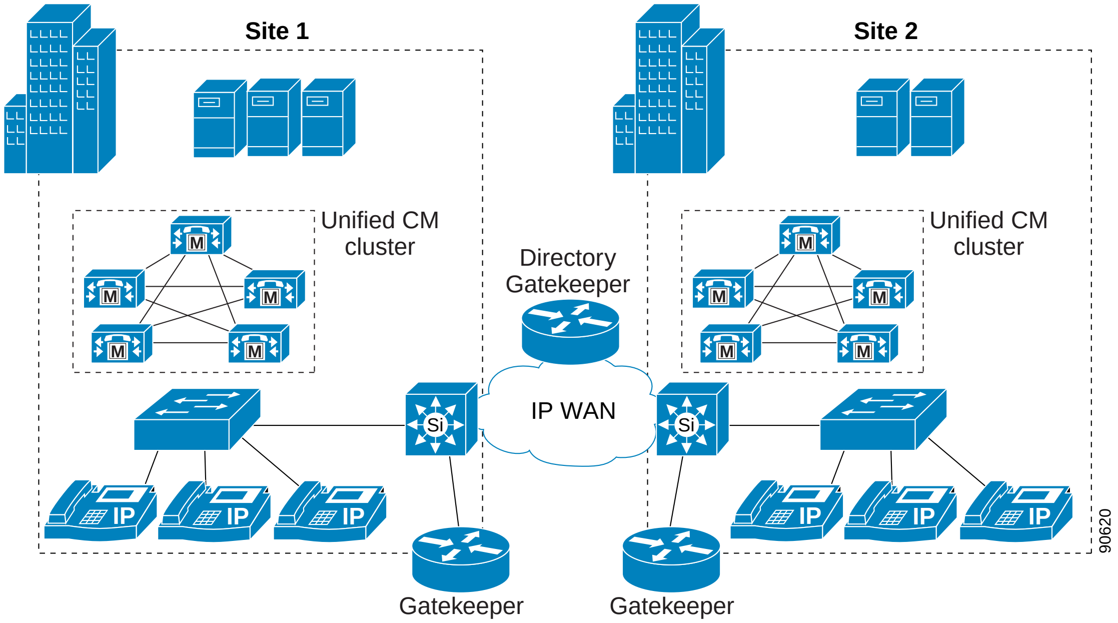
Distributed Gatekeeper Configuration with Directory Gatekeeper

Because there is no gatekeeper protocol available to update gatekeeper routing tables, use of a directory gatekeeper can help make distributed gatekeeper configurations more scalable and more manageable. Implementing a directory gatekeeper makes gatekeeper configurations at each site simpler and moves most of the configuration for inter-zone communication into the directory gatekeeper.

Without a directory gatekeeper, you would have to add an entry in every gatekeeper on the network every time you add a new zone on one of the gatekeepers. However, with a directory gatekeeper, you can add the new zone in the local gatekeeper and the directory gatekeeper only. If the local gatekeeper cannot resolve a call request locally, it forwards that request to the directory gatekeeper with a matching zone prefix.

Figure 9-58 illustrates a Unified CM distributed call processing environment with distributed gatekeepers for local call routing and a directory gatekeeper to provide call routing between gatekeepers.

Figure 9-58 Distributed Gatekeepers with a Directory Gatekeeper



Example 9-8 shows the gatekeeper configuration for Site 1 in Figure 9-58. Because the Site 1 and Site 2 gatekeeper configurations are almost identical in this example, only Site 1 is covered here.

Example 9-8 Gatekeeper Configuration for Site 1, with Directory Gatekeeper

```
gatekeeper
 zone local GK-Site1 customer.com 10.1.10.100
 zone remote DGK customer.com 10.1.10.101
 zone prefix GK-Site1 408.......
 zone prefix DGK ..........
 bandwidth remote 160
 gw-type-prefix 1#* default-technology
 arq reject-unknown-prefix
 no shutdown
```

The following notes also apply to Example 9-8:

- A local zone is configured for registration of local Unified CM cluster trunks.
- A remote zone is configured for the directory gatekeeper.
- Zone prefixes are configured for both zones for inter-zone call routing.
- The directory gatekeeper zone prefix is configured with 10 dots. This pattern matches any unresolved 10-digit dial strings. Multiple zone prefixes can be configured for a single zone, allowing matches on different length dial strings. A wildcard (*) can also be used for a directory gatekeeper zone prefix, but this method can cause call routing issues in some instances.
- The **bandwidth remote** command is used to limit bandwidth between the local zone and any other remote zone.
- The **gw-type-prefix 1# default-technology** command allows all locally unresolved calls to be forwarded to a device registered with a technology prefix of 1# in the local zone. In this example, all Unified CM trunks have been configured to register with a 1# prefix.

- The **arq reject-unknown-prefix** command guards against potential call routing loops across redundant Unified CM trunks.

Example 9-9 shows the directory gatekeeper configuration for the example in Figure 9-58.

Example 9-9 Directory Gatekeeper Configuration

```
gatekeeper
 zone local DGK customer.com 10.1.10.101
 zone remote GK-Site1 customer.com 10.1.10.100
 zone remote GK-Site2 customer.com 10.1.11.100
 zone prefix GK-Site1 408*
 zone prefix GK-Site2 212*
 lrq forward-queries
 no shutdown
```

The following notes also apply to Example 9-9:

- A local zone is configured for the directory gatekeeper.
- Remote zones are configured for each remote gatekeeper.
- Zone prefixes are configured for both remote zones for inter-zone call routing. The wildcard (*) is used in the zone prefix to simplify configuration. Calls will not be routed to the DGK zone, so no prefix is required for it.
- The **lrq forward-queries** command allows the directory gatekeeper to forward an LRQ received from another gatekeeper.

Calling Privileges in Cisco IOS with H.323 Dial Peers

Use the class of restriction (COR) functionality to implement calling privileges with Cisco IOS-based systems using H.323, including H.323 gateways, SRST, and Cisco Unified Communications Manager Express (Unified CME). This functionality provides flexibility in network design, allows administrators to block calls for all users (for example, calls to 900 numbers), and applies different calling privileges to call attempts from different originators (for example, it allows some users but not others to call international numbers).

The fundamental mechanism at the center of the COR functionality relies on the definition of incoming and outgoing *corlists* that are associated to existing dial peers, where the concepts of incoming and outgoing are relative to the Cisco IOS router (as in the case of dial peers). Each corlist is defined to include a number of members, which are simply tags previously defined within Cisco IOS.

When a call goes through the router, an incoming dial peer and an outgoing dial peer are selected based on the Cisco IOS dial peer routing logic. If corlists are associated with the selected dial peers, the following additional check is performed before extending the call:

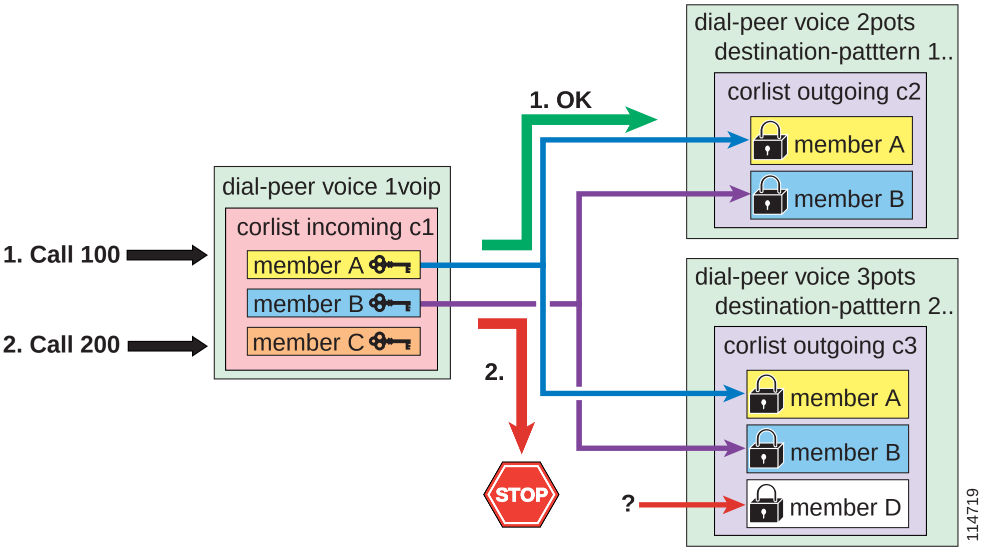
- If the members of the outgoing corlist associated with the outgoing dial peer are a subset of the members of the incoming corlist associated with the incoming dial peer, the call is permitted.
- If the members of the outgoing corlist associated with the outgoing dial peer are *not* a subset of the members of the incoming corlist associated with the incoming dial peer, the call is rejected.

If no corlist statements are applied to some dial peers, the following properties apply:

- When no incoming corlist is configured on a dial-peer, the default incoming corlist is used. The default incoming corlist has the highest possible priority, and it therefore allows this dial-peer to access all other dial-peers, regardless of their outgoing corlist.
- When no outgoing corlist is configured on a dial-peer, the default outgoing corlist is used. The default outgoing corlist has the lowest possible priority, and it therefore allows all other dial-peers to access this dial-peer, regardless of their incoming corlist.

This behavior is best illustrated with an example as shown in Figure 9-59, where one VoIP dial-peer and two POTS dial-peer are defined.

Figure 9-59 Example of COR Behavior



The VoIP dial-peer is associated with the c1 incoming corlist, with members A, B, and C. You can think of members of the incoming corlist as "keys."

The first POTS dial-peer has a destination-pattern of 1.. and is associated with the c2 outgoing corlist, with members A and B. The second POTS dial-peer has a destination-pattern of 2.. and is associated with the c3 outgoing corlist, with members A, B, and D. You can think of members of the outgoing corlists as "locks."

For the call to succeed, the incoming corlist of the incoming dial-peer must have all the "keys" needed to open all the "locks" of the outgoing corlist of the outgoing dial-peer.

In the example shown in Figure 9-59, a first VoIP call with destination 100 is received by the router. The Cisco IOS call routing logic matches the incoming call leg with the VoIP dial-peer and the outgoing call leg with the first POTS dial-peer. The COR logic is then applied; because the c1 incoming corlist has all the keys needed for the c2 outgoing corlist locks (A and B), the call succeeds.

A second VoIP call with destination 200 is then received by the router. The Cisco IOS call routing logic matches the incoming call leg with the VoIP dial-peer and the outgoing call leg with the second POTS dial-peer. The COR logic is then applied; because the c1 incoming corlist is missing one "key" for the c3 outgoing corlist (D), the call is rejected.

Follow these steps when configuring the COR functionality in Cisco IOS:

Step 1 Define "tags" to be used as corlist members with the command **dial-peer cor custom**.

Step 2 Define corlists with the command **dial-peer cor list** *corlist-name*.

Step 3 Associate corlists with existing VoIP or POTS dial-peers by using the command **corlist** {**incoming** | **outgoing**} *corlist-name* within the dial-peer configuration.

With Cisco IOS Release 12.2(8)T and later, you can apply the COR functionality to SRST-controlled IP phones. Because IP phones register with the SRST router dynamically, SRST has no prior knowledge of the individual phones until they lose connectivity to the Unified CM cluster. Therefore, the COR feature configuration for SRST is based on the phone DNs instead. When the IP phones register with the SRST router, they communicate their DN to it, thus allowing the SRST router to assign them to the appropriate corlist.

Configure COR for IP phones controlled by SRST by using the command **cor** {**incoming** | **outgoing**} *corlist-name* {*corlist-number starting-number – ending-number* | **default**} within the **call-manager-fallback** configuration mode.

The following limitations apply to this command:

- The maximum number of **cor** {**incoming** | **outgoing**} statements under **call-manager-fallback** is 5 (plus the default statement) in SRST version 2.0, available on Cisco IOS Release 12.2(8)T or later.
- The maximum number of **cor** {**incoming** | **outgoing**} statements under **call-manager-fallback** is 20 (plus the default statement) in SRST version 3.0, available on Cisco IOS Release 12.3(4)T) or later.

The COR functionality can also be deployed with Cisco Unified Communications Manager Express (Unified CME), using Cisco IOS Release 12.2(8)T and later. Because the individual IP phones are specifically configured within Unified CME, you can apply corlists directly to the IP phones themselves by using the command **cor** {**incoming** | **outgoing**} *corlist-name* within the **ephone-dn** *dn-tag* configuration mode of each IP phone.

Refer to the section on Building Classes of Service in Cisco IOS with H.323, page 9-71, for an example of how to apply all these concepts in practice.

For more details on configuration of Cisco SRST and Unified CME, refer to the *Cisco SRST System Administrator Guide* and the *Cisco Unified Communications Manager Express System Administrator Guide*, both available at

http://www.cisco.com

Digit Manipulation in Cisco IOS with H.323 Dial Peers

In Cisco IOS routers running H.323, digit manipulation is performed via voice translation profiles, which are used to manipulate the calling number (ANI) or called number (DNIS) digits for a voice call or to change the numbering type of a call.

Voice translation profiles are available starting with Cisco IOS Release 12.2(11)T, and they are used to convert a telephone number into a different number before the call is matched to an incoming dial peer or before the call is forwarded by the outgoing dial peer. For example, within your company you might dial a five-digit extension to reach an employee at another site. If the call is routed through the PSTN to reach the other site, the originating gateway must use voice translation profiles to convert the five-digit extension into the ten-digit format that is recognized by the central office switch.

You configure voice translation profiles by using the **voice translation-rule** and **voice translation-profile** Cisco IOS commands, which use regular expressions to define the digit strings to be modified. You then specify if the manipulations should be associated to calling numbers, called numbers, or redirected called numbers. After you define a voice translation profile, you can apply it to any of the following elements:

- All incoming POTS call legs that terminate on a specific voice port
- All incoming VoIP call legs into the router
- Outgoing call legs associated with a specific VoIP or POTS dial peer
- All incoming or outgoing call legs that terminate on the IP phones controlled by SRST
- Incoming call legs for calls originated by all IP phones controlled by SRST

Note Voice translation profiles using the **voice translation-rule** command replace and enhance the functionality previously provided by the **translation-rule** command. The syntax of the new command differs from that used by the old command. For more details, refer to **voice translation-rule** in the *Cisco IOS Voice Command Reference*, Release 12.2(11)T or later, available at http://www.cisco.com.

A typical application of voice translation profiles is to enable the preservation of on-net inter-site dialing habits from a branch site even when the IP WAN is unavailable and the router is running in SRST mode. For example, consider a simple deployment with a central site in San Jose and three remote sites in San Francisco, New York, and Dallas. Table 9-10 shows the DID ranges and the internal site codes for this example.

Table 9-10 Example of DID Ranges and Site Codes for Translation Rule Application

	San Jose	San Francisco	New York	Dallas
DID Range	(408) 555-1XXX	(415) 555-1XXX	(212) 555-1XXX	(972) 555-1XXX
Site Code	1	2	3	4

Inter-site calls are normally placed across the IP WAN by dialing the on-net access code 8 followed by the one-digit site code and the four-digit extension of the called party. To preserve these dialing habits even when the IP WAN is down and Cisco SRST is active, the internal numbers must be converted back into the E.164 numbers before being sent to the PSTN. A sample configuration for the San Francisco router is as follows:

```
voice translation-rule 1
  rule 1 /^81/ /91408555/
  rule 2 /^83/ /91212555/
  rule 3 /^84/ /91972555/

voice translation-profile on-net-xlate
  translate called 1

call-manager-fallback
  translation-profile outgoing on-net-xlate

dial-peer voice 2 pots
  destination-pattern 91[2-9]..[2-9]......
port 1/0:0
  direct-inward-dial
  forward-digits 11
```

With this configuration, when the San Francisco site is in SRST mode and a user dials 831000, the router will match **rule 2** of **voice translation-rule 1** and translate the called number to 912125551000. This new number will then be used to match the outgoing dial peer (**dial-peer voice 2**).

For a detailed description of dial peers and their configuration, refer to *Configuring Dial Plans, Dial Peers, and Digit Manipulation*, part of the *Cisco IOS Voice, Video, and Fax Configuration Guide, Release 12.2*, available at

http://www.cisco.com

For a thorough explanation of regular expression syntax in Cisco IOS, refer to the information on *Regular Expressions* in the *Cisco IOS Terminal Services Configuration Guide*, available at

http://www.cisco.com/en/US/docs/ios/termserv/configuration/guide/tsv_reg_express_ps6441_TSD_Products_Configuration_Guide_Chapter.html

Service Advertisement Framework (SAF) Call Control Discovery (CCD)

This section highlights some aspects of the Cisco Service Advertisement Framework (SAF) Call Control Discovery (CCD) service configuration in Unified CM and related Unified Communications subsystems. For more details on this subject, refer to the section on Call Routing and Dial Plan Distribution Using Call Control Discovery for the Service Advertisement Framework, page 5-52, and to the latest version of the *Cisco Unified Communications Manager Administration Guide*, available at

http://www.cisco.com/en/US/products/sw/voicesw/ps556/prod_maintenance_guides_list.html

By participating in the framework as a CCD advertiser, a Unified CM cluster injects information into the network about the DN ranges it hosts. This information is sent to a Service Advertisement Framework Forwarder (SAF Forwarder), which learns the new routes and shares them with other participating SAF Forwarders and CCD requestors in the network.

By participating in the framework as a CCD requestor, a Unified CM cluster learns the DN ranges advertised by other call agents in the network from a SAF Forwarder.

SAF Forwarders

SAF Forwarders are configured on Cisco IOS routers and require Cisco IOS Release 15.0(1) or higher. For more information on configuring SAF Forwarders, consult the Cisco IOS *Service Advertisement Framework Configuration Guide*, available at

http://www.cisco.com/en/US/products/ps10591/products_installation_and_configuration_guides_list.html

SAF Forwarder Configuration

Within Unified CM, you need to configure both a SAF Security Profile (**Advanced Features** > **SAF** > **SAF Security Profile**) and a SAF Forwarder (**Advanced Features** > **SAF** > **SAF Forwarder**).

The SAF Security Profile configuration page in Unified CM features User Name and User Password fields. These entries must match the SAF Forwarder configuration in the Cisco IOS command line interface (CLI).

Also, the Unified CM Client Label as configured under the SAF Forwarder configuration page must match one of the external-client statements from the SAF Forwarder's CLI configuration. For example:

```
router eigrp
  service-family ipv4 autonomous-system 1
  !
  topology base
```

```
   external-client sample_client_label
  exit-sf-topology
 exit-service-family
!

service-family external-client listen ipv4 5050
 external-client sample_client_label
  username sample_user_name
  password sample_user_password
  keepalive 10000
!
```

For more details, refer to the Cisco IOS *Service Advertisement Framework Configuration Guide*, available at

http://www.cisco.com/en/US/products/ps10591/products_installation_and_configuration_guides_list.html

Requesting Service

The SAF CCD learned DN ranges are populated in a dedicated CCD call routing partition in the requesting cluster. Learned DN ranges are associated with one or more CCD trunks. Calls made to CCD learned DN ranges are placed through CCD trunks associated with the requesting service. The association of CCD trunks with the requesting service is done through the Selected SAF Trunks filed under the CCD Requesting Service Info page in Unified CM, located under **Call Routing** > **Call Control Discovery** > **Requesting Service**.

The CCD records exchanged between a cluster and a SAF Forwarder include information about the DN range, the IP address of the call agent node hosting the DN range, the digit manipulation rules to adapt a DN when rerouting the call to the PSTN, and the IP signaling protocol required to call this DN range.

For example, Cluster A hosts DN range 8555XXXX, whose corresponding DID range on the PSTN is +1415555XXXX. The IP address of the Cluster A subscriber associated with the CCD trunk designated to receive IP calls for this DN range is 10.1.1.1. The protocol required to reach this DN range is SIP. The CCD record associated with this DN range can be represented as follows:

DN Range	ToDID	IP	Protocol
8555XXXX	1:+1415	10.1.1.1	SIP

- DN range

  If a user dials 85551234, a match would be made on 8555XXXX and a call to 85551234 would be extended to the cluster that advertised the pattern.

- ToDID

  This field represents the rules allowing the DN range to be reached across the PSTN. If a user dials 85551234 and the call cannot be routed through a CCD trunk, the ToDID rules are applied and the destination number is transformed to a form compatible to the PSTN. For example, the rule 1:+1415 applied to the range 8555XXXX would require the removal of one digit and the prefixing of +1415. The resulting +14155551234 would allow routing of the call to any gateway in the cluster of origin, provided that it is provisioned to route calls in the +E.164 form and that gateways are provisioned with appropriate called party transformation patterns to adapt the globalized +E.164 form of the number to a localized form acceptable to the PSTN carrier.

- IP

  The IP address of the destination DN's hosting call agent node is used when placing a call across the associated CCD trunk, in the cluster of origin.

- Protocol

  In this case, SIP is the protocol advertised by the call agent hosting the DN range. The other possible choice is H.323.

To view which SAF CCD records were discovered by a particular cluster, use the Cisco Unified CM Real-Time Monitoring Tool (RTMT). It offers information about the discovered DN ranges as well as information about the SAF Forwarder associations with the cluster.

Dial Plan Considerations for Business Edition 3000

Cisco Business Edition 3000 comes with a highly simplified front end with a drawer user interface (GUI). New concepts such as Site and Usage profiles have been introduced. Although the underlying concept of dial plan still remains the same with the line/device approach and calling search spaces, Business Edition 3000 implements the dial plan by using the concepts of sites and profiles. The calling privileges at the device level are defined by the Site, and restrictions to the calling privileges at line level are defined by the Usage profile, as follows:

- Site

  Site represent the geographic grouping of endpoints, including phones, users, gateways, and so forth. Privileges are given to a Site, which define the highest level of calling privileges each user at that site can achieve. This does not mean that every user will have those privileges; instead, the privileges govern the capability of each location to make calls. For example, a location can have calling privileges to dial international numbers, but that does not necessarily mean that every user at the site can make international calls.

- Usage profile

  A usage profile allows the system administrator to configure most of the user settings for a phone in one place. The administrator can edit an existing usage profile, duplicate an existing usage profile to create a new profile, or add an entirely new usage profile. Each usage profile has a unique name. After configuring usage profiles, the system administrator can assign them to users or to departments, so that the settings in the usage profile apply to the phones that belong to an individual user or to an entire department.

In the Usage profile, you can configure calling privileges for users; phone features such as barge, Cisco Extension Mobility, and so on; phone hardware functionality; phone applications that may display on the phone; and the phone button template, which controls the order of the buttons and the feature buttons that display on the phone.

Note Cisco Business Edition 3000 supports a maximum of 10 usage profiles.

The combination of Site privileges and Usage calling privileges define the actual capability for a user to dial numbers. For example, a Site can be allocated privileges to dial international calls as the highest level of calling, which essentially means that users can dial any number including international, long distance, and local calls. But if a user is attached to a usage profile that limits the dialing privilege to local calls only, then that user will be able to dial only local calls even though the site has privileges to dial international calls.

As another example, assume that the user is associated to a Usage profile and a Site that both have calling privileges to dial international calls. If this user moves to another site where the Site level privileges are limited to dialing local calls only, then this user will not be able to dial beyond local calls because the current site level calling privileges do not allow it.

Essentially, the lower of the Site level calling privileges and the Usage profile calling privileges determines the calling privileges for a user.

Translation Rules

Translation rules allow Cisco Business Edition 3000 to manipulate an incoming phone number that is part of your system and transform it to an extension before routing the call. Any call coming into the system or generated by IP phones is matched against the configured translation rule, and if the number matches, the translation is performed.

Note Support for wild cards in translation rules is not available with Business Edition 3000.

Logical Partitioning

Every phone is associated to a site based upon the IP address configured on the phone. Every site is mapped to one or more subnet(s). If the phone IP address lies within one of those subnets, then the phone belongs to the site to which that subnet is mapped. If the phone acquires an IP address that is not defined in the system, then phone is assumed to be part of the central site. However, if a teleworker site is configured, then any phone whose configured IP address does not lie within one of the configured subnets will be assumed to be a teleworker phone.

The configured sites provide support for logical partitioning. While configuring sites, the administrator is required to configure the PSTN privileges. If access to the central site PSTN is not configured, the users at a remote site will not be allowed to be part of any conversation where a PSTN call leg is involved. Also, the remote site phones will not be able to initiate PSTN calls.

CHAPTER 10

Emergency Services

Revised: March 31, 2011; OL-27282-05

Emergency services are of great importance in the proper deployment of a voice system. This chapter presents a summary of the following major design considerations essential to planning for emergency calls:

This chapter presents some information specific to the 911 emergency networks as deployed in Canada and the United States. Many of the concepts discussed here are adaptable to other locales. Please consult with your local telephony network provider for appropriate implementation of emergency call functionality.

In the United States, some states have already enacted legislation covering the 911 functionality required for users in a multi-line telephone system (MLTS). The National Emergency Number Association (NENA) has also produced the *NENA Technical Requirements Document on Model Legislation E9-1-1 for Multi-Line Telephone Systems*, available online at

http://www.nena.org/

The Federal Communications Commission (FCC) has also drafted a proposed new section to Title 47, Part 68, Section 319, which is available at

http://www.apcointl.org/about/pbx/worddocs/mltspart68.doc

This chapter assumes that you are familiar with the generic 911 functionality available to residential PSTN users in North America. For more information on the subject, refer to the NENA website at

http://www.nena.org/

What's New in This Chapter

Table 10-1 lists the topics that are new in this chapter or that have changed significantly from previous releases of this document.

Table 10-1 New or Changed Information Since the Previous Release of This Document

New or Revised Topic	Described in	Revision Date
No changes for Cisco Unified Communications System Release 9.0		June 28, 2012

911 Emergency Services Architecture

This section highlights some of the functionality requirements for emergency calls in multi-line telephone systems (MLTS). In the context of this section, emergency calls are 911 calls serviced by the North American public switched telephone network (PSTN).

Any emergency services architecture usually consists of the following elements:

- A distressed caller should be able to dial the emergency services from a fixed line, a mobile phone, a public phone, or any device capable of making the voice call.
- An emergency services call handler must be available to respond to the emergency request and dispatch the needed services such as police, fire, and medical.
- The call handler should be able to identify the location of the distressed caller as precisely as possible to provide help.
- An emergency services network is needed to route the call to the nearest emergency services call handler with jurisdiction for the location of the caller.

The following sections explain some of the important architectural components of 911 emergency services architecture.

Public Safety Answering Point (PSAP)

The public safety answering point (PSAP) is the party responsible for answering the 911 call and arranging the appropriate emergency response, such as sending police, fire, or ambulance teams. The physical location of the phone making the 911 call is the primary factor in determining the appropriate PSAP for answering that call. Generally, each building is serviced by one local PSAP.

To determine the responsible PSAP for a given location, contact a local public safety information service such as the local fire marshal or police department. Also, the phone directory of the local exchange carrier usually lists the agency responsible for servicing 911 calls in a given area.

Typical Situation

- For a given street address, there is only one designated PSAP.
- For a given street address, all 911 calls are routed to the same PSAP.

Exceptional Situation

- The physical size of the campus puts some of the buildings in different PSAP jurisdictions.
- Some of the 911 calls need to be routed to an on-net location (campus security, building security).

Selective Router

The selective router is a node in the emergency services network that determines the appropriate PSAP for call delivery based on caller's geographic area and the automatic number identification (ANI). The Local Exchange Carrier (LEC) usually operates the selective router. Hence it is imperative to ensure that the enterprise IP telephony network is designed in such a way that the caller is routed to the appropriate selective router based on its location.

Automatic Location Identifier Database

Location information of the caller is an important part of the 911 services infrastructure. The Automatic Location Identifier (ALI) database maintains the location information for the particular geographical location served by the LEC. For every 911 call, the PSAP searches the ALI database to retrieve the caller's location based on the ANI of the calling number. The addresses are stored in the Master Street Address Guide (MSAG) format in the ALI database. The ALI database is maintained on behalf of the local emergency services administration by a contracted third party, generally the incumbent Local Exchange Carrier (LEC).

Private Switch ALI

Private Switch ALI (PS/ALI) is an enhancement to 911 emergency response systems that enables MLTS operators to provide more specific address and location information for each telephony endpoint. The service allows a customer-generated address table to be loaded into the ALI database so that each station of an MLTS system can be uniquely identified if a call is placed to 911 from that telephone number. The station-specific or location-specific automatic number identification (ANI) generated by the switching system can be passed directly to the E911 system to pinpoint the precise location of the caller. The PSAP operator can then direct emergency response personnel to the correct address, building, floor, room, or even cubicle, thereby streamlining operations and increasing accuracy.

911 Network Service Provider

After identifying the responsible PSAPs, you must also identify the 911 network service providers to which each PSAP is connected. It is commonly assumed that PSAPs receive 911 phone calls from the PSTN, but that is not the case. Instead, 911 calls are carried over dedicated, regionally significant networks, and each PSAP is connected to one or more such regional networks. In the majority of cases, the incumbent Local Exchange Carrier (LEC) is the 911 network service provider for a PSAP. Some exceptions include military installations, university campuses, federal or state parks, or other locations where the public safety responsibility falls outside the jurisdiction of the local authorities and/or where a private network is operated by an entity other than a public local exchange carrier.

If you are in doubt about the 911 network service provider for a given PSAP, contact the PSAP directly to verify the information.

Typical Situation

- For a given street address, the 911 network service provider is the incumbent Local Exchange Carrier (LEC). For a location served by Phone Company X, the corresponding PSAP is also served by Phone Company X.
- All 911 calls are routed directly to an off-net location, or all 911 calls are routed directly to an on-net location.

Exceptional Situation

- The local exchange carrier (LEC) through which the MLTS interfaces to the PSTN is *not* the same LEC that serves as 911 network service provider to the PSAP. (For example, the phone system is served by Phone Company X, but the PSAP is connected to Phone Company Y.) This situation might require either a special arrangement between the LECs or special, dedicated trunks between the phone system and the PSAP's 911 network service provider.
- Some LECs may not accept 911 calls on their networks. If this is the case, the only two options are to change LECs or to establish trunks (dedicated to 911 call routing) connected to a LEC that can route 911 calls to the appropriate PSAPs.
- Some (or all) of the 911 calls have to be routed to an on-net location such as campus security or building security. This situation can easily be accommodated during the design and implementation phases, but only if the destination of 911 calls for each phone has been properly planned and documented.

Interface Points into the Appropriate 911 Networks

For larger telephony systems, 911 connectivity might require many interface points. Typically, more than one E911 selective router is used within a LEC's territory, and these routers usually are *not* interconnected.

For example, an enterprise with a large campus could have the following situation:

- Building A located in San Francisco
- Building B located in San Jose
- San Francisco Police Department and San Jose Police Department are the appropriate PSAPs
- San Francisco Police Department and San Jose Police Department are served by the same 911 network service provider
- However, San Francisco Police Department and San Jose Police Department are served by different E911 selective routers operated by that same 911 network service provider!

This type of situation would require two separate interface points, one per E911 selective router. The information pertaining to the E911 selective router territories is generally kept by the incumbent LEC, and the local account representative for that LEC should be able to provide an enterprise customer with the pertinent information. Many LECs also provide the services of 911 subject matter experts who can consult with their own account representatives on the proper mapping of 911 access services.

Typical Situation

- For single-site deployments or campus deployments, there is usually only one PSAP for 911 calls.
- If access to only one PSAP is required, then only one interface point is required. Even if access to more than one PSAP is required, they might be reachable from the same E911 selective router, through the same centralized interface. If the enterprise's branch sites are linked via a WAN (centralized call processing), it is desirable to give each location its own local (that is, located inside each branch office) access to 911 to prevent 911 isolation during WAN failure conditions where Survivable Remote Site Telephony (SRST) operation is activated.

Exceptional Situation

- The physical size of the campus puts some of the buildings in different PSAP jurisdictions, *and*
- Some of the 911 calls have to be routed to different E911 selective routers, through different interface points.

Note Some of the information required to establish the geographical territories of PSAPs and E911 selective routers is available online or from various competitive local exchange carrier (CLEC) information web sites. (For example, https://clec.att.com/clec/hb/shell.cfm?section=782 provides some valuable data about the territory covered by AT&T in California and Nevada.) However, Cisco strongly recommends that you obtain proper confirmation of the appropriate interface points from the LEC prior to the design and implementation phases of 911 call routing.

Interface Type

In addition to providing voice communications, the interfaces used to present 911 calls to the network must also provide identification data about the calling party.

Automatic Number Identification (ANI) refers to the North American Numbering Plan number of the calling party, which is used by networks to route a 911 call to the proper destination. This number is also used by the PSAP to look up the Automatic Location Identification (ALI) associated with a call.

911 calls are source-routed, which means that they are routed according to the calling number. Even though different locations are all dialing the same number (911), they will reach different PSAPs based on their location of origin, which is represented by the ANI (calling number).

You can implement 911 call functionality with either of the following interface types:

- Dynamic ANI assignment
- Static ANI assignment

While dynamic ANI assignment scales better (because it supports multiple ANIs) and lends itself to all but the smallest of applications, static ANI assignment can be used in a wider variety of environments, from the smallest to the largest systems.

Dynamic ANI (Trunk Connection)

The dynamic aspect of ANI refers to the fact that a system has many phones sharing access to the 911 network across the same interface, and the ANI transmitted to the network might need to be different for each call.

There are two main types of dynamic ANI interfaces:

- Integrated Services Digital Network Primary Rate Interface (ISDN-PRI, or simply PRI)
- Centralized Automatic Message Accounting (CAMA).

PRI

This type of interface usually connects a telephony system to a PSTN Class 5 switch. The calling party number (CPN) is used at call setup time to identify the E.164 number of the calling party.

Most LECs treat the CPN differently when a call is made to 911. Depending upon the functionality available in the Class 5 switch and/or upon LEC or government policy, the CPN may not be used as the ANI for 911 call routing. Instead, the network may be programmed to use the listed directory number (LDN) or the bill-to number (BTN) for ANI purposes.

If the CPN is not used for ANI, then 911 calls coming from a PRI interface all look the same to the 911 network because they all have the same ANI, and they are all routed to the same destination (which might not be the appropriate one).

Some LECs offer a feature to provide CPN transparency through a PRI interface for 911 calls. With this feature, the CPN presented to the Class 5 switch at call setup is used as ANI to route the call. The feature name for this functionality varies, depending on the LEC. (For example, SBC calls it Inform 911 in California.)

Note The CPN *must* be a routable North American Numbering Plan number, which means that the CPN must be entered in the routing database of the associated E911 selective router.

Note For Direct Inward Dial (DID) phones, the DID number could be used as the ANI for 911 purposes, but only if it is properly associated with an Emergency Service Number in the 911 service provider's network. For non-DID phones, use another number. (See Emergency Location Identification Number Mapping, page 10-10, for more information.)

Many Class 5 switches are connected to E911 selective routers through trunks that do not support more than one area code. In such cases, if PRI is used to carry 911 calls, then the only 911 calls that will be routed properly are those whose CPN (or ANI) have the same Numbering Plan Area (NPA) as the Class 5 switch.

Example

An MLTS is connected to a Class 5 switch in area code 514 (NPA = 514). If the MLTS were to send a 911 call on the PRI trunk, with a CPN of **450**.555.1212, the Class 5 switch would send the call to the E911 selective router with an ANI of **514**.555.1212 (instead of the correct **450**.555.1212), yielding inappropriate routing and ALI lookup.

To use PRI properly as a 911 interface, the system planner must ensure that the CPN will be used for ANI and must properly identify the range of numbers (in the format NPA NXX TNTN) acceptable on the link. For example, if a PRI link is defined to accept ANI numbers within the range 514 XXX XXXX, then only calls that have a Calling Party Number with NPA = 514 will be routed appropriately.

CAMA

Centralized Automatic Message Accounting (CAMA) trunks also allow the MLTS to send calls to the 911 network, with the following differences from the PRI approach:

- CAMA trunks are connected directly into the E911 selective router. Extra mileage charges may apply to cover the distance between the E911 selective router and the MLTS gateway point.
- CAMA trunks support 911 calls only. The capital and operational expenses associated with the installation and operation of CAMA trunks support 911 traffic only.
- CAMA trunks for the MLTS market may be limited to a fixed area code, and the area code is typically implied (that is, not explicitly sent) in the link protocol. The connection assumes that all calls share the same deterministic area code, therefore only 7 or 8 digits are sent as ANI.

Static ANI (Line Connection)

Static ANI provides a line (rather than a trunk) connection to the PSTN, and the ANI of the line is associated with all 911 calls made on that line, regardless to the CPN of the calling phone. A plain old telephone service (POTS) line is used for this purpose.

POTS lines are one of the simplest and most widely supported PSTN interfaces. A POTS line usually comes fully configured to accept 911 calls. In addition, the existing E911 infrastructure supports 911 calls from POTS lines very well.

The POTS approach has the following attributes:

- The operational costs associated with a POTS line are low.
- The POTS line can even serve as a backup line in case of power failure.
- The POTS line number can be used as the callback number entered into the ALI database.
- POTS lines represent the lowest cost 911 support for locations where user density does not justify local PRI or CAMA access into the PSTN.
- POTS lines are ubiquitous in PSTN installations.

All outgoing 911 calls through this type of interface are treated the same by the E911 network, and the tools that enable Cisco Unified Communications Manager to control the ANI presented to the E911 network (such as calling party transformation masks) are irrelevant because the ANI can be only the POTS line's number.

Cisco Emergency Responder

Ease of administration for moves, adds, and changes is one of the key advantages of IP telephony technology. To provide for moves, adds, and changes that automatically update 911 information without user intervention, Cisco has developed a product called the Cisco Emergency Responder (Cisco ER).

Cisco ER provides the following primary functionality:

- Dynamic association of a phone to an Emergency Response Location (ERL), based on the detected physical location of the phone.
- Dynamic association of the Emergency Location Identification Number (ELIN) to the calling phone, for callback purposes. In contrast to non-ER scenarios outlined in preceding sections, Cisco ER enables the callback to ring the exact phone that initiated the 911 call.
- On-site notification to designated parties (by pager, web page, or phone call) to inform them that there is an emergency call in progress. The pager and web page notifications include the calling party name and number, the ERL, and the date and time details associated with the call. Phone notification provides the information about the calling number from which the emergency call was placed.

For more information on ERLs and ELINs, see Emergency Response Location Mapping, page 10-10, and Emergency Location Identification Number Mapping, page 10-10. For more information on Cisco ER, see Cisco Emergency Responder Design Considerations, page 10-15, and refer to the Cisco ER product documentation available online at

http://www.cisco.com/en/US/products/sw/voicesw/ps842/tsd_products_support_series_home.html

The key functionality of Cisco ER relies on the detection of the phone's location by discovery of the network port (Layer 2 port, such as a Fast Ethernet switched port) from which the phone made the 911 call. The discovery mechanism relies on two main assumptions:

- The wired infrastructure of the enterprise is well established and does not change sporadically.
- The infrastructure is available for Cisco ER to browse; that is, Cisco ER can establish Simple Network Management Protocol (SNMP) sessions to the underlying network infrastructure and can scan the network ports for the discovery of connected phones.

Once Cisco ER discovers the originating port for the call, it associates the call with the pre-established ERL for the location of that port. This process also yields an association with a pre-established ELIN for the location and the selection of the appropriate egress point to the E911 infrastructure, based on the originating ERL.

Cisco ER also provides the capability to configure ERLs for IP subnets and to assign IP phone location by IP address. This capability may be used to locate wireless IP phones, IP softphones, and third-party SIP phones registered to Cisco Unified CM, which Cisco ER cannot locate by connected switch port. It may also be used instead of, or in addition to, connected switch port locations for wired Cisco IP phones. If both connected switch port and IP subnet locations are available for an IP phone, Cisco ER will prefer the connected switch port location because it is usually more specific than the IP subnet location. Using both connected switch port and IP subnet locations is a best practice because it provides assurance that an appropriate ERL will be assigned, even in case of any delay or error in detecting the connected switch port.

Cisco ER allows for the use of two or more ELINs per ERL. The purpose of this enhancement is to cover the specific case of more than one 911 call originating from a given ERL within the same general time period, as illustrated by the following examples.

Example 1

- Phone A and phone B are both located within ERL X, and ERL X is associated with ELIN X.
- Phone A makes a 911 call at 13:00 hours. ELIN X is used to route the call to PSAP X, and PSAP X answers and releases the call. Then, at 13:15 hours, phone B makes a 911 call. ELIN X is again used to route the call to PSAP X.
- PSAP X, after releasing the call from phone B, decides to call back phone A for further details pertaining to phone A’s original call. The PSAP dials ELIN X, and gets phone B (instead of the desired phone A).

To work around this situation, Cisco ER allows you to define a pool of ELINs for each ERL. This pool provides for the use, in a round-robin fashion, of a distinct ELIN for each successive call. With the definition of two ELINs for ERL X in our example, we now have the situation described in Example 2.

Example 2

- Phone A and phone B are both located within ERL X. ERL X is associated with both ELIN X1 and ELIN X2.
- Phone A makes a 911 call at 13:00 hours. ELIN X1 is used to route the call to PSAP X, and PSAP X answers and releases the call. Then, at 13:15 hours, phone B makes a 911 call, and ELIN X2 is used to route this call to PSAP X.
- PSAP X, after releasing the call from phone B, decides to call back phone A for further details pertaining to phone A’s original call. The PSAP dials ELIN X1 and gets phone A.

Of course, if a third 911 call were made but there were only two ELINs for the ERL, the situation would allow for callback functionality to properly reach only the last two callers in the sequence.

High Availability for Emergency Services

It is very important for emergency services to always be available to the user even under the most critical of conditions. Therefore, high availability planning must be done carefully when deploying emergency services in an enterprise.

Cisco Emergency Responder supports clustering with a maximum of two servers in active/standby mode. The data is synchronized between the primary and the secondary Cisco ER servers. To ensure that calls are routed to the secondary server if the primary server is unavailable, the system administrator must follow certain provisioning guidelines for configuring CTI route points and the DNs associated to those CTI route points in Cisco Unified CM. For more details on configuration, refer to the *Cisco Emergency Responder Administration Guide*, available at

http://www.cisco.com/en/US/products/sw/voicesw/ps842/prod_maintenance_guides_list.html

If both of the Cisco ER servers are unavailable, a local route group (LRG) may be used to route the call to the appropriate PSAP with an appropriate ELIN/ERL (which might be less specific than what Cisco ER could have provided). Alternatively, the call may be routed to an internal security office to determine the caller's location. In either case, this provisioning must be done in Cisco Unified CM.

Apart from Cisco ER redundancy, Cisco Unified CM redundancy and gateway/trunk redundancy should also be considered where possible to route the 911 emergency calls and to avoid any single point of failure.

Capacity Planning for Cisco ER Clustering

In a Cisco ER cluster, the quantity of phones roaming outside the tracking domain of their home Cisco ER group is a scalability factor that you must kept within the limits set forth in the section on *Network Hardware and Software Requirements* in the *Cisco Emergency Responder Administration Guide*, available at:

http://www.cisco.com/en/US/products/sw/voicesw/ps842/prod_maintenance_guides_list.html

With the Cisco MCS 7845 server platform, a Cisco ER cluster can support a maximum of 3000 roaming phones. For deployments that have to exceed this limit (for instance, large campus deployments with multiple Unified CM clusters), phone movement can be tracked by IP subnets. By defining the IP subnets in each of the Cisco ER groups and by assigning each ERL with one ELIN per Cisco ER group, you can virtually eliminate roaming phones because all phones in the campus will be part of the tracking domain of their respective Cisco ER group.

To ensure proper sizing, use the Cisco Unified Communications Sizing Tool (Unified CST). This tool is available only to Cisco partners and employees, with appropriate login required, at http://tools.cisco.com/cucst. If you do not have access to this sizing tool, work with your Cisco account team or partner integrator to size your system appropriately.

Design Considerations for 911 Emergency Services

When planning 911 emergency services for a multi-line telephone system (MLTS) deployment, first establish all of the physical locations where phone services are needed. The locations can be classified as follows:

- Single building deployments, where all users are located in the same building
- Single campus deployments, where the users are located in a group of buildings situated in close proximity
- Multisite deployments, where users are distributed over a wide geographical area and linked to the telephony call processing site through WAN connectivity

The locations, or type of deployment, affect the criteria used to design and implement 911 services. The following sections describe the key criteria, along with typical and exceptional situations for each. When analyzing and applying these criteria, consider how they are affected by the phone locations in your network.

Emergency Response Location Mapping

The National Emergency Number Association (NENA) has proposed model legislation to be used by state and federal agencies in enacting the rules that govern 911 in enterprise telephony systems. One of the concepts in the NENA proposal is that of the emergency response location (ERL), which is defined as:

> *A location to which a 911 emergency response team may be dispatched. The location should be specific enough to provide a reasonable opportunity for the emergency response team to quickly locate a caller anywhere within it.*

Rather than having to identify each phone's location individually, the requirement allows for the grouping of phones into a "zone," the ERL. The maximum size of the ERL may vary, depending upon local implementation of the legislation, but we will use 7000 square feet (sq ft) as a basis for discussion in this section. (The concepts discussed here are independent of the maximum ERL size that may be allowed in any given state or region.)

An emergency location identification number (ELIN) is associated with each ERL. The ELIN is a fully qualified E.164 number, used to route the call within the E911 network. The ELIN is sent to the E911 network for any 911 call originating from the associated ERL. This process allows more than one phone to be associated with the same fully qualified E.164 number for 911 purposes, and it can be applied to DID and non-DID phones alike.

Note This document does not attempt to present the actual requirements of any legislation. Rather, the information and examples presented here are for the purposes of discussion only. The system planner is responsible for verifying the applicable local requirements.

For example, assume a building has a surface area of 70,000 sq ft and 100 phones. In planning for 911 functionality, the building can be divided into 10 zones (ERLs) of 7000 sq ft each, and each phone can be associated with the ERL where it is located. When a 911 call is made, the ERL (which could be the same for multiple phones) is identified by sending the associated ELIN to the PSAP. If the phones were evenly distributed in this example, each group of 10 phones would have the same ERL and, therefore, the same ELIN.

The various legislations define a minimum number of phones (for example, 49) and a minimum surface area (for example, 40,000 sq ft) below which the requirements for MLTS 911 are not applicable. But even if the legislation does not require 911 functionality for a given enterprise, it is always best practice to provision for it.

Emergency Location Identification Number Mapping

In general, you must associate a single fully qualified E.164 number, known as the emergency location identification number (ELIN), with each ERL. (However, if using Cisco Emergency Responder, you can configure more than one ELIN per ERL.) The ELIN is used to route the call across the E911 infrastructure and is used by the PSAP as the index into the ALI database.

ELINs must meet the following requirements:

- They must be routable across the E911 infrastructure. (See the examples in the section on Interface Type, page 10-5.) If an ELIN is not routable, 911 calls from the associated ERL will, at best, be handled according to the default routing programmed in the E911 selective router.
- Once the ERL-to-ELIN mapping of an enterprise is defined, the corresponding ALI records must be established with the LEC so that the ANI and ALI database records serving the PSAP can be updated accurately.

The ELIN mapping process can be one of the following, depending on the type of interface to the E911 infrastructure for a given ERL:

- Dynamic ANI interface

  With this type of interface, the calling party number identification passed to the network is controlled by the MLTS. The telephony routing table of the MLTS is responsible for associating the correct ELIN with the call, based on the calling phone's ERL. Within Cisco Unified Communications Manager, the calling party number can be modified by using transformation masks for calls made to 911. For example, all phones located in a given ERL can share the same calling search space that lists a partition containing a translation pattern (911) and a calling party transformation mask that would replace the phone's CPN with the ELIN for that location.

- Static ANI interface

  With this type of interface, the calling party number identification passed to the network is controlled by the PSTN. This is the case if the interface is a POTS line. The ELIN is the phone number of the POTS line, and no further manipulation of the phone's calling party identification number is possible.

PSAP Callback

The PSAP might have to reach the caller after completion of the initial conversation. The PSAP's ability to call back relies on the information that it receives with the original incoming call.

The delivery of this information to the PSAP is a two-part process:

1. The Automatic Number Identification (ANI) is first sent to the PSAP. The ANI is the E.164 number used to route the call. In our context, the ANI received at the PSAP is the ELIN that the MLTS sent.
2. The PSAP then uses the ANI to query a database and retrieve the Automatic Location Identification (ALI). The ALI provides the PSAP attendant with information such as:
   - Caller's name
   - Address
   - Applicable public safety agency
   - Other optional information, which could include callback information. For example, the phone number of the enterprise's security service could be listed, to aid in the coordination of rescue efforts.

Typical Situation

- The ANI information is used for PSAP callback, which assumes that the ELINs are dialable numbers.
- The ELINs are PSTN numbers associated with the MLTS. If someone calls the ELIN from the PSTN, the call will terminate on an interface controlled by the MLTS.
- It is the responsibility of the MLTS system administrator to program the call routing so that calls made to any ELIN in the system will ring a phone (or multiple phones) in the immediate vicinity of the associated ERL.

- Once the ERL-to-ELIN mapping is established, it needs be modified only when there are changes to the physical situation of the enterprise. If phones are simply added, moved, or deleted from the system, the ERL-to-ELIN mapping and its associated ANI/ALI database records need not be changed.

Exceptional Situation

- Callback to the immediate vicinity of the originating ERL may be combined with (or even superseded by) routing the callback to an on-site emergency desk, which will assist the PSAP in reaching the original caller and/or provide additional assistance with the emergency situation at hand.
- The situation of the enterprise could change, for example, due to area code splits, city or county service changes requiring a new distribution of the public safety responsibilities, new buildings being added, or any other change that would affect the desired routing of a call for 911 purposes. Any of these evens could require changes in the ERL-to-ELIN mapping and the ANI/ALI database records for the enterprise.

Dial Plan Considerations

It is highly desirable to configure a dial plan so that the system easily recognizes emergency calls, whether an access code (for example, 9) is used or not. The emergency string for North America is generally 911. Cisco strongly recommends that you configure the system to recognize both the strings 911 and 9911.

Cisco also strongly recommends that you explicitly mark the emergency route patterns with Urgent Priority so that Unified CM does *not* wait for the inter-digit timeout (Timer T.302) before routing the call.

Other emergency call strings may be supported concurrently on your system. Cisco highly recommends that you provide your telephony system users with training on the selected emergency call strings.

Also, it is highly desirable that users be trained to react appropriately if they dial the emergency string by mistake. In North America, 911 may be dialed in error by users trying to access a long distance number through the use of 9 as an access code. In such a case, the user should remain on the line to confirm that there is no emergency, and therefore no need to dispatch emergency personnel. Cisco ER's on-site notification capabilities can help in identifying the phone at the origin of such spurious 911 calls by providing detailed accounts of all calls made to 911, including calls made by mistake.

In a multisite deployment, the dial plan configuration should ensure that the emergency calls are always routed through the PSTN gateway local to the site, thereby making sure that the emergency call is routed to the nearest PSAP within the jurisdiction. One of the mechanism to achieve this could be to use the Local Route Group feature of Cisco Unified CM.

Also, in a multisite deployment it is very important to make sure that the emergency number is always reachable and routed through the local PSTN gateway for the mobility users (extension mobility and device mobility) independent of the implemented Class of Service (CoS). If the site/device approach is being used, the device calling search space (CSS) could be used to route the emergency calls.

Cisco recommends enabling Calling Party Modification on Cisco Emergency Responder. When this feature is enabled, the calling party number is replaced with the ELIN by Cisco ER for the emergency call. If Calling Party Modification is not enabled, either the DID will be sent to the PSAP or Cisco Unified CM must be configured to replace the calling party with the ELIN.

Note Except for manually configured phones, E.164 numbers are supported with Cisco Emergency Responder. For manually configured phones, Cisco recommends configuring the phones on Cisco Unified CM with the E.164 numbers without the leading "+".

Gateway Considerations

Consider the following factors when selecting the gateways to handle emergency calls for your system:

- Gateway Placement, page 10-13
- Gateway Blocking, page 10-13
- Answer Supervision, page 10-14
- Answer Supervision, page 10-14

Gateway Placement

Within the local exchange carrier (LEC) networks, 911 calls are routed over a locally significant infrastructure based on the origin of the call. The serving Class 5 switches are connected either directly to the relevant PSAP for their location or to an E911 selective router, which itself is connected to a group of PSAPs significant for its region.

With Cisco's IP-based enterprise telephony architecture, it is possible to route calls on-net to gateways that are remotely situated. As an example, a phone located in San Francisco could have its calls carried over an IP network to a gateway situated in San Jose, and then sent to the LEC's network.

For 911 calls, it is critical to chose the egress point to the LEC network so that emergency calls are routed to the appropriate local PSAP. In the example above, a 911 call from the San Francisco phone, if routed to a San Jose gateway, could not reach the San Francisco PSAP because the San Jose LEC switch receiving the call does not have a link to the E911 selective router serving the San Francisco PSAP. Furthermore, the San Jose area 911 infrastructure would not be able to route the call based on a San Francisco calling party number.

As a rule of thumb, route 911 calls to a gateway physically co-located with the originating phone. Contact the LEC to explore the possibility of using a common gateway to aggregate the 911 calls from multiple locations. Be aware that, even if the 911 network in a given region lends itself to using a centralized gateway for 911 calls, it might be preferable to rely on gateways co-located with the calling phones to prevent 911 call routing from being impacted during WAN failures.

Gateway Blocking

It is highly desirable to protect 911 calls from "all trunks busy" situations. If a 911 call needs to be connected, it should be allowed to proceed even if other types of calls are blocked due to lack of trunking resources. To provide for such situations, you can dedicate an explicit trunk group just for 911 calls.

It is acceptable to route emergency calls exclusively to an emergency trunk group. Another approach is to send emergency calls to the same trunk group as the regular PSTN calls (if the interface permits it), with an alternative path to a dedicated emergency trunk group. This latter approach allows for the most flexibility.

As an example, we can point emergency calls to a PRI trunk group, with an alternate path (reserved exclusively for emergency calls) to POTS lines for overflow conditions. If we put 2 POTS lines in the alternate trunk group, we are guarantying that a minimum of two simultaneous 911 calls can be routed in addition to any calls that were allowed in the main trunk group.

If the preferred gateway becomes unavailable, it may be acceptable to overflow emergency calls to an alternate number so that an alternate gateway is used. For example, in North America calls dialed as 911 could overflow to an E.164 (non-911) local emergency number. This approach does not take advantage of the North American 911 network infrastructure (that is, there is no selective routing, ANI, or ALI services), and it should be used only if it is acceptable to the applicable public safety authorities and only as a last resort to avoid blocking the emergency call due to a lack of network resources.

Answer Supervision

Under normal conditions, calls made to an emergency number should return answer supervision upon connection to the PSAP. The answer supervision may, as with any other call, trigger the full-duplex audio connection between the on-net caller and the egress interface to the LEC's network.

With some North American LECs, answer supervision might not be returned when a "free" call is placed. This may be the case for some toll-free numbers (for example, 800 numbers). In exceptional situations, because emergency calls are considered "free" calls, answer supervision might not be returned upon connection to the PSAP. You can detect this situation simply by making a 911 test call. Upon connection to the PSAP, if audio is present, the call timer should record the duration of the ongoing call; if the call timer is absent, it is very likely that answer supervision was not returned. If answer supervision is not returned, Cisco highly recommends that you contact the LEC and report this situation because it is most likely not the desired functionality.

If this situation cannot be rectified by the Local Exchange Carrier, it would be advisable to configure the egress gateway *not* to require answer supervision when calls are placed to the LEC's network, and to cut through the audio in both directions so that progress indicator tones, intercept messages, and communications with the PSAP are possible even if answer supervision is not returned.

By default, Cisco IOS-based H.323 gateways must receive answer supervision in order to connect audio in both directions. To forego the need for answer supervision on these gateways, use the following commands:

- **progress_ind alert enable 8**

  This command provides the equivalent of receiving a progress indicator value of 8 (in-band information now available) when alerting is received. This command allows the POTS side of the gateway to connect audio toward the origin of the call.

- **voice rtp send-recv**

  This command allows audio cut-through in both the backward and forward directions before a Connect message is received from the destination switch. This command affects all Voice over IP (VoIP) calls when it is enabled.

Be advised that, in situations where answer supervision is not provided, the call detail records (CDRs) will not accurately reflect the connect time or duration of 911 calls. This inaccuracy can impede the ability of a call reporting system to document the relevant statistics properly for 911 calls.

In all cases, Cisco highly recommends that you test 911 call functionality from all call paths and verify that answer supervision is returned upon connection to the PSAP.

Cisco Emergency Responder Design Considerations

Device mobility brings about special design considerations for emergency calls. Cisco Emergency Responder (Cisco ER) can be used to track device mobility and to adapt the system's routing of emergency calls based on a device's dynamic physical location.

Device Mobility Across Call Admission Control Locations

In a centralized call processing deployment, Cisco ER can detect IP phone relocation and reassign relocated IP phones to appropriate ERLs automatically. However, Cisco Unified CM location-based call admission control for a relocated phone might not properly account for the WAN bandwidth usage of the phone in the new location, yielding possible over-subscription or under-subscription of WAN bandwidth resources. For example, if you physically move a phone from Branch A to Branch B, the phone's call admission control location remains the same (Location_A), and it is possible that calls made to 911 from that phone would be blocked due to call admission control denial if all available bandwidth to Location_A is in use for other calls. To avoid such blocking of calls, manual intervention might be required to adapt the device's location and region parameters.

Cisco Unified CM device mobility provides a way to update the phone's configuration automatically (including its calling search space and location information) in Unified CM to reflect its new physical location. If device mobility is not used, manual configuration changes may be necessary in Cisco Unified CM.

Default Emergency Response Location

If Cisco ER cannot directly determine the physical location of a phone, it assigns a default emergency response location (ERL) to the call. The default ERL points all such calls to a specific PSAP. Although there is no universal recommendation as to where calls should be sent when this situation occurs, it is usually desirable to choose a PSAP that is centrally located and that offers the largest public safety jurisdiction. It is also advisable to populate the ALI records of the default ERL’s emergency location identification numbers (ELINs) with contact information for the enterprise's emergency numbers and to offer information about the uncertainty of the caller's location. In addition, it is advisable to mark those ALI records with a note that a default routing of the emergency call has occurred. Alternatively, the call may be routed to an internal security office to determine the caller's location.

Cisco Emergency Responder and Extension Mobility

Cisco ER supports Extension Mobility within a Cisco Unified CM cluster. It can also support Extension Mobility Cross-Cluster (EMCC), provided that both Cisco Unified CM clusters are supported either by a common Cisco ER server or group, or by two Cisco ER servers or groups configured as a Cisco ER cluster. In either case, the Cisco Unified CM clusters must not be configured to use the Adjunct Calling Search Space (CSS) associated with EMCC for 911 calls, but must be configured to use Cisco ER for all 911 calls in both Cisco Unified CM clusters.

Soft Clients

In cases where soft clients such as Cisco IP Communicator are used within the enterprise, Cisco ER can provide device mobility support. However, if the soft client is used outside the boundaries of the enterprise (for example, VPN access from a home office or hotel), Cisco ER will not be able to determine the location of the caller. Furthermore, it is unlikely that the Cisco system would have a gateway properly situated to allow sending the call to the appropriate PSAP for the caller's location.

It is a matter of enterprise policy to allow or not to allow the use of soft clients for 911 calls. It may be advisable to disallow 911 calls by policy for soft clients that can roam across the internet. Nevertheless, if such a user were to call 911, the best-effort system response would be to route the call to either an on-site security force or a large PSAP close to the system's main site.

The following paragraph is an example notice that you could issue to users to warn them that emergency call functionality is not guaranteed to soft client users:

> *Emergency calls should be placed from phones that are located at the site for which they are configured (for example, your office). A local safety authority might not answer an emergency call placed from a phone that has been removed from its configured site. If you must use this phone for emergency calls while away from your configured site, be prepared to provide the answering public safety authority with specific information regarding your location. Use a phone that is locally configured to the site (for example, your hotel phone or your home phone) for emergency calls when traveling or telecommuting.*

Cisco ER also supports integration with Intrado V9-1-1, an emergency call delivery service that can reach almost any PSAP in the United States. With the combination of Cisco ER and Intrado V9-1-1, users of IP phones and softphones outside the enterprise can provide and update their locations through a web page provided by Cisco ER. Emergency calls from an off-premises location will then be delivered by Intrado to the appropriate PSAP for the caller's location.

Test Calls

For any enterprise telephony system, it is a good idea to test 911 call functionality, not only after the initial installation, but regularly, as a preventive measure.

The following suggestions can help you carry out the testing:

- Contact the PSAP to ask for permission before doing any tests, and provide them with the contact information of the individuals making the tests.
- During each call, indicate that it is *not* an actual emergency, just a test.
- Confirm the ANI and ALI that the call taker has on their screen.
- Confirm the PSAP to which the call was routed.
- Confirm that answer supervision was received by looking at the call duration timer on the IP phone. An active call timer is an indication that answer supervision is working properly.

PSAP Callback to Shared Directory Numbers

Cisco ER handles the routing of inbound calls made to emergency location identification numbers (ELINs). In cases where the line from which a 911 call was made is a shared directory number, the PSAP callback will cause all shared directory number appearances to ring. Any of the shared appearances can then answer the call, which means that it may not be the phone from which the 911 call originated.

Cisco Emergency Responder Deployment Models

Enterprise telephony systems based on multiple Unified CM clusters can benefit from the functionality of Cisco Emergency Responder (Cisco ER).

The *Cisco Emergency Responder Administration Guide* provides detailed descriptions of the terms used herein, as well as the background information required to support the following discussion. Of specific interest is the chapter on *Planning for Cisco Emergency Responder*. This documentation is available at

http://www.cisco.com/en/US/products/sw/voicesw/ps842/prod_maintenance_guides_list.html

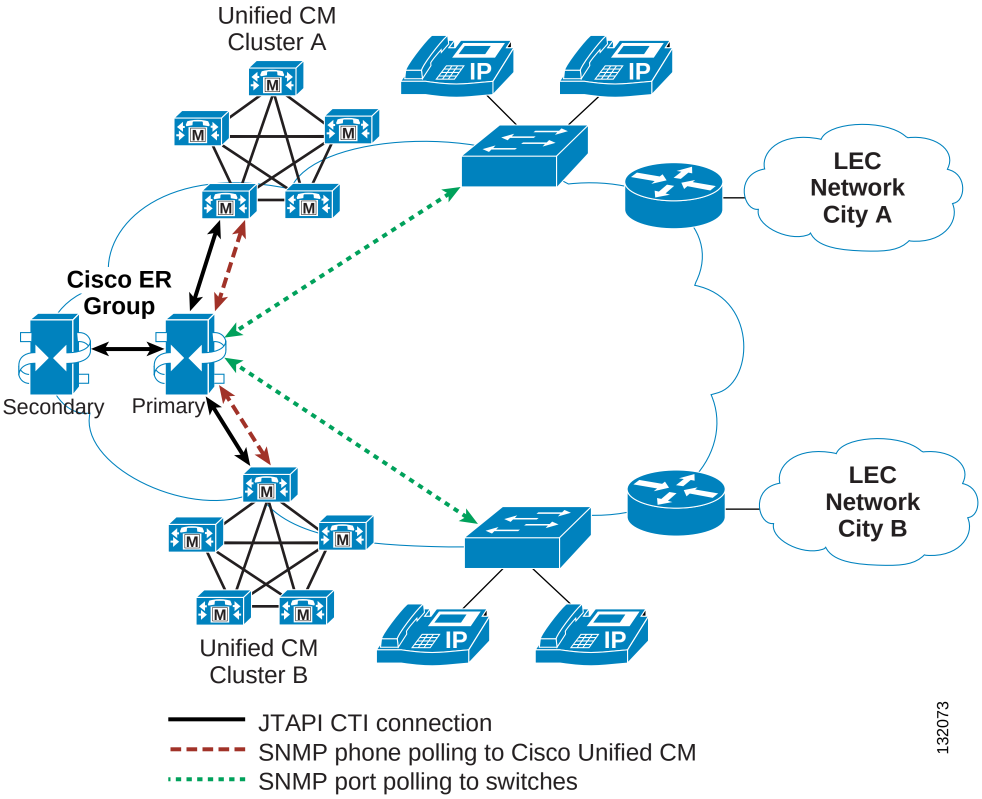
Note Cisco Emergency Responder does not support Cisco Unified Communication Manager Express (Unified CME) or Survivable Remote Mode Telephony (SRST). In case of SRST deployment, configure the appropriate dial-peer to route the 911 calls to the PSTN with the Site Published Number.

Single Cisco ER Group

A single Emergency Responder group can be deployed to handle emergency calls from two or more Unified CM clusters. The design goal is to ensure that any phone's emergency call is routed to the Cisco ER group, which will assign an ELIN and route the call to the appropriate gateway based on the phone's location.

One advantage of using a single Cisco ER group is that all ERLs and ELINs are configured into a single system. A phone registered on any cluster will be located by the single Cisco ER group because that group is responsible for polling all of the system's access switches. Figure 10-1 illustrates a single Cisco ER group interfaced with two Unified CM clusters.

Figure 10-1 A Single Cisco ER Group Connected to Two Unified CM Clusters



The single Cisco ER group in Figure 10-1 interfaces with the following components:

- Each Unified CM cluster via SNMP, to collect information about their respective configured phones
- All of the enterprise's switches via SNMP, so that any cluster's phone, connected to any switch, can be located. This connection is not required if the phone locations are being identified based on IP subnets. For details on configuring IP subnet-based ERLs, refer to the chapter on Configuring Cisco Emergency Responder in the *Cisco Emergency Responder Administration Guide*, available at

  http://www.cisco.com/en/US/products/sw/voicesw/ps842/prod_maintenance_guides_list.html
- Each Unified CM cluster via JTAPI, to allow for the call processing required by any phone that dials 911 – for example, identification of the calling phone's ERL, assignment of the ELIN, redirection of the call to the proper gateway (based on the calling phone's location), and the handling of the PSAP callback functionality

The version of the JTAPI interface used by Cisco Emergency Responder is determined by the version of the Unified CM software to which it is connected. At system initialization, Cisco ER interrogates the Unified CM cluster and loads the appropriate JTAPI Telephony Service Provider (TSP). Because there can be only one version of JTAPI TSP on the Cisco ER server, all Unified CM clusters to which a single Cisco ER group is interfaced *must* run the same version of Unified CM software.

For some deployments, this software version requirement might present some difficulties. For instance, during a Unified CM upgrade, different clusters will be running different versions of software, and some of the clusters will be running a version of JTAPI that is not compatible with the version running on the Cisco ER servers. When this situation occurs, emergency calls from the cluster running a version of JTAPI different than that of the Cisco ER group might receive the call treatment provided by the Call Forward Busy settings of the emergency number's CTI Route Point.

When considering if a single Cisco ER group is appropriate for multiple Unified CM clusters, apply the following guidelines:

- Make Unified CM upgrades during an acceptable maintenance window when emergency call volumes are as low as possible (for example, after hours, when system use is at a minimum).
- Use a single Cisco ER group only if the quantity and size of the clusters allow for minimizing the amount of time when dissimilar versions of JTAPI are in use during software upgrades.

For example, a deployment with one large eight-server cluster in parallel with a small two-server cluster could be considered for use with a single Cisco ER group. In this case, it would be best to upgrade the large cluster first, thus minimizing the number of users (those served by the small cluster) that might be without Cisco ER service during the maintenance window of the upgrade. Furthermore, the small cluster's users can more appropriately be served by the temporary static routing of emergency calls in effect while Cisco ER is not reachable because they can be identified by the single ERL/ELIN assigned to all non-ER calls made during that time.

Multiple Cisco ER Groups

Multiple Cisco ER groups can also be deployed to support multi-cluster systems. In this case, each ER group interfaces with the following components:

- A Unified CM cluster via the following methods:
  - SNMP, to collect information about its configured phones
  - JTAPI, to allow for the call processing associated with redirection of the call to the proper gateway or, in the case of roaming phones, the proper Unified CM cluster
- The access switches (via SNMP) to which most of the phones associated with the Unified CM of the Cisco ER group are most likely to be connected

This approach allows Unified CM clusters to run different versions of software because each is interfaced to a separate Cisco ER group.

To allow phones to roam between various parts of the network and still be tracked by Cisco ER, you might have to configure the Cisco ER groups into a Cisco ER cluster. For details on Cisco ER clusters and groups, refer to the chapter on *Planning for Cisco Emergency Responder* in the *Cisco Emergency Responder Administration Guide*, available at

http://www.cisco.com/en/US/products/sw/voicesw/ps842/prod_maintenance_guides_list.html

Figure 10-2 presents a sample topology illustrating some of the basic concepts behind Cisco ER clustering.

Figure 10-2 Multiple Cisco ER Groups

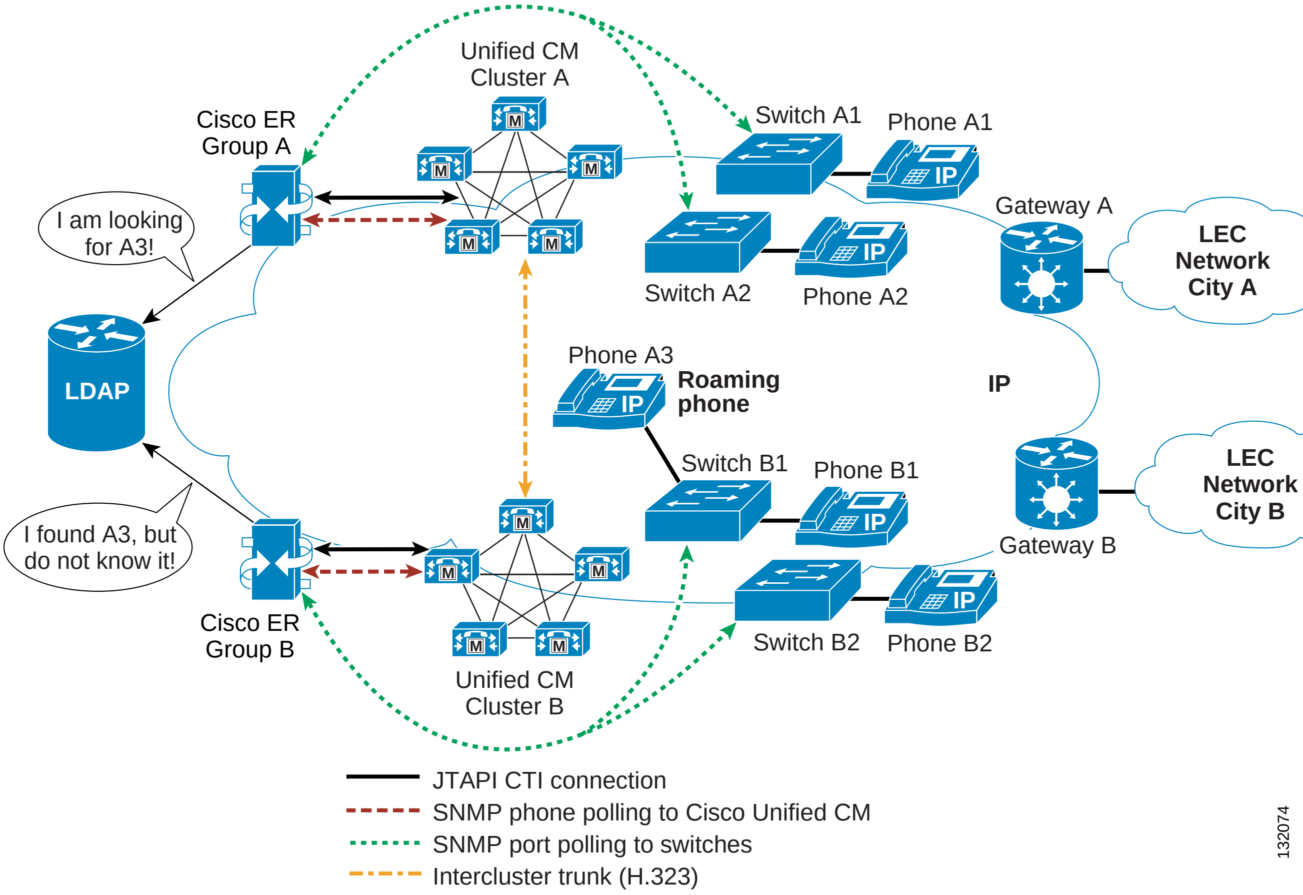


Figure 10-2 illustrates the following topology:

- Cisco ER group A is interfaced to Unified CM cluster A to access switches A1 and A2, and it is deemed to be the home Cisco ER group of all phones registered to Unified CM cluster A.
- Likewise, Cisco ER group B is interfaced to Unified CM cluster B to access switches B1 and B2, and it is deemed to be the home Cisco ER group of all phones registered to Unified CM cluster B.

Note Emergency Responder requires all ER groups in an ER cluster to run the same version of software.

Phone Movements Within the Tracking Domain of a Cisco ER Group

The emergency call processing for phones moving between access switches controlled by the same home Cisco ER group is the same as the processing done for a deployment with a single Unified CM cluster. For example, a phone moving between access switches A1 and A2 remains registered with Unified CM cluster A, and its location is determined by Cisco ER group A both before and after the move. The phone is still under full control of Cisco ER group A, for both the discovery of the phone by Unified CM cluster A and the determination of the phone's location by switch A2. The phone is therefore not considered to be an unlocated phone.

Phone Movements Between the Various Tracking Domains of a Cisco ER Cluster

A Cisco ER cluster is essentially a collection of Cisco ER groups that share location information. Each group shares the location of any phone it finds on an access switch or in an IP subnet.

Cisco ER groups also share information about phones that cannot be located within a Cisco ER group's tracking domain (in switches or IP subnets) but which are known to be registered in the group’s associated Unified CM cluster. Such phones are deemed *unlocated*.

If a phone is roaming between access switches monitored by different Cisco ER groups, those groups must be configured in a Cisco ER cluster so they can exchange information about the phone's location. For example, phone A3 is registered with Unified CM cluster A, but it is connected to an access switch controlled by Cisco ER group B. Cisco ER group A is aware that phone A3 is registered with Unified CM cluster A, but group A cannot locate phone A3 in any of the site A switches. Therefore, phone A3 is deemed *unlocated* by Cisco ER group A.

Cisco ER group B, on the other hand, has detected the presence of phone A3 in one of the switches that it monitors. Because the phone is not registered with Unified CM cluster B, phone A3 is advertised through the Cisco ER LDAP database as an *unknown* phone.

Because the two Cisco ER groups are communicating through an LDAP database, they can determine that Cisco ER group B's *unknown* phone A3 is the same as Cisco ER group A's *unlocated* phone A3.

The Unlocated Phone page in Cisco ER group A will display the phone's host name along with the remote Cisco ER group (in this, case Cisco ER group B).

Emergency Call Routing within a Cisco ER Cluster

Cisco ER clustering also relies on route patterns that allow emergency calls to be redirected between pairs consisting of a Unified CM cluster and a Cisco ER. For more details, refer to the section on *Creating Route Patterns for Inter-Cisco Emergency Responder-Group Communications* in the *Cisco Emergency Responder Administration Guide*, available at

http://www.cisco.com/en/US/products/sw/voicesw/ps842/prod_maintenance_guides_list.html

If phone A3 places an emergency call, the call signaling flow will be as follows:

1. Phone A3 sends the emergency call string to Unified CM cluster A for processing.
2. Unified CM cluster A sends the call to Cisco ER group A for redirection.
3. Cisco ER group A determines that phone A3 is located in Cisco ER group B's tracking domain, so it redirects the call to a route pattern that points to Unified CM cluster B.
4. Unified CM cluster A sends the call to Unified CM cluster B.
5. Unified CM cluster B sends the call to Cisco ER group B for redirection.
6. Cisco ER group B identifies the ERL and ELIN associated with phone A3's location and redirects the call to Unified CM cluster B. The calling number is transformed into the ELIN associated with the ERL of phone A3, and the called number is modified to route the call to the proper gateway.
7. Unified CM cluster B routes the call according to the new called number information obtained from Cisco ER group B.
8. Unified CM cluster B sends the call out the gateway toward the Emergency PSTN network.

WAN Deployment of Cisco Emergency Responder

The Cisco Emergency Responder Group can be located remotely from the Cisco Unified CM cluster (that is, over the WAN). Also, the primary and secondary Cisco ER servers can be placed in geographically separate sites over the WAN. For such deployments, the recommended round-trip time (RTT) is 40 msec or less, and the minimum bandwidth required between the Cisco ER Servers is 1.544 Mbps.

ALI Formats

In multi-cluster configurations, there might be instances where the physical locations of ERLs and ELINs defined in a single Cisco ER group span the territory of more than one phone company. This condition can lead to situations where records destined for different phone companies have to be extracted from a common file that contains records for multiple LECs.

Cisco ER exports this information in ALI records that conform to National Emergency Number Association (NENA) 2.0, 2.1, and 3.0 formats. However, many service providers do not use NENA standards. In such cases, you can use the ALI Formatting Tool (AFT) to modify the ALI records generated by Cisco ER so that they conform to the formats specified by your service provider. That service provider can then use the reformatted file to update their ALI database.

The ALI Formatting Tool (AFT) enables you to perform the following functions:

- Select a record and update the values of the ALI fields. AFT allows you to edit the ALI fields to customize them to meet the requirements of various service providers. You service provider can then read the reformatted ALI files and use them to update their ELIN records.
- Perform bulk updates on multiple ALI records. Using the bulk update feature, you can apply common changes to all the records that you have selected, to one area code, or to one area code and one city code.
- Selectively export ALI records based on area code, city code, or a four-digit directory number. By selecting to export all the ALI records in an area code, for example, you can quickly access all the ELIN records for each service provider, thereby easily supporting multiple service providers.

Given the flexibility of the AFT, a single Cisco ER group can export ALI records in multiple ALI database formats. For a Cisco ER group serving a Unified CM cluster with sites in the territories of two LECs, the basic approach is as follows:

1. Obtain an ALI record file output from Cisco Emergency Responder in standard NENA format. This file contains the records destined for multiple LECs.
2. Make a copy of the original file for each required ALI format (one copy per LEC).
3. Using the AFT of the first LEC (for example, LEC-A), load a copy of the NENA-formatted file and delete the records of all the ELINs associated with the other LECs. The information to delete can usually be identified by NPA (or area code).
4. Save the resulting file in the required ALI format for LEC-A, and name the file accordingly.
5. Repeat steps 3 and 4 for each LEC.

For more information about the ALI formatting tools, refer to the online documentation available at

http://www.cisco.com/en/US/products/sw/voicesw/ps842/prod_maintenance_guides_list.html

For LECs not listed at this URL, the output from Unified CM can be formatted using standard text file editing tools, such as spreadsheet programs and standard text editors.

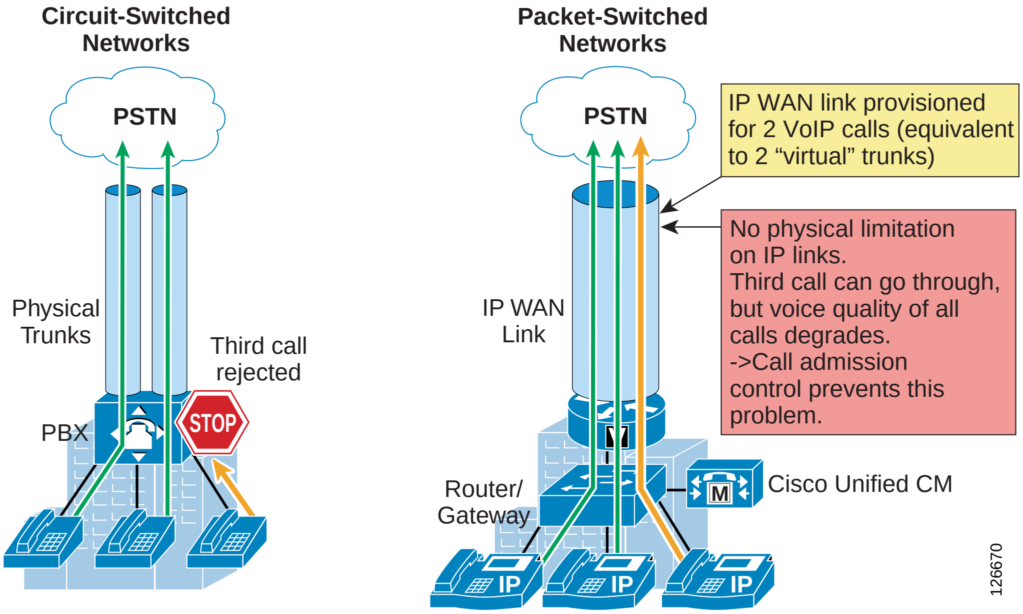
CHAPTER 11

Call Admission Control

Revised: June 28, 2012; OL-27282-05

The call admission control function is an essential component of any IP telephony system, especially those that involve multiple sites connected through an IP WAN. In order to better understand what call admission control does and why it is needed, consider the example in Figure 11-1.

Figure 11-1 Why Call Admission Control is Needed



Circuit-Switched Networks
PSTN
Physical Trunks
Third call rejected
PBX
STOP
Packet-Switched Networks
PSTN
IP WAN link provisioned for 2 VoIP calls (equivalent to 2 "virtual" trunks)
No physical limitation on IP links. Third call can go through, but voice quality of all calls degrades. ->Call admission control prevents this problem.
IP WAN Link
Router/ Gateway
Cisco Unified CM
M
IP
IP
IP
126670

As shown on the left side of Figure 11-1, traditional TDM-based PBXs operate within circuit-switched networks, where a circuit is established each time a call is set up. As a consequence, when a legacy PBX is connected to the PSTN or to another PBX, a certain number of physical trunks must be provisioned. When calls have to be set up to the PSTN or to another PBX, the PBX selects a trunk from those that are available. If no trunks are available, the call is rejected by the PBX and the caller hears a network-busy signal.

Now consider the IP telephony system shown on the right side of Figure 11-1. Because it is based on a packet-switched network (the IP network), no circuits are established to set up an IP telephony call. Instead, the IP packets containing the voice samples are simply routed across the IP network together with other types of data packets. Quality of Service (QoS) is used to differentiate the voice packets from the data packets, but bandwidth resources, especially on IP WAN links, are not infinite. Therefore, network administrators dedicate a certain amount of "priority" bandwidth to voice traffic on each IP WAN link. However, once the provisioned bandwidth has been fully utilized, the IP telephony system must reject subsequent calls to avoid oversubscription of the priority queue on the IP WAN link, which would cause quality degradation for all voice calls. This function is known as call admission control, and it is essential to guarantee good voice quality in a multisite deployment involving an IP WAN.

To preserve a satisfactory end-user experience, the call admission control function should always be performed during the call setup phase so that, if there are no network resources available, a message can be presented to the end-user or the call can be rerouted across a different network (such as the PSTN).

This chapter discusses the following main topics:

- Call Admission Control Principles, page 11-3

  This section defines the two fundamental approaches to call admission control in an IP-based telephony system: topology-aware and topology-unaware call admission control.

- Call Admission Control Architecture, page 11-12

  This section describes the call admission control mechanisms available through the various components of a Cisco IP Communications system, such as Cisco Unified Communications Manager Enhanced Locations call admission control, Cisco IOS gatekeeper, RSVP, and RSVP SIP Preconditions.

- Design Considerations for Call Admission Control, page 11-93

  This section shows how to apply and combine the mechanisms described in the previous sections, based on the IP WAN topology.

What's New in This Chapter

Table 11-1 lists the topics that are new in this chapter or that have changed significantly from previous releases of this document.

Table 11-1 New or Changed Information Since the Previous Release of This Document

New or Revised Topic	Described in:	Revision Date
Enhanced Locations CAC	Unified CM Enhanced Locations Call Admission Control, page 11-12	June 28, 2012
RSVP Agent support for RTCP, BFCP and FECC	RSVP Agent Support for RTCP, BFCP and FECC Negotiation, page 11-67	June 28, 2012
Migration to RSVP Agent call admission control	Migrating to RSVP Call Admission Control, page 11-70	June 28, 2012
Example WAN topologies for call admission control	Design Considerations for Call Admission Control, page 11-93	June 28, 2012
Other updates for Cisco Unified Communications System Release 9.0	Various sections throughout this chapter	June 28, 2012

Call Admission Control Principles

As mentioned previously, call admission control is a function of the call processing agent in an IP-based telephony system, so in theory there could be as many call admission control mechanisms as there are IP-based telephony systems. However, most of the existing call admission control mechanisms fall into one of the following two main categories:

- Topology-unaware call admission control — Based on a static configuration within the call processing agent
- Topology-aware call admission control — Based on communication between the call processing agent and the network about the available resources

The remainder of this section first analyzes the principles of topology-unaware call admission control and its limitations, and then presents the principles of topology-aware call admission control.

Topology-Unaware Call Admission Control

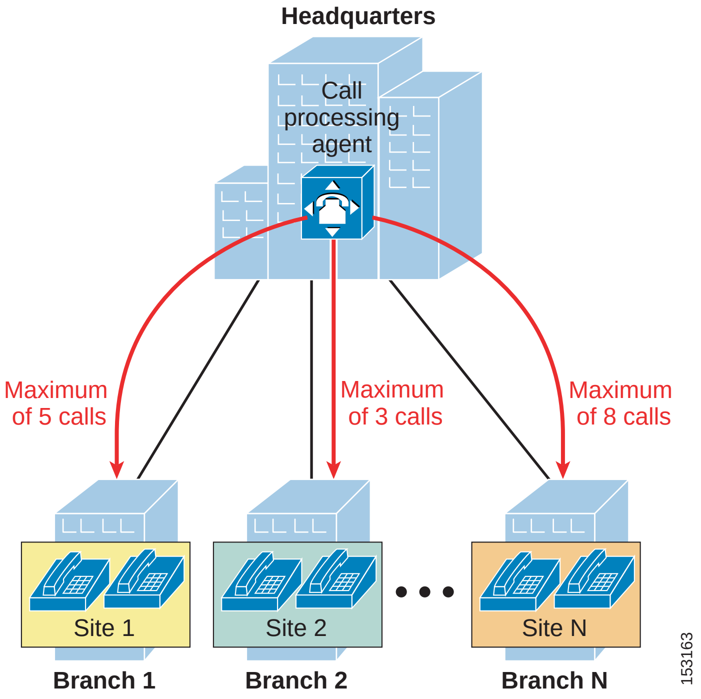
We define as topology-unaware call admission control any mechanism that is based on a static configuration within a call processing agent or IP-based PBX, aimed at limiting the number of simultaneous calls to or from a remote site connected via the IP WAN.

As shown in Figure 11-2, most of these mechanisms rely on the definition of a logical "site" entity, which generally corresponds to a geographical branch office connected to the enterprise IP WAN.

After assigning all the devices located at each branch office to the corresponding site entity, the administrator usually configures a maximum number of calls (or a maximum amount of bandwidth) to be allowed in or out of that site.

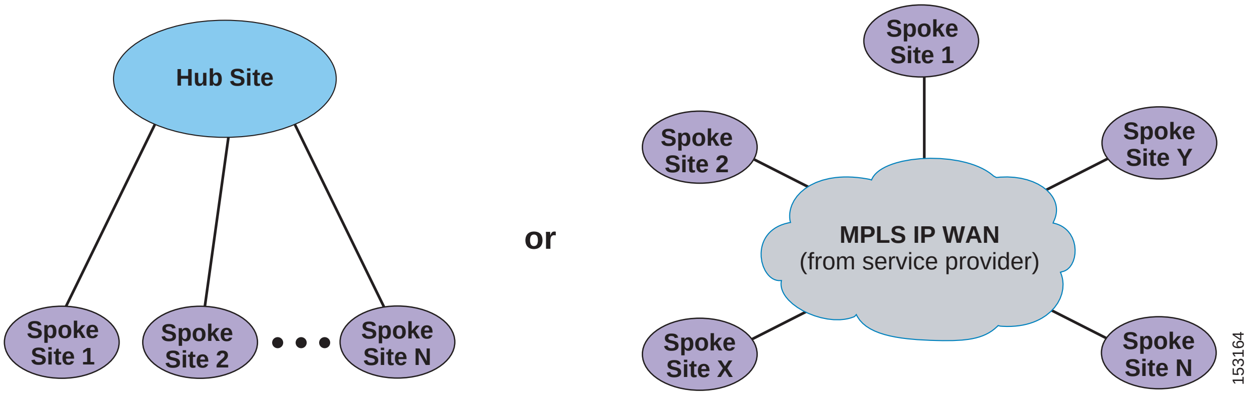
Each time a new call needs to be established, the call processing agent checks the sites to which the originating and terminating endpoints belong, and verifies whether there are available resources to place the call (in terms of number of calls or amount of bandwidth for both sites involved). If the check succeeds, the call is established and the counters for both sites are decremented. If the check fails, the call processing agent can decide how to handle the call based on a pre-configured policy. For example, it could send a network-busy signal to the caller device, or it could attempt to reroute the call over a PSTN connection.

Figure 11-2 Principles of Topology-Unaware Call Admission Control



Because of their reliance on static configurations, topology-unaware call admission control mechanisms can generally be deployed only in networks with a relatively simple IP WAN topology. In fact, most of these mechanisms mandate a simple hub-and-spoke topology or a simple MPLS-based topology (where the MPLS service is provided by a service provider), as shown in Figure 11-3.

Figure 11-3 Domain of Applicability of Topology-Unaware Call Admission Control



In a hub-and-spoke network or MPLS-based network such as those shown in Figure 11-3, each spoke site is assigned to a "site" within the call processing agent, and the number of calls or amount of bandwidth for that "site" is configured to match the bandwidth available for voice (and/or video) on the IP WAN link that connects the spoke to the IP WAN.

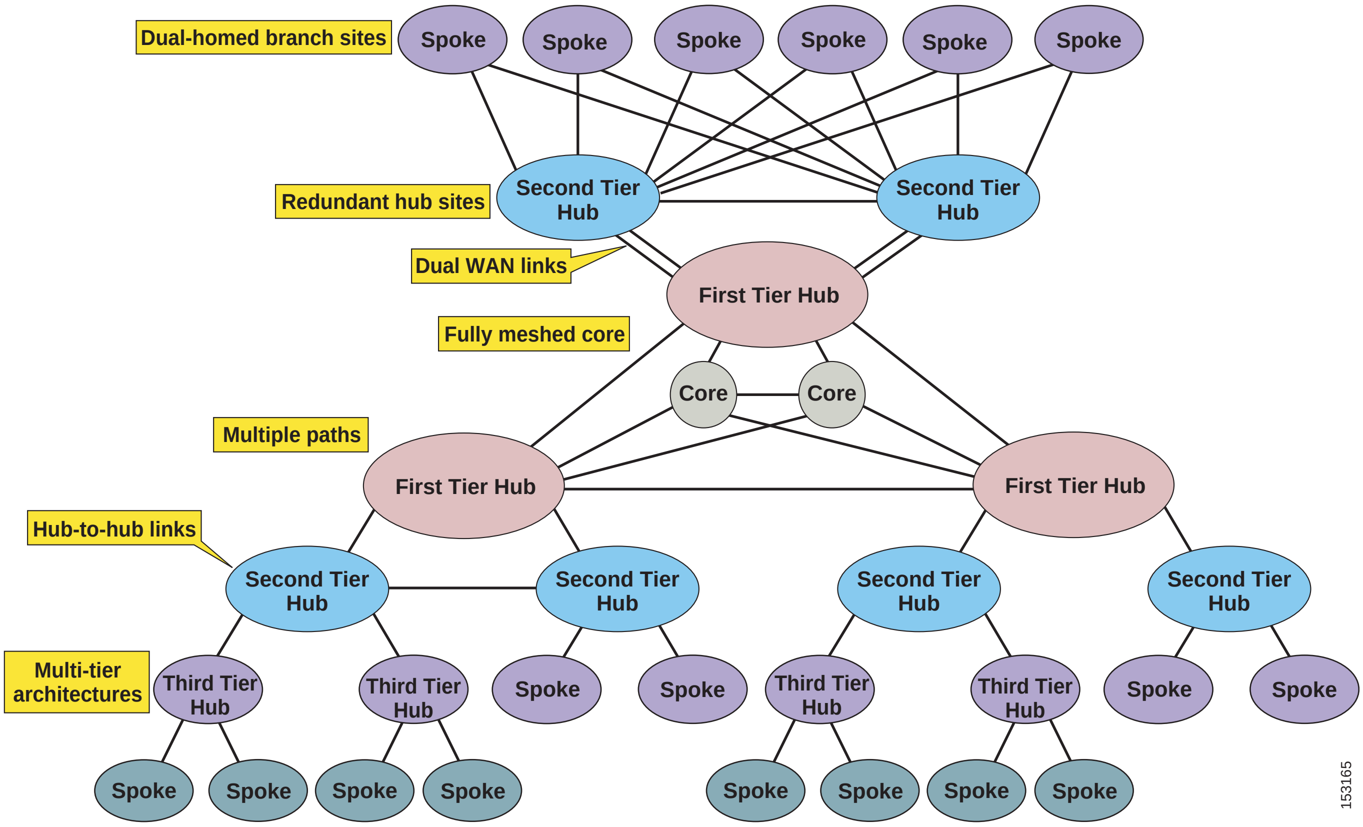
Notice the absence of redundant links from the spoke sites to the hub site and of links directly connecting two spoke sites. The next section explains why such links create problems for topology-unaware call admission control.

Limitations of Topology-Unaware Call Admission Control

In today's enterprise networks, high availability is a common requirement, and it often translates into a desire to provide redundancy for the IP WAN network connectivity.

When considering the IP WAN topology in a typical enterprise network, you are likely to encounter a number of characteristics that complicate the assumption of a pure hub-and-spoke topology. Figure 11-4 shows several of these network characteristics in a single diagram. Obviously, only the largest enterprise networks present all these characteristics at once, but it is highly likely that most IP WAN networks feature at least one of them.

Figure 11-4 Topology Characteristics of Typical Enterprise Networks

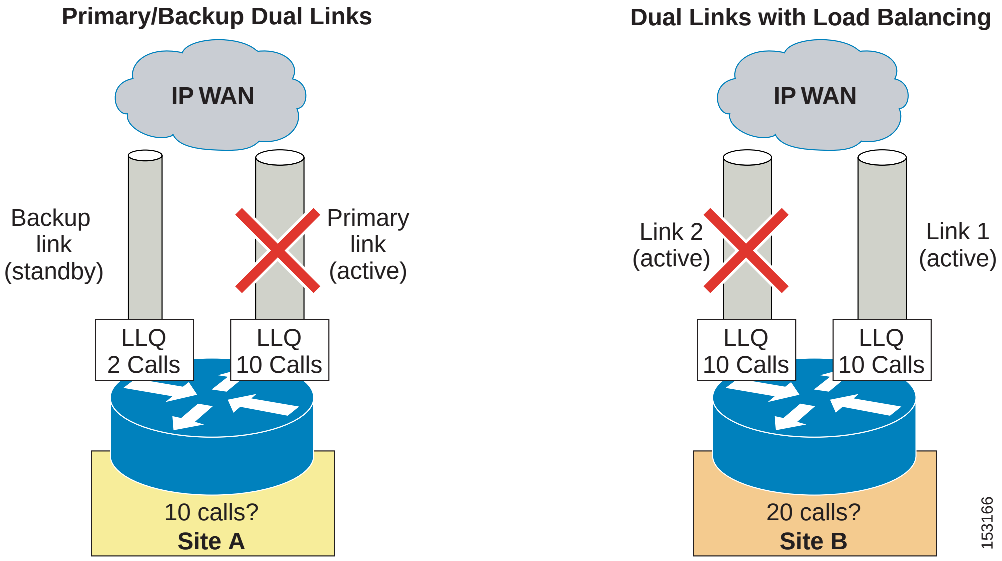


As explained in the section on Design Considerations for Call Admission Control, page 11-93, it is sometimes possible to adapt a topology-unaware call admission control mechanism to a complex network topology, but there are limitations in terms of when this approach can be used and what behavior can be achieved. For example, consider the simple case of a branch site connected to a hub site via the IP WAN, where redundancy is a network requirement. Typically, redundancy can be achieved in one of the following ways:

- A single router with a primary and a backup link to the IP WAN
- A single router with two active WAN links in a load-balancing configuration
- Two router platforms, each connected to the IP WAN, with load-balanced routing across them

The examples Figure 11-5 attempt to apply a topology-unaware call admission control mechanism to the case of a single router with a primary and backup link and the case of a single router with two active load-balanced links. (The case of two router platforms has the same call admission control implications as the latter example.)

Figure 11-5 Topology-Unaware Call Admission Control in Presence of Dual Links



For the first example in Figure 11-5, branch office A is normally connected to the IP WAN via a primary link, whose Low Latency Queuing (LLQ) bandwidth is provisioned to allow a maximum of 10 simultaneous calls. When this primary link fails, a smaller backup link becomes active and preserves the connectivity to the IP WAN. However, the LLQ bandwidth of this backup link is provisioned to allow only up to 2 simultaneous calls.

In order to deploy a topology-unaware call admission control mechanism for this branch office, we must define a "site" A in the call processing agent and configure it for a certain number of calls (or amount of bandwidth). If we choose to use 10 calls as the maximum for site A, the backup link can be overrun during failures of the primary link, thereby causing bad voice quality for all active calls. If, on the other hand, we choose 2 calls as the maximum, we will not be able to use the bandwidth provisioned for the remaining 8 calls when the primary link is active.

Now consider branch office B, which has two active links connecting it to the IP WAN. Each of these links is provisioned to allow a maximum of 10 simultaneous calls, and the routing protocol automatically performs load-balancing between them. When deploying a topology-unaware call admission control mechanism for this branch office, we must define a "site" B in the call processing agent and configure it for a certain number of calls (or amount of bandwidth). Similar to the case of branch office A, if we choose to add up the capacity of the two links and use 20 calls as the maximum for site B, there is a potential to overrun the LLQ on one of the two links during failures of the other one. For example, if link #2 fails, the system still allows 20 simultaneous calls to and from site B, which are now all routed via link #1, thus overrunning it and causing poor voice quality for all calls. On the other hand, if site B is configured for a maximum of 10 simultaneous calls, the available LLQ bandwidth is never fully utilized under normal conditions (when both links are operational).

These two simple examples show how IP WAN bandwidth provisioning in real enterprise networks is often too complex to be summarized in statically configured entries within the call processing agent. Deploying topology-unaware call admission control in such networks forces the administrator to make assumptions, develop workarounds, or accept sub-optimal use of network resources.

The optimal way to provide call admission control in the presence of a network topology that does not conform to a simple hub-and-spoke is to implement topology-aware call admission control, as described in the following section.

Note Some IP telephony systems augment classic topology-unaware call admission control with a feedback mechanism based on observed congestion in the network, which forces calls through the PSTN when voice quality deteriorates. This approach is still not equivalent to true topology-aware call admission control because it is performed after the calls have already been established and because the call processing agent still does not have knowledge of exactly where congestion is occurring. As mentioned at the beginning of the chapter, in order to be effective, call admission control must be performed before the call is set up.

Topology-Aware Call Admission Control

We define as topology-aware call admission control any mechanism aimed at limiting the number of simultaneous calls across IP WAN links that can be applied to any network topology and can dynamically adjust to topology changes.

To accomplish these goals, topology-aware call admission control must rely on real-time communications about the availability of network resources between a call processing agent (or IP-based PBX) and the network. Because the network is a distributed entity, real-time communications require a signaling protocol.

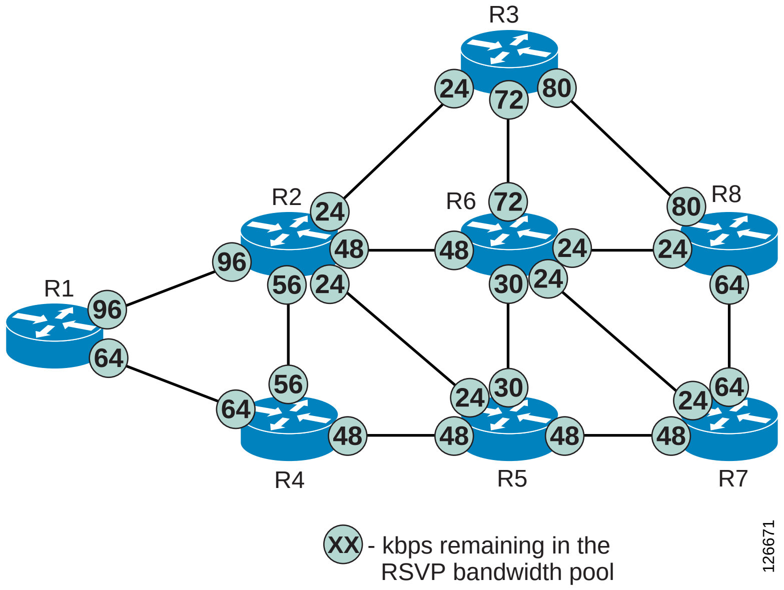
The Resource Reservation Protocol (RSVP) is the first significant industry-standard signaling protocol that enables an application to reserve bandwidth dynamically across an IP network. Using RSVP, applications can request a certain amount of bandwidth for a data flow across a network (for example, a voice call) and can receive an indication of the outcome of the reservation based on actual resource availability.

In the specific case of call admission control for voice or video calls, an IP-based PBX can synchronize the call setup process with RSVP reservations between the two remote sites and can make a routing decision based on the outcome of the reservations. Because of its distributed and dynamic nature, RSVP is capable of reserving bandwidth across any network topology, thus providing a real topology-aware call admission control mechanism.

To better understand the basic principles of how RSVP performs bandwidth reservation in a network, consider the simple example depicted in Figure 11-6. This example does not analyze the exact message exchanges and protocol behaviors, but rather focus on the end results from a functionality perspective. For more information on the RSVP message exchanges, see RSVP Principles, page 11-42.

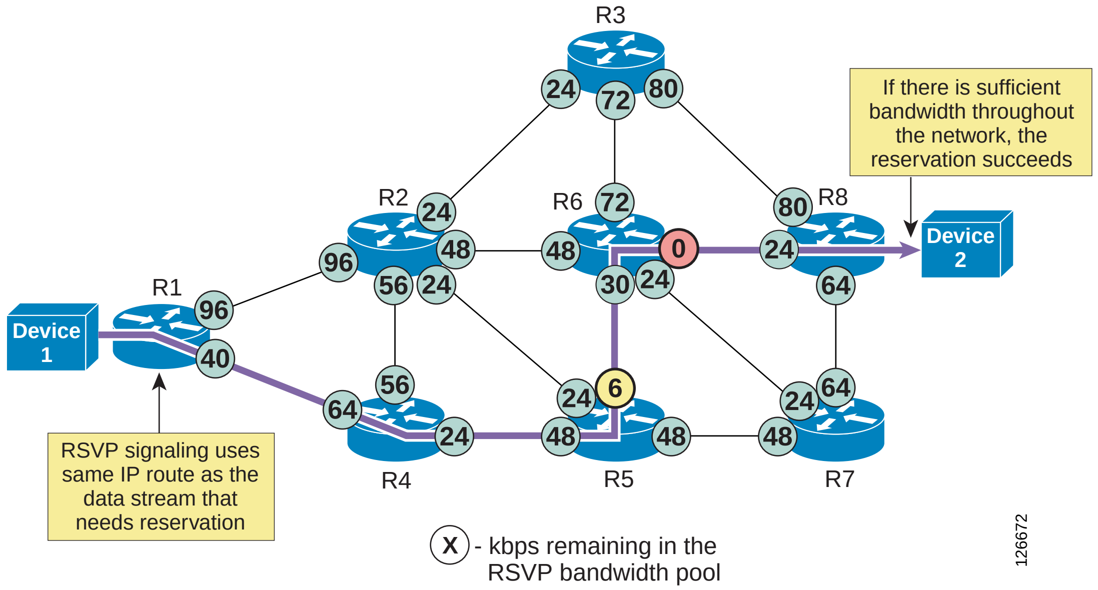
Assume that RSVP is enabled on each router interface in the network shown in Figure 11-6 and that the numbers shown in the circles represent the amount of available RSVP bandwidth remaining on each interface.

Figure 11-6 Sample Network to Show RSVP Principles



Now consider an RSVP-enabled application that wants to reserve a certain amount of bandwidth for a data stream between two devices. This scenario is depicted in Figure 11-7, which shows a particular data stream that requires 24 kbps of bandwidth from Device 1 to Device 2.

Figure 11-7 RSVP Signaling for a Successful Reservation

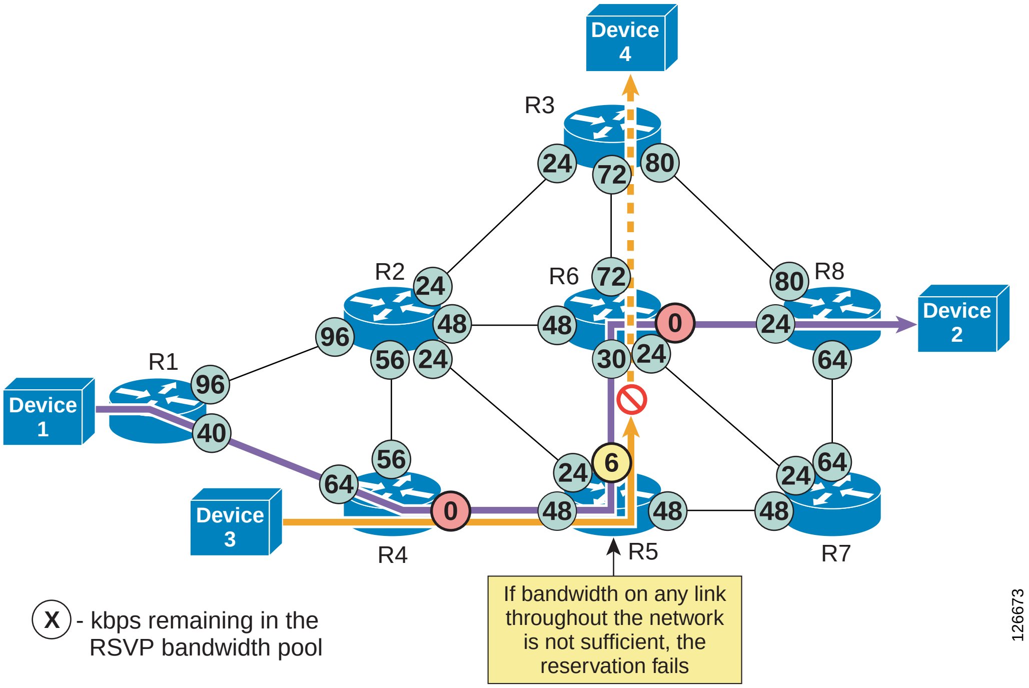

R3
24
72
80
If there is sufficient bandwidth throughout the network, the reservation succeeds
R2
24
R6
72
80
R8
96
48
48
0
24
Device 2
56
24
30
24
64
R1
Device 1
96
40
56
64
24
6
24
48
48
24
64
48
RSVP signaling uses same IP route as the data stream that needs reservation
R4
R5
R7
X - kbps remaining in the RSVP bandwidth pool
126672

The following considerations apply to Figure 11-7:

- RSVP does not perform its own routing; instead it uses underlying routing protocols to determine where it should carry reservation requests. As routing changes paths to adapt to topology changes, RSVP adapts its reservations to the new paths wherever reservations are in place.
- The RSVP protocol attempts to establish an end-to-end reservation by checking for available bandwidth resources on all RSVP-enabled routers along the path from Device 1 to Device 2. As the RSVP messages progress through the network, the available RSVP bandwidth gets decremented by 24 kbps on the outbound router interfaces, as shown in Figure 11-7.
- The available bandwidth on all outbound interfaces is sufficient to accept the new data stream, so the reservation succeeds and the application is notified.
- RSVP reservations are unidirectional (in this case, the reservation is established from Device 1 to Device 2, and not vice versa). In the presence of bidirectional applications such as voice and videoconferencing, two reservations must be established, one in each direction.
- RSVP provides transparent operation through router nodes that do not support RSVP. If there are any routers along the path that are not RSVP-enabled, they simply ignore the RSVP messages and pass them along like any other IP packet, and a reservation can still be established. (See RSVP Principles, page 11-42, for details on protocol messages and behaviors.) However, in order to have an end-to-end QoS guarantee, you have to ensure that there is no possibility of bandwidth congestion on the links controlled by the non-RSVP routers.

After a reservation has been successfully established between Device 1 and Device 2, now assume that another application requests a 24-kbps reservation between Device 3 and Device 4, as depicted in Figure 11-8.

Figure 11-8 RSVP Signaling for an Unsuccessful Reservation


Device 4
R3
24
72
80
R2
24
R6
72
R8
80
Device 2
96
48
48
0
24
56
24
30
24
64
R1
96
Device 1
40
56
64
24
6
24
64
Device 3
0
48
48
48
R4
R5
R7
X - kbps remaining in the RSVP bandwidth pool
If bandwidth on any link throughout the network is not sufficient, the reservation fails
126673

The following considerations apply to Figure 11-8:

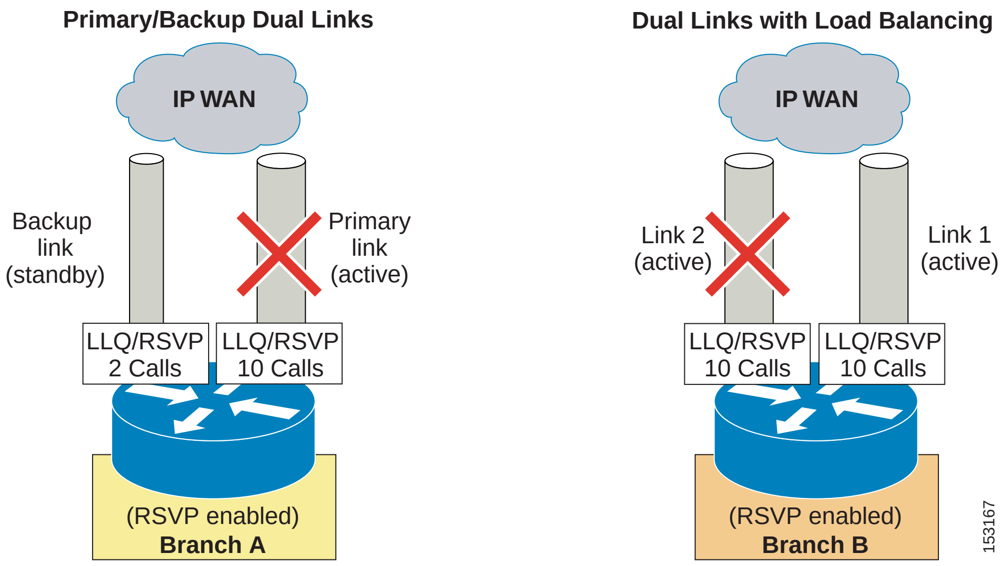
- The RSVP protocol attempts to establish an end-to-end reservation by checking for available bandwidth resources on all RSVP-enabled routers along the path from Device 3 to Device 4. As the RSVP messages progress through the network, the available RSVP bandwidth gets decremented by 24 kbps on the outbound router interfaces, as shown in Figure 11-8.
- In this example, the available bandwidth on R5's outbound interface toward R6 is not sufficient to accept the new data stream, so the reservation fails and the application is notified. The available RSVP bandwidth on each outbound interface along the path is then restored to its previous value.
- The application can then decide what to do. It could abandon the data transfer or decide to send it anyway with no QoS guarantees, as best-effort traffic.

We can now apply the topology-aware call admission control approach based on RSVP to the examples of dual-connected branch offices A and B introduced in the previous section.

As shown in Figure 11-9, branch office A has a primary link with an LLQ provisioned for 10 calls, while the backup link can accommodate only 2 calls. With this approach, RSVP is configured on both router interfaces so that the RSVP bandwidth matches the LLQ bandwidth. Branch A is also configured within the call processing agent to require RSVP reservations for all calls to or from other branches. Now calls are admitted or rejected based on the outcome of the RSVP reservations, which automatically follow the path determined by the routing protocol. Under normal conditions (when the primary link is active), up to 10 calls will be admitted; during failure of the primary link, only up to 2 calls will be admitted.

Policies can typically be set within the call processing agent to determine what to do in the case of a call admission control failure. For example, the call could be rejected, rerouted across the PSTN, or sent across the IP WAN as a best-effort call with a different DSCP marking.

Figure 11-9 Topology-Aware Call Admission Control for Dual Links



Similar considerations apply to branch B, connected to the IP WAN via two load-balanced links, as shown on the right side of Figure 11-9. RSVP is enabled on each of the two router interfaces, with a bandwidth value that matches the LLQ configuration (in this case, enough bandwidth for 10 calls). Branch B is also configured within the call processing agent to request RSVP reservations for calls to or from other branches. Again, calls are admitted or rejected based on the actual bandwidth available along the path chosen by the routing protocol. So in a case of perfectly even load-balancing across the two links, up to 20 calls could be admitted under normal conditions (when both links are operational); if one of the two links fails, only up to 10 calls would be admitted.

In the case that one of the two links failed while more than 10 calls were active, some calls would fail to re-establish a reservation on the new path. At this point, the call processing agent would be notified and could react based on the configured policy (for example, by dropping the extra calls or by remarking them as best-effort calls).

In conclusion, topology-aware call admission control allows administrators to protect call quality with any network topology, to automatically adjust to topology changes, and to make optimal use of the network resources under all circumstances.

Special Considerations for MPLS Networks

From the call admission control perspective, a network based on MPLS differs from one based on traditional Layer 2 WAN Services with respect to support for RSVP in the "hub" of the network. Hub sites of traditional Layer 2 wide-area networks consist, in most cases, of an enterprise-controlled router that can be enabled to participate in RSVP. Because the entire network (cloud) is the "hub site" in MPLS networks, there is no enterprise-controlled hub location to enable RSVP. (For more information, see MPLS Clouds, page 11-94.) Therefore, to provide topology-aware call admission control in an MPLS environment, the Customer Edge (CE) devices of the network must be configured for RSVP support.

Because RSVP must be enabled on the CE, control of this equipment is important. If this equipment is not under the control of the enterprise, you must work with your service provider to determine if they will enable RSVP on your WAN interface and if that implementation will support advanced features such as RSVP application ID.

RSVP messages will transparently pass across the RSVP-unaware MPLS cloud, so this does not pose a problem with end-to-end RSVP capability. Configuring RSVP on the CE WAN interface will ensure that its priority queue will not be overrun. Because RSVP reservations are unidirectional, the following rules must be observed to protect the priority queue on the Provider Edge (PE) router when RSVP is not enabled in the MPLS cloud:

- The media streams must be the same size in both directions.
- The media has to be symmetrically routed.

RSVP PATH messages record the egress IP address of the RSVP-aware routers they traverse. The information in the PATH message is used to send the RSVP RESV message back via the same route. Because of this mechanism, the WAN link between CE and PE must have routable IP addresses or the RSVP Reservations will fail.

If your MPLS network does not comply with these rules, contact your local Cisco account team for further assistance before implementing RSVP.

Call Admission Control Architecture

There are several mechanisms that perform the call admission control function in a Cisco IP Communications system. This section provides design and configuration guidelines for all of these mechanisms, according to their category:

- Topology-unaware mechanisms
  - Unified CM Enhanced Locations Call Admission Control, page 11-12
  - Cisco IOS Gatekeeper Zones, page 11-40
- Topology-aware mechanisms
  - Unified CM RSVP-Enabled Locations, page 11-62
  - RSVP SIP Preconditions, page 11-73

Unified CM Enhanced Locations Call Admission Control

Cisco Unified CM 9.*x* provides Enhanced Locations call admission control (CAC) to support complex WAN topologies as well as distributed deployments of Unified CM for call admission control where multiple clusters manage devices in the same physical sites using the same WAN uplinks. The Enhanced Locations CAC feature also supports immersive video, allowing the administrator to control call admissions for immersive video calls such as TelePresence separately from other video calls.

To support more complex WAN topologies Unified CM has implemented a locations-based network modeling functionality. This provides Unified CM with the ability to support multi-hop WAN connections between calling and called parties. This network modeling functionality has also been incrementally enhanced to support multi-cluster distributed Unified CM deployments. This allows the administrator to effectively "share" locations between clusters by enabling the clusters to communicate with one another to reserve, release, and adjust allocated bandwidth for the same locations across clusters. In addition, an administrator has the ability to provision bandwidth separately for immersive video calls such as TelePresence by allocating a new field to the Locations configuration called **immersive video bandwidth**.

There are also a number of tools to administer and troubleshoot Enhanced Locations CAC. The CAC enhancements and design are discussed in detail in this chapter, but the troubleshooting and serviceability tools are discussed in separate product documentation.

Network Modeling with Locations, Links, and Weights

Enhanced Locations CAC is a model-based static CAC mechanism. Enhanced Locations CAC involves using the administration interface in Unified CM to configure Locations and Links to model the "Routed WAN Network" in an attempt to represent how the WAN network topology routes media between groups of endpoints for end-to-end audio, video, and immersive calls. Although Unified CM provides configuration and serviceability interfaces in order to model the network, it is still a "static" CAC mechanism that does not take into account network failures and network protocol rerouting such as RSVP CAC. Therefore, the model needs to be updated when the WAN network topology changes. Enhanced Locations CAC is also call oriented, and bandwidth deductions are per-call not per-stream, so asymmetric media flows where the bit-rate is higher in one direction than in the other will always deduct for the highest bit rate. In addition, unidirectional media flows will be deducted as if they were bidirectional media flows.

The administrator builds the network model using locations and links. Enhanced Locations CAC incorporates the following configuration components:

- Locations — A Location represents a LAN. It could contain endpoints or simply serve as a transit location between links for WAN network modeling.
- Links — Links interconnect locations and are used to define bandwidth available between locations. Links logically represent the WAN link and are configured in the Location user interface (UI).
- Weights — A weight provides the relative priority of a link in forming the effective path between any pair of locations. The effective path is the path used by Unified CM for the bandwidth calculations, and it has the least cumulative weight of all possible paths. Weights are used on links to provide a "cost" for the "effective path" and are pertinent only when there is more than one path between any two locations.
- Path — A path is a sequence of links and intermediate locations connecting a pair of locations. Unified CM calculates shortest paths (least cost) from each location to all other locations and builds the paths. Only one "effective path" is used between a pair of locations.
- Effective Path — The effective path is the path with the least cumulative weight.
- Bandwidth Allocation — The amount of bandwidth allocated in the model for each type of traffic: audio, video, and immersive video (TelePresence).
- Locations Bandwidth Manager (LBM) — The active service in Unified CM that assembles a network model from configured location and link data in one or more clusters, determines the effective paths between pairs of locations, determines whether to admit calls between a pair of locations based on the availability of bandwidth for each type of call, and deducts (reserves) bandwidth for the duration of each call that is admitted.
- Locations Bandwidth Manager Hub — A Locations Bandwidth Manager (LBM) service that has been designated to participate directly in intercluster replication of fixed locations, links data, and dynamic bandwidth allocation data. LBMs assigned to an LBM hub group discover each other through their common connections and form a fully-meshed intercluster replication network. Other LBM services in a cluster with an LBM hub participate indirectly in intercluster replication through the LBM hubs in their cluster.

Locations and Links

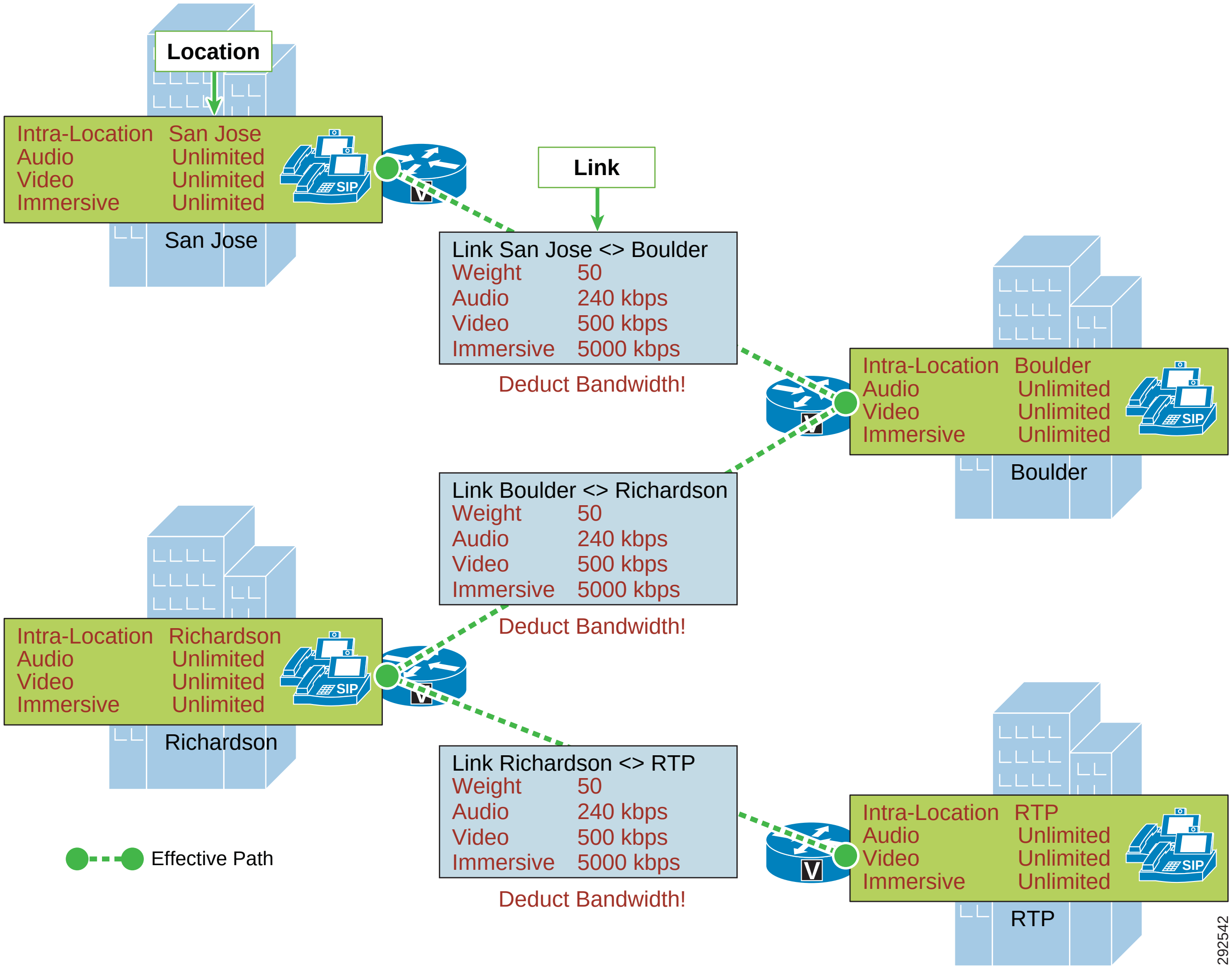
Unified CM uses the concept of locations to represent a physical site and to create an association with media devices such as endpoints, voice messaging ports, trunks, gateways, and so forth, through direct configuration on the device itself, through a device pool, or even through device mobility. Unified CM 9.*x* also uses a new locations configuration parameter called *links*. Links interconnect locations and are used to define bandwidth available between locations. Links logically represent the WAN links. This section describes locations and links and how they are used.

The location configuration itself consists of three main parts: links, intra-location bandwidth parameters, and RSVP locations settings. The RSVP locations settings are not considered here for Enhanced Locations CAC because they apply only to RSVP implementations. In the configuration, the link bandwidth parameters are displayed first while the intra-location bandwidth parameters are hidden and displayed by selecting the **Show advanced** link.

The intra-location bandwidth parameters allow the administrator to configure bandwidth allocations for three call types: audio, video, and immersive. They limit the amount of traffic within, as well as to or from, any given location. When any device makes or receives a call, bandwidth is deducted from the applicable bandwidth allocation for that call type. This feature allows administrators to effectively limit the amount of bandwidth used on the LAN or transit location. In most networks today that consist of at least 100BASE-T or Gigabit LANs, there is little or no reason to limit bandwidth on those LANs. However, there are some deployments that can benefit from limiting high-bandwidth video calls. A simple example might be an enterprise site with video deployed pervasively on the desktop and/or endpoints. If user calls are mostly all video-enabled, it is easy to see how a large number of 1 to 2 Mbps video calls might utilize a large percentage of available bandwidth on a LAN, and an administrator might consider limiting the number of video calls to a smaller percentage of that available LAN bandwidth. Keep in mind that this utilization might occur only during the busy hour of business or during a specific time of the year when specific traffic levels spike, and the bandwidth limit would be reached only during that time when it would be needed to ensure that the LAN is not over-subscribed with video traffic. It is also noteworthy to mention that video devices can be enabled to **Retry Video Call as Audio** if a video call to that device fails for any reason. This is configured on the video endpoint configuration page in Unified CM and is applicable to video endpoints or trunks receiving calls. It should also be noted that for some video endpoints **Retry Video Call as Audio** is enabled by default and not configurable on the endpoint.

The link bandwidth parameters allow the administrator to characterize the provisioned bandwidth for audio, video, and immersive calls between "adjacent locations" (that is, locations that have a link configured between them). This feature offers the administrator the ability to create a string of location pairings in order to model a multi-hop WAN network. To illustrate this, consider a simple three-hop WAN topology connecting four physical sites, as shown in Figure 11-10. In this topology we want to create links between San Jose and Boulder, between Boulder and Richardson, and between Richardson and RTP. Note that when we create a link from San Jose to Boulder, for example, the inverse link (Boulder to San Jose) also exists. Therefore, the administrator needs to create the link pairing only once from either location configuration page. In the example in Figure 11-10, each of the three links has the same settings: a weight of 50, 240 kbps of audio bandwidth, 500 kbps of video bandwidth, and 5000 kbps (or 5 MB) of immersive bandwidth.

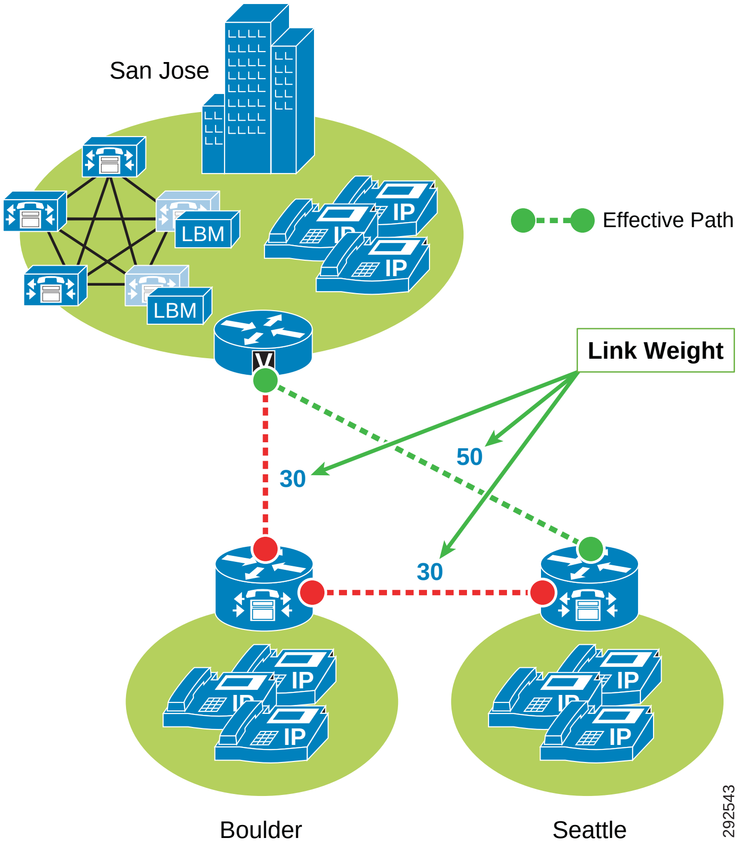
Figure 11-10 Simple Link Example with Three WAN Hops



When a call is made between San Jose and RTP, Unified CM calculates the bandwidth of the requested call, which is determined by the region pair between the two devices (see Locations, Links, and Region Settings, page 11-18) and verifies the effective path between the two locations. That is to say, Unified CM verifies the locations and links that make up the path between the two locations and accordingly deducts bandwidth from each link and (if applicable) from each location in the path. The intra-location bandwidth also is deducted along the path if any of the locations has configured a bandwidth value other than unlimited.

Weight is configurable on the link only and provides the ability to force a specific path choice when multiple paths between two locations are available. When multiple paths are configured, only one will be selected based on the cumulative weight, and this path is referred to as the *effective path*. This weight is static and the effective path does not change dynamically. Figure 11-11 illustrates weight configured on links between three locations: San Jose, Boulder, and Seattle.

Figure 11-11 Cumulative Path Weights



San Jose to Seattle has two paths, one direct link between the locations and another path through the Boulder location (link San Jose/Boulder and link Boulder/Seattle). The weight configured on the direct link between San Jose and Seattle is 50 and is less than the cumulative weight of links San Jose/Boulder and Boulder/Seattle which is 60 (30+30). Thus, the direct link is chosen as the effective path because the cumulative link weight is 50.

When you configure a device in Unified CM, the device can be assigned to a location. A location can be configured with links to other locations in order to build a topology. The locations configured in Unified CM are virtual locations and not real, physical locations. As mentioned, Unified CM has no knowledge of the actual physical topology of the network. Therefore, any changes to the physical network must be made manually in Unified CM to map the real underlying network topology with the Unified CM locations model. If a device is moved from one physical location to another, the system administrator must either perform a manual update on its location configuration or else implement the device mobility feature so that Unified CM can correctly calculate bandwidth allocations for calls to and from that device. Each device is in location **Hub_None** by default. Location Hub_None is an example location that typically serves as a hub linking two or more locations, and it is configured by default with unlimited intra-location bandwidth allocations for audio, video, and immersive bandwidth.

Unified CM allows you to define separate voice, video, and immersive video bandwidth pools for each location and link between locations. Typically the locations intra-location bandwidth configuration is left as a default of **Unlimited** while the link between locations is set to a finite number of kilobits per second (kbps) to match the capacity of a WAN links between physical sites. If the location's intra-location audio, video, and immersive bandwidths are configured as **Unlimited**, there will be unlimited bandwidth available for all calls (audio, video, and immersive) within that location and transiting that location. On the other hand, if the bandwidth values are set to a finite number of kilobits per second (kbps), Unified CM will track all calls within the location and all calls that use the location as a transit location (a location that is in the calculation path but is not the originating or terminating location in the path).

For video calls, the video location bandwidth takes into account both the audio and the video portions of the video call. Therefore, for a video call, no bandwidth is deducted from the audio bandwidth pool. The same applies to immersive video calls.

The devices that can specify membership in a location include:

- IP phones
- CTI ports
- H.323 clients
- CTI route points
- Conference bridges
- Music on hold (MoH) servers
- Gateways
- Trunks

The Enhanced Locations call admission control mechanism also takes into account the mid-call changes in call type. For example, if an inter-site video call is established, Unified CM will subtract the appropriate amount of video bandwidth from the respective locations and links in the path. If this video call changes to an audio-only call as the result of a transfer to a device that is not capable of video, Unified CM will return the allocated bandwidth to the video pool and allocate the appropriate amount of bandwidth from the audio pool along the same path. Calls that change from audio to video will cause the opposite change of bandwidth allocation.

Table 11-2 lists the amount of bandwidth requested by the static locations algorithm for various call speeds. For an audio call, Unified CM counts the media bit rates plus the Layer 3 overhead. For example, a G.711 audio call consumes 80 kbps (64k bit rate + L3 overhead) deducted from the location's and link's audio bandwidth allocation. For a video call, Unified CM counts only the media bit rates for both the audio and video streams. For example, for a video call at a bit rate of 384 kbps, Unified CM will allocate 384 kbps from the video bandwidth allocation.

Table 11-2 Amount of Bandwidth Requested by the Locations and Links Bandwidth Deduction Algorithm

Call Speed	Static Location and Link Bandwidth Value
G.711 audio call (64 kbps)	80 kbps
G.729 audio call (8 kbps)	24 kbps
128 kbps video call	128 kbps
384 kbps video call	384 kbps
512 kbps video call	512 kbps
768 kbps video call	768 kbps

For a complete list of codecs and location and link bandwidth values, refer to the bandwidth calculations information in the *Call Admission Control* section of the *Cisco Unified Communications Manager System Guide*, available at

http://www.cisco.com/en/US/products/sw/voicesw/ps556/prod_maintenance_guides_list.html

For example, assume that the link configuration for the location Branch 1 to Hub_None allocates 256 kbps of available audio bandwidth and 384 kbps of available video bandwidth. In this case the path from Branch 1 to Hub_None can support up to three G.711 audio calls (at 80 kbps per call) or ten G.729 audio calls (at 24 kbps per call), or any combination of both that does not exceed 256 kbps. The link between locations can also support different numbers of video calls depending on the video and audio codecs being used (for example, one video call requesting 384 kbps of bandwidth or three video calls with each requesting 128 kbps of bandwidth).

When a call is placed from one location to the other, Unified CM deducts the appropriate amount of bandwidth from the effective path of locations and links from one location to another. Using Figure 11-10 as an example, a G.729 call between San Jose and RTP locations causes Unified CM to deduct 24 kbps from the available bandwidth at the links between San Jose and Boulder, between Boulder and Richardson, and between Richardson and RTP. When the call has completed, Unified CM returns the bandwidth to those same links over the effective path. If there is not enough bandwidth at any one of the links over the path, the call is denied by Unified CM and the caller receives the network busy tone. If the calling device is an IP phone with a display, that device also displays the message "Not Enough Bandwidth."

When an inter-location call is denied by call admission control, Unified CM can automatically reroute the call to the destination through the PSTN connection by means of the Automated Alternate Routing (AAR) feature. For detailed information on the AAR feature, see Automated Alternate Routing, page 9-117.

Note AAR is invoked only when Enhanced Locations call admission control denies the call due to a lack of network bandwidth along the effective path. AAR is not invoked when the IP WAN is unavailable or other connectivity issues cause the called device to become unregistered with Unified CM. In such cases, the calls are redirected to the target specified in the Call Forward No Answer field of the called device.

Locations, Links, and Region Settings

Locations work in conjunction with regions to define the characteristics of a call over the effective path of locations and links. Regions define the type of compression or bit rate (8 kbps or G.729, 64 kbps or G.722/G.711, and so forth) that is used between devices, and location links define the amount of available bandwidth for the effective path between devices. You assign each device in the system to both a region (by means of a device pool) and a location (by means of a device pool or by direct configuration on the device itself).

You can configure locations in Unified CM to define:

- Physical sites (for example, a branch office) or transit sites (for example, an MPLS cloud) — A location represents a LAN. It could contain endpoints or simply serve as a transit location between links for WAN network modeling.
- Link bandwidth between adjacent locations — Links interconnect locations and are used to define bandwidth available between locations. Links logically represent the WAN link between physical sites.
  - Audio Bandwidth — The amount of bandwidth that is available in the WAN link for voice and fax calls being made from devices in the location to the configured adjacent location. This bandwidth value is used by Unified CM for Enhanced Locations call admission control.
  - Video Bandwidth — The amount of video bandwidth that is available in the WAN link for video calls being made from devices in the location to the configured adjacent location. This bandwidth value is used by Unified CM for Enhanced Locations call admission control.

  - Immersive Video Bandwidth — The amount of immersive bandwidth that is available in the WAN link for TelePresence calls being made from devices in the location to the configured adjacent location. This bandwidth value is used by Unified CM for Enhanced Locations call admission control.

- Intra-location bandwidth
  - Audio Bandwidth — The amount of bandwidth that is available in the WAN link for voice and fax calls being made from devices in the location to the configured adjacent location. This bandwidth value is used by Unified CM for Enhanced Locations call admission control.
  - Video Bandwidth — The amount of video bandwidth that is available in the WAN link for video calls being made from devices in the location to the configured adjacent location. This bandwidth value is used by Unified CM for Enhanced Locations call admission control.
  - Immersive Video Bandwidth — The amount of immersive bandwidth that is available in the WAN link for TelePresence calls being made from devices in the location to the configured adjacent location. This bandwidth value is used by Unified CM for Enhanced Locations call admission control.
- The settings for RSVP call admission control between locations — Possible settings are No Reservation, Optional, Optional (Video Desired), Mandatory, and Mandatory (Video Desired).

You can configure regions in Unified CM to define:

- The Max Audio Bit Rate used for intraregion calls
- The Max Audio Bit Rate used for interregion calls
- The Max Video Call Bit Rate (Includes Audio) used for intraregion and interregion calls. This also includes the maximum bit rate for immersive calls when applied to TelePresence endpoints.
- The link loss type for interregion calls (Possible link loss types are Low Loss and Lossy)

Unified CM Support for Locations and Regions

Cisco Unified Communications Manager supports 2,000 locations and 2,000 regions with Cisco MCS-7845 servers. To deploy up to 2,000 locations and regions, you must configure the following service parameters in the **Clusterwide Parameters** > **(System - Location and Region) and Clusterwide Parameters** > **(System - RSVP)** configuration menus:

- Default Intraregion Max Audio Bit Rate
- Default Interregion Max Audio Bit Rate
- Default Intraregion Max Video Call Bit Rate (Includes Audio)
- Default Interregion Max Video Call Bit Rate (Includes Audio)
- Default Intraregion and Interregion Link Loss Type

When adding regions, you should select **Use System Default** for the Max Audio Bit Rate and Max Video Call Bit Rate values. If you are using RSVP call admission control, you should also select **Use System Default** for the RSVP parameter.

Changing these values for individual regions and locations from the default has an impact on server initialization and publisher upgrade times. Hence, with a total of 2,000 regions and 2,000 locations, you can modify up to 200 of them to use non-default values. With a total of 1,000 or fewer regions and locations, you can modify up to 500 of them to use non-default values. Table 11-3 summarizes these limits.

Table 11-3 Number of Allowed Non-Default Regions and Locations

Number of non-default regions and locations	Maximum number of regions	Maximum number of locations
0 to 200	2,000	2,000
200 to 500	1,000	1,000

Note The Max Audio Bit Rate is used by both voice calls and fax calls. If you plan to use G.729 as the interregion codec, use T.38 Fax Relay for fax calls. If you plan to use fax pass-through over the WAN, change the default Interregion Max Audio Bit Rate to 64 kbps (G.722 or G.711), or else add a region for fax machines to each location with a non-default bit rate of 64 kbps (G.722 or G.711), subject to the limits in Table 11-3.

Note Irrespective of the MCS model you are using, your Cisco Partner or Cisco Systems Engineer should always use the Cisco Unified Communications Sizing Tool (http://tools.cisco.com/cucst) to validate all designs that incorporate a large number of remote sites, because there are many interdependent variables that can affect Unified CM cluster scalability (such as regions, locations, gateways, media resources, and so forth). Use the Sizing Tool to accurately determine the number of servers or clusters required to meet your design criteria.

Locations Bandwidth Manager

The Locations Bandwidth Manager (LBM) is a Unified CM Feature Service managed from the serviceability web pages and responsible for all of the Enhanced Locations CAC bandwidth functions. The LBM can run on any Unified CM subscriber or as a standalone service on a dedicated Unified CM server in the cluster. A minimum of one instance of LBM must run in each cluster to enable Enhanced Locations CAC in the cluster. For most installations, Cisco recommends the LBM are:

- Locations and links path assembly
- Bandwidth calculations over the effective paths in the assembly
- Servicing bandwidth requests from the Cisco CallManager service (Unified CM call control)
- Replication of bandwidth information to other LBMs within the cluster and between clusters when intercluster Enhanced Locations CAC is enabled
- Providing configured and dynamic information to serviceability
- Updating Location Real-Time Monitoring Tool (RTMT) counters
- Using Extensible Markup Language (XML) over TCP for communication to/from the Cisco CallManager service as well as between LBMs.

The LBM Service is enabled by default when upgrading to Cisco Unified CM 9.x from earlier releases. For new installations, the LBM service must be activated manually.

During initialization, the LBM reads local locations information from the database, such as: locations audio, video, and immersive bandwidth values; intra-location bandwidth data; and location-to-location link audio, video, and immersive bandwidth values and weight (inter-location bandwidth data). Using the link data, each LBM in a cluster creates a local assembly of the paths from one location to every other location. This is referred to as the *assembled topology*. In a cluster, each LBM accesses the same data and thus creates the same local copy of the assembled topology during initialization.

At runtime, the LBM applies reservations along the computed paths in the local assembled topology of locations and links, and it replicates the reservations to other LBMs in the cluster. If intercluster Enhanced Locations CAC is configured and activated, the LBM replicates the assembled topology to other clusters (see Intercluster Enhanced Locations CAC, page 11-22, for more details).

By default the Cisco CallManager service communicates with the local LBM service; however, LBM groups can be used to manage this communication. LBM groups provide an active and standby LBM in order to create redundancy for Unified CM call control. Figure 11-12 illustrates LBM redundancy.

Figure 11-12 Locations Bandwidth Manager Redundancy

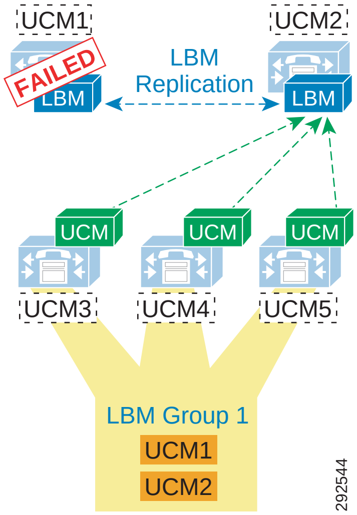


Figure 11-12 shows five Unified CM servers: UCM1 and UCM2 are dedicated LBM servers (only LBM service enabled); UCM3, UCM4, and UCM5 are Unified CM subscribers (Cisco CallManager service enabled). An LBM Group has been configured with UCM1 as active and UCM2 as standby, and it is applied to subscribers UCM3, UCM4, and UCM5. This configuration allows for UCM3, UCM4, and UCM5 to query UCM1 for all bandwidth requests. If UCM1 fails for any reason, the subscribers will fail-over to the standby UCM2.

The order in which the Unified CM Cisco CallManager service uses the LBM is as follows:

- LBM Group designation
- Local LBM
- Service parameter **Call Treatment when no LBM available** (Default = **allow calls**)

Enhanced Locations CAC Design and Deployment Recommendations and Considerations

- The Locations Bandwidth Manager (LBM) is a Unified CM Feature Service.
- LBM is responsible for modeling the topology and servicing Unified CM bandwidth requests.
- All LBMs are fully meshed within the cluster.

- The Enhanced Locations CAC LBM replication network is used to replicate the modeled topology as well as the bandwidth allocations within the cluster and across multiple clusters.
- Recommendations for LBM Group usage are as follows:
  - Manage how the Cisco CallManager service interacts with LBM (co-resident or dedicated).
  - Minimize LBM full-mesh bandwidth requirements in clustering over the WAN or dual data center deployments.
  - Deploy a minimum of two LBMs per call processing site for redundancy, either co-resident or dedicated.
  - Off-load active LBMs to inactive stand-by subscribers.
- Current recommendation is to deploy the LBM service co-resident with a Unified CM subscriber running the Cisco CallManager call processing service.

LBM Group Recommendations

- Configure each Unified CM subscriber to have a local LBM running and active.
- A minimum of two LBMs in a redundant LBM group configuration should be active at each call processing site, such as in clustering over the WAN designs.
- For load reduction of active subscribers, use dedicated LBMs or enable LBM on the inactive stand-by subscribers in 1:1 Unified CM redundancy models.

Intercluster Enhanced Locations CAC

Intercluster Enhanced Locations CAC extends the concept of network modeling across multiple clusters. In intercluster Enhanced Locations CAC, each cluster manages its locally configured topology of locations and links and then propagates this local topology to other remote clusters that are part of the LBM intercluster replication network. Upon receiving a remote cluster's topology, the LBM assembles this into its own local topology and creates a global topology. Through this process the global topology is then identical across all clusters, providing each cluster a global view of enterprise network topology for end-to-end CAC. Figure 11-13 illustrates the concept of a global topology with a simplistic hub-and-spoke network topology as an example.

Figure 11-13 Example of a Global Topology for a Simple Hub-and-Spoke Network

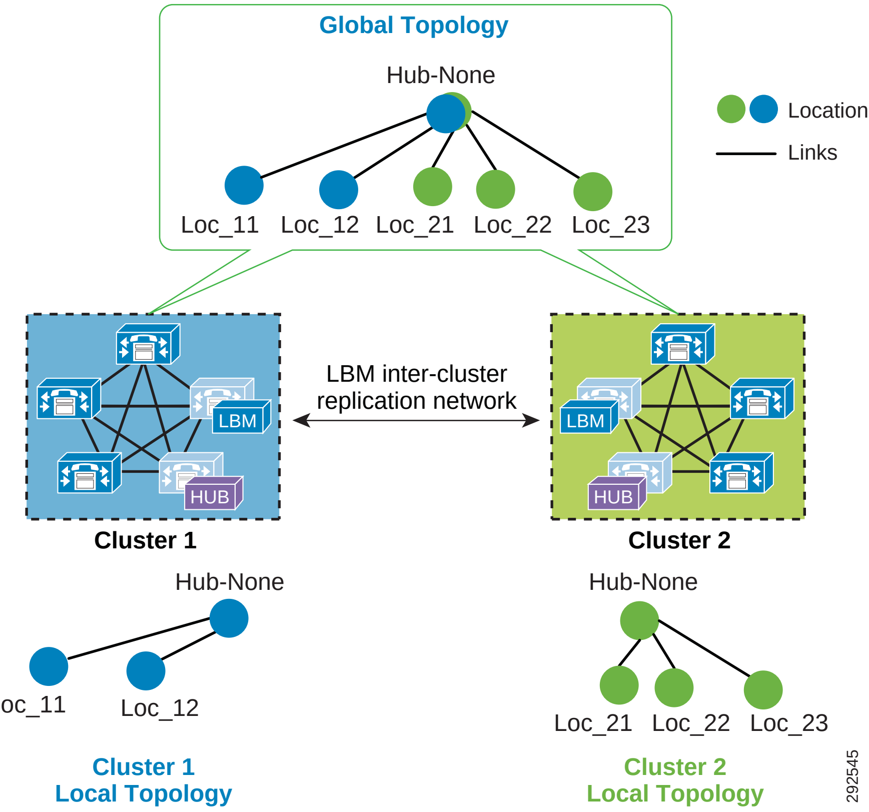
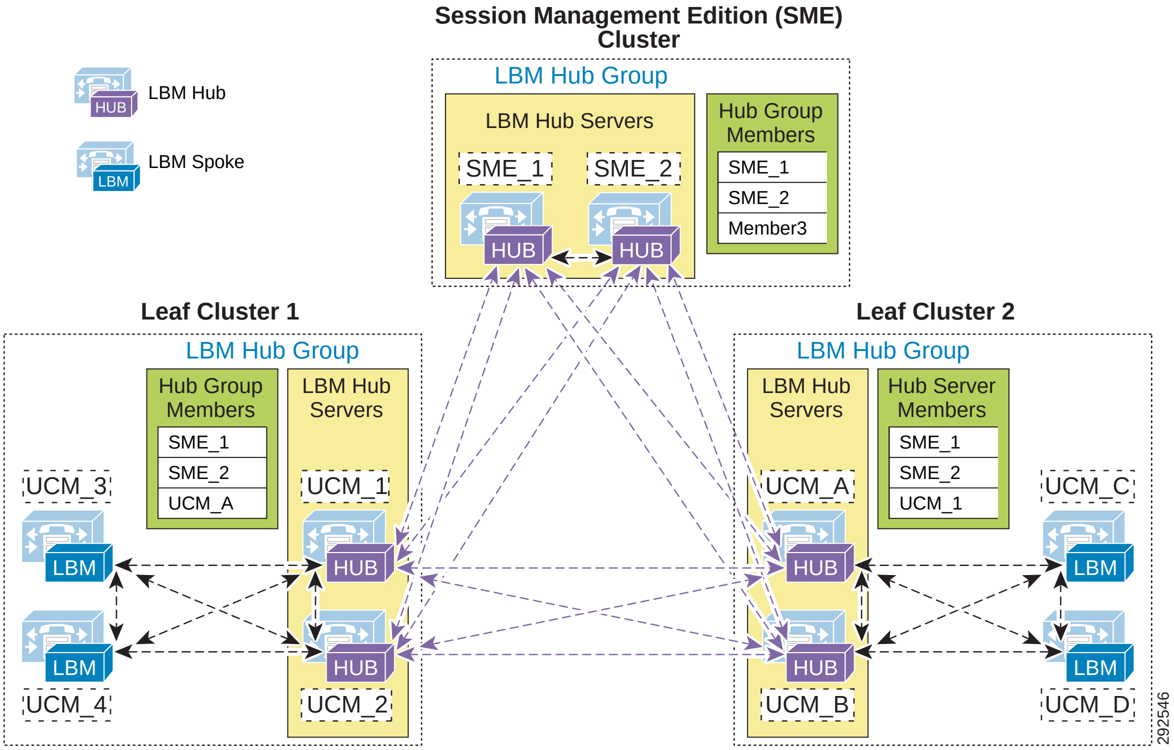


Figure 11-13 shows two clusters, Cluster 1 and Cluster 2, each with a locally configured hub-and-spoke network topology. Cluster 1 has configured Hub_None with links to Loc_11 and Loc_12, while Cluster 2 has configured Hub_None with links to Loc_21, Loc_22, and Loc_23. Upon enabling intercluster Enhanced Locations CAC, Cluster 1 sends its local topology to Cluster 2, as does Cluster 2 to Cluster 1. After each cluster obtains a copy of the remote cluster's topology, each cluster overlays the remote cluster's topology over their own. The overlay is accomplished through common locations, which are locations that are configured with the same name. Because both Cluster 1 and Cluster 2 have the common location Hub_None with the same name, each cluster will overlay the other's network topology with Hub_None as a common location, thus creating a global topology where Hub_None is the hub and Loc_11, Loc_12, Loc_21, Loc_22 and Loc_23 are all spoke locations. This is an example of a simple network topology, but more complex topologies would be processed in the same way.

LBM Hub Replication Network

The intercluster LBM replication network is a network of designated LBMs that create a full-mesh with one another and replicate their local cluster's topology. In turn, each receives all remote clusters' topologies in order to create the global topology. The designated LBMs for the intercluster replication network are called LBM hubs, and the LBMs that replicate only within a cluster are called LBM spokes. The LBM hubs are designated in configuration through the LBM hub group. The LBM hub group has two main configuration areas called hub group members and hub group usage information. The hub group members are LBM hubs in remote clusters that are part of the LBM replication network. A maximum of three members can be configured. The members designated in the LBM hub group members serve as bootstrap servers for the entire intercluster replication network, providing each LBM hub in each cluster with the connectivity details of other remote clusters with whom they are connected. The LBM Hub group usage information consists of the LBM hubs and spokes in the local cluster. Moving an LBM service into or out of the LBM Hub group determines the hub or spoke role. (See Cisco Unified Communications Manager product documentation for further information on the LBM hub group configuration.) Once the LBM hub group is configured on each cluster in the designated LBM, hubs will create the full mesh intercluster replication network. Figure 11-14 illustrates an intercluster replication network configuration with LBM hub groups set up between three clusters (Leaf Cluster 1, Leaf Cluster 2 and a Session Management Edition (SME) Cluster) to form the intercluster replication network.

Figure 11-14 Example Intercluster Replication Network for Three Cluster



In Figure 11-14, two LBM servers from each cluster have been designated as the LBM hubs for their cluster. These LBM hub servers form the intercluster LBM replication network. The LBM hub group members configured in each LBM hub group are designated as SME_1 and SME_2. These two LBM servers from the SME cluster serve as points of contact for the entire intercluster LBM replication network. This means that each LBM in each cluster connects to SME_1, replicates its local topology to SME_1, and gets the remote topology from SME_1. They also get the connectivity information for the other leaf clusters from SME_1, connect to the other remote clusters, and replicate their topologies. This creates the full-mesh replication network. If SME_1 is unavailable, the LBM hubs will connect to SME_2. If SME_2 is unavailable, Leaf Cluster 1 LBMs will connect to UCM_A and Leaf Cluster 2 LBMs will connect to UCM_1 as a backup measure in case the SME Cluster is unavailable. This is just an example configuration to illustrate the components of the intercluster LBM replication network.

The LBM has the following roles with respect to the LBM intercluster replication network:

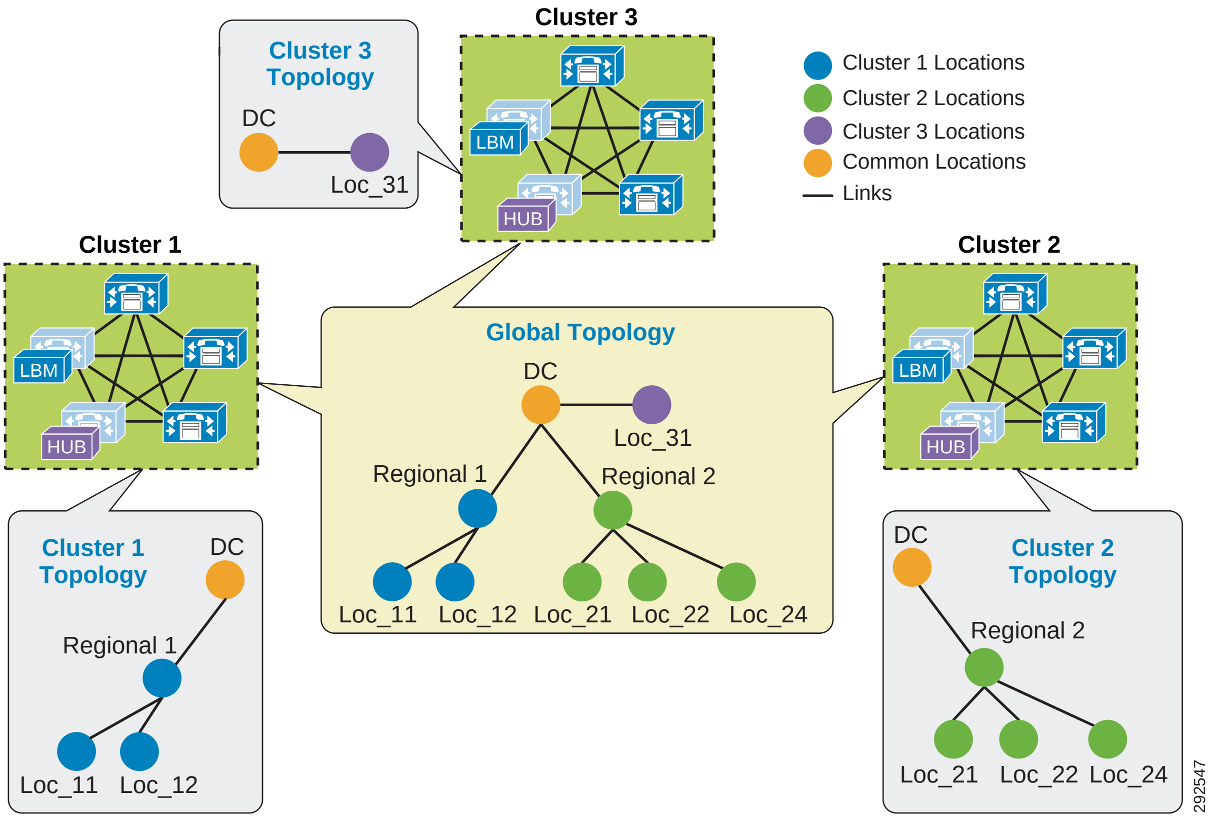
- LBM Hub group members
  - Remote hub servers responsible for interconnecting all LBM hubs in the replication network
  - Can be any hub in the network
  - Can indicate up to 3 per hub group
- LBM Hub Servers (Local LBMs)
  - Communicate directly to other remote hub servers as part of the intercluster LBM replication network
- LBM Spoke Servers (Local LBMs)
  - Communicate directly to local LBM hubs in the cluster and indirectly to the remote LBM hubs through the local LBM hubs
- LBM Hub Replication Network — Bandwidth deduction and adjustment messages
  - If a cluster has multiple LBM hubs, the LBM hub with the lowest IPv4 (entire) address will function as the sender of messages to other remote clusters. Only one hub per cluster will forward messages to remote clusters. This limits the amount of replication traffic in the intercluster replication network.
  - The LBM hub that functions as the sender for messages in the cluster selects one LBM hub from each cluster and forwards messages to that LBM.
  - The LBM hubs that receive messages from remote clusters, in turn forward the received messages to the LBM spokes in their local cluster.
  - Forwarded messages have a unique random string associated with them that allows receivers to determine if a messages has already been received and thus drop messages that they have received twice to prevent any replication storm or looping.
  - Other LBM hubs in the cluster that receive the forwarded message will not forward on to LBM spokes because the message is not directly from a remote cluster. This avoids hubs sending duplicate messages from remote clusters.

Common Locations (Shared Locations) and Links

As mentioned previously, common locations are locations that are named the same across clusters. Common locations play a key role in how the LBM creates the global topology and how it associates a single location across multiple clusters. A location with the same name between two or more clusters is considered the same location and is thus a shared location across those clusters. So if a location is meant to be shared between multiple clusters, it is required to have exactly the same name. After replication, the LBM will check for configuration discrepancies across locations and links. Any discrepancy in bandwidth value or weight between common locations and links can be seen in serviceability, and the LBM calculates the locations and link paths with the most restrictive values for bandwidth and the lowest value (least cost) for weight.

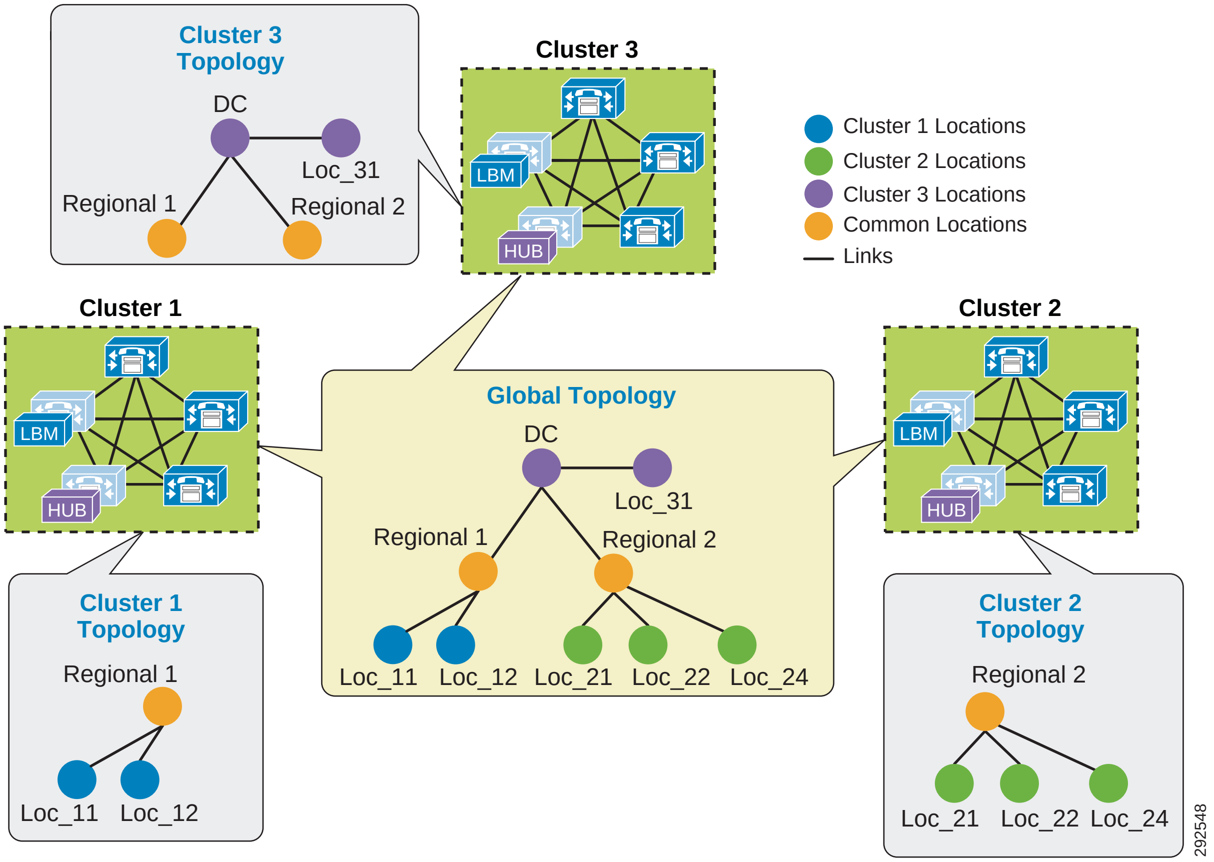
Common locations and links can be configured across clusters for a number of different reasons. You might have a number of clusters that manage devices in the same physical site and use the same WAN uplinks, and therefore the same location needs to be configured on each cluster in order to associate that location to the local devices on each cluster. You might also have clusters that manage their own topology, yet these topologies interconnect at specific locations and you will have to configure these locations as common locations across each cluster so that, when the global topology is being created, the clusters have the common interconnecting locations and links on each cluster to link each remote topology together effectively. Figure 11-15 illustrates linking topologies together and shows the common topology that each cluster shares.

Figure 11-15 Using Common Locations and Links to Create a Global Topology



In Figure 11-15, Cluster 1 has devices in locations Regional 1, Loc_11, and Loc_12, but it requires configuring DC and a link from Regional 1 to DC in order to link to the rest of the global topology. Cluster 2 is similar, with devices in Regional 2 and Loc_21, Loc_22, and Loc_23, and it requires configuring DC and a link from DC to Regional 2 to map into the global topology. Cluster 3 has devices in Loc_31 only, and it requires configuring DC and a link to DC from Loc_31 to map into Cluster 1 and Cluster 2 topologies. Alternatively, Regional 1 and Regional 2 could be the common locations configured on all clusters instead of DC, as is illustrated in Figure 11-16.

Figure 11-16 Alternative Topology Using Different Common Locations



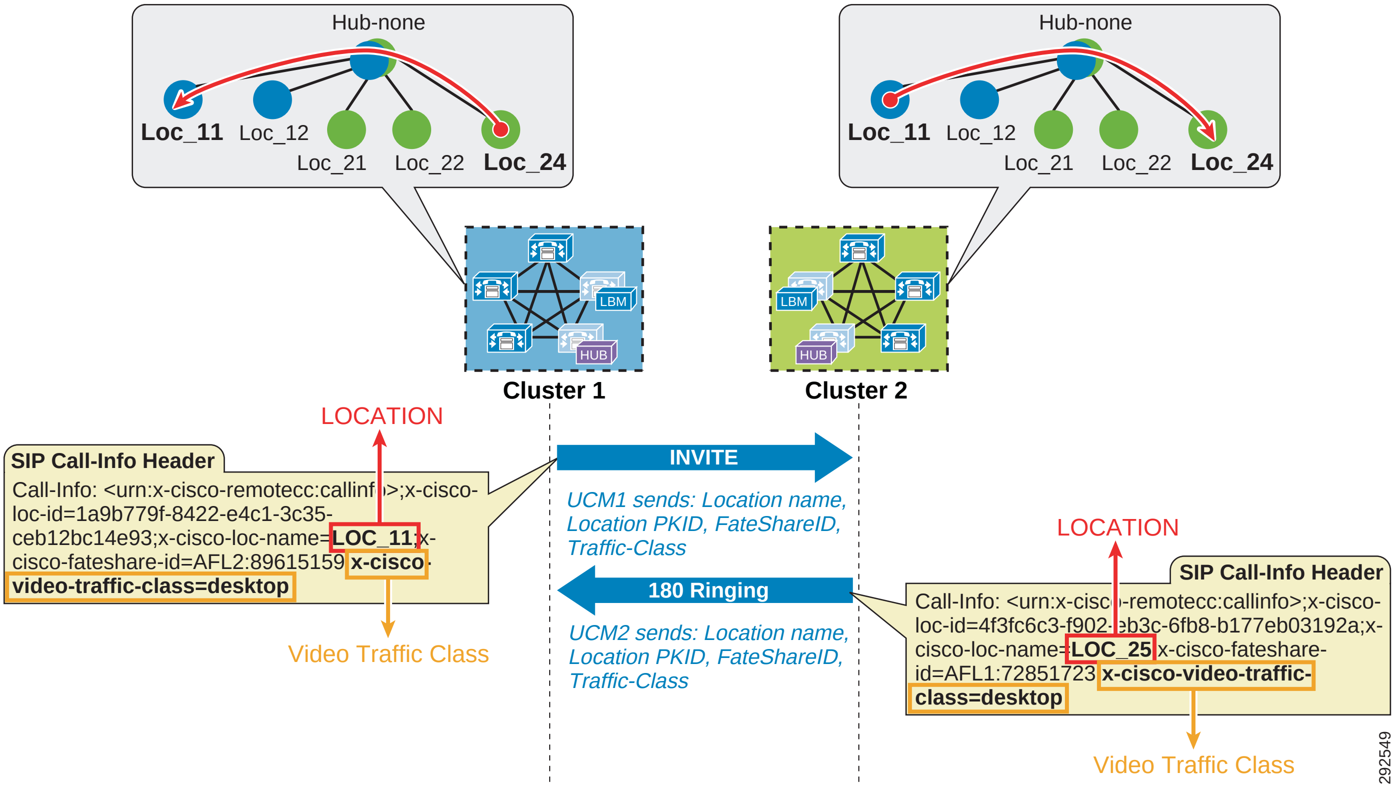
The key to topology mapping from cluster to cluster is to ensure that at least one cluster has a common location with another cluster so that the topologies interconnect accordingly.

Shadow Location

The *shadow location* is used to enable a SIP trunk to pass Enhanced Locations CAC information such as location name and Video-Traffic-Class (discussed below), among other things, required for Enhanced Locations CAC to function across clusters. In order to pass this location information across clusters, the SIP intercluster trunk (ICT) must be assigned to the "shadow" location. Similar to the "phantom" location, it cannot have a link to other locations, and therefore no bandwidth can be reserved between the shadow location and other locations. Any device other than a SIP ICT that is assigned to the shadow location will be treated as if it was associated to Hub_None. That is important to know because if a device other than a SIP ICT ends up in the shadow location, bandwidth deductions will be made from that device as if it were in Hub_None, and that could have varying effects depending on the location and links configuration.

When the SIP ICT is enabled for Enhanced Locations CAC, it passes information in the SIP Call-Info header that allows the originating and terminating clusters to process the location bandwidth deductions end-to-end. Figure 11-17 illustrates an example of a call between two clusters and some details about the information passed. This is only to illustrate how location information is passed from cluster to cluster and how bandwidth deductions are made.

Figure 11-17 Shadow Location Used to Pass Information Between Clusters


Hub-none
Loc_11
Loc_12
Loc_21
Loc_22
Loc_24
Hub-none
Loc_11
Loc_12
Loc_21
Loc_22
Loc_24
LBM
HUB
Cluster 1
LBM
HUB
Cluster 2
LOCATION
SIP Call-Info Header
Call-Info: <urn:x-cisco-remotecc:callinfo>;x-cisco-loc-id=1a9b779f-8422-e4c1-3c35-ceb12bc14e93;x-cisco-loc-name=LOC_11;x-cisco-fateshare-id=AFL2:89615159;x-cisco-video-traffic-class=desktop
Video Traffic Class
INVITE
UCM1 sends: Location name, Location PKID, FateShareID, Traffic-Class
180 Ringing
UCM2 sends: Location name, Location PKID, FateShareID, Traffic-Class
LOCATION
SIP Call-Info Header
Call-Info: <urn:x-cisco-remotecc:callinfo>;x-cisco-loc-id=4f3fc6c3-f902-eb3c-6fb8-b177eb03192a;x-cisco-loc-name=LOC_25;x-cisco-fateshare-id=AFL1:72851723;x-cisco-video-traffic-class=desktop
Video Traffic Class
292549

In Figure 11-17, Cluster 1 sends an invite to Cluster 2 and populates the call-info header with the calling parties location name and Video-Traffic-Class, among other pertinent information such as unique call-ID. When Cluster 2 receives the invite with the information, it looks up the terminating party and performs a CAC request on the path between the calling party's and called party's locations from the global topology that it has in memory from LBM replication. If it is successful, Cluster 2 will replicate the reservation and extend the call to the terminating device and return a 180 ringing with the location information of the called party back to Cluster 1. When Cluster 1 receives the 180 ringing, it gets the terminating device's location name and goes through the same bandwidth lookup process using the same unique call-ID that it calculates from the information passed in the call-info header. If it is successful, it too continues with the call flow. Because both clusters use the same information in the call-info header, they will deduct bandwidth for the same call using the same call-ID, thus avoiding any double bandwidth deductions.

Location and Link Management Cluster

In order to avoid configuration overhead, a Location and Link Management Cluster can be configured to manage all locations and links in the global topology. All other locations uniquely configure the locations that they require for location-to-device association and do not configure links or any bandwidth values other than unlimited. It should be noted that the Location and Link Management Cluster is a design concept and is simply any cluster that is configured with the entire global topology of locations and links, while all other clusters in the LBM replication network are configured only with locations with unlimited bandwidth values and no configured links. When intercluster Enhanced Locations CAC is enabled and the LBM replication network is configured, all clusters replicate their view of the network. The designated Location and Link Management Cluster has the entire global topology with locations, links, and bandwidth values; and once those values are replicated, all clusters use those values because they are the most restrictive. This design alleviates configuration overhead in deployments where a large number of common locations are required across multiple clusters.

Recommendations

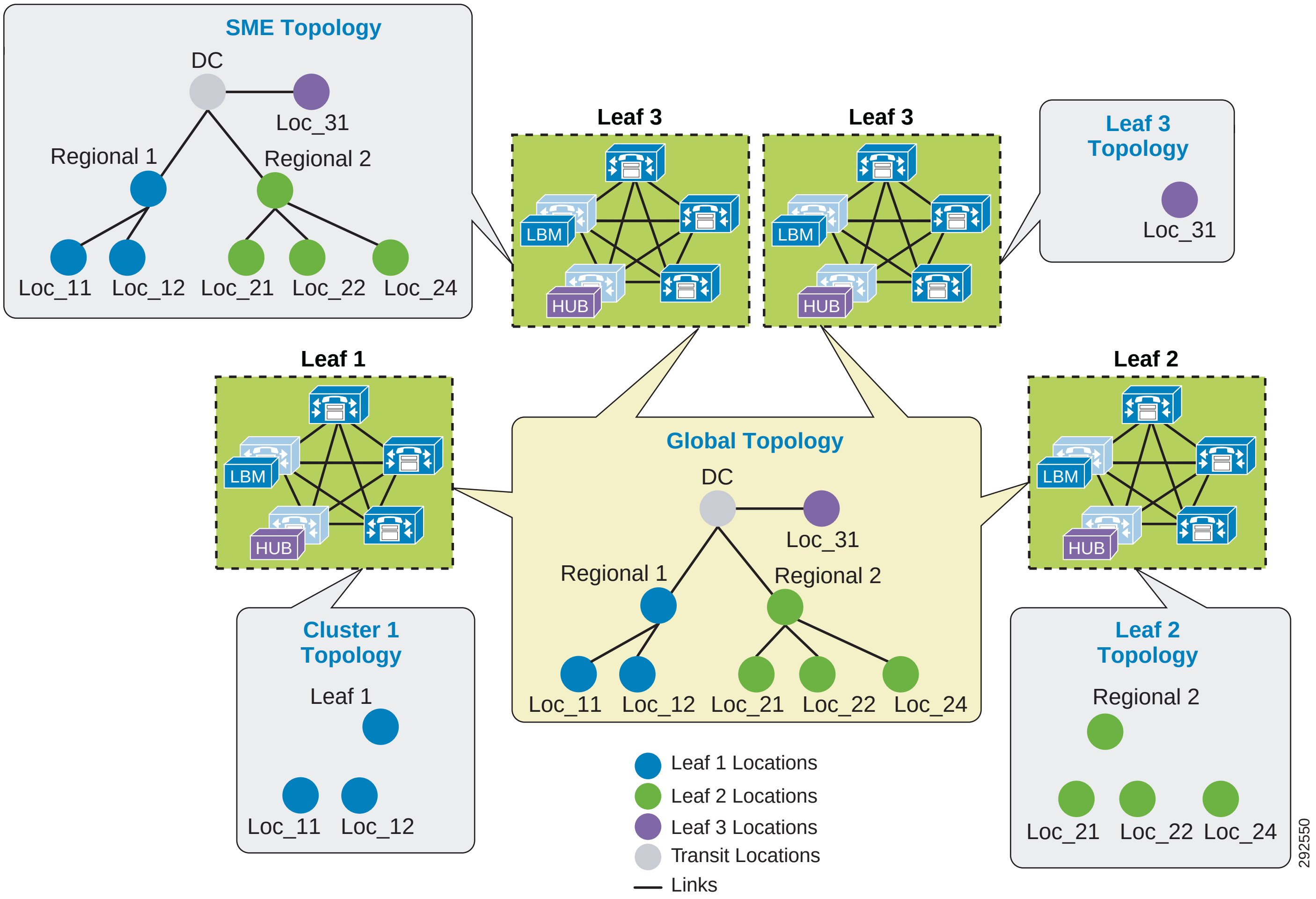
- Management cluster in the LBM replication network
  - All links and locations are managed in the management cluster.
  - Locations, bandwidth values, links, and weights
- Other clusters in the LBM replication network
  - Use unlimited bandwidth values on the intra-locations bandwidth parameters.
  - Do not configure links.
- LBM will always use the lowest most restrictive bandwidth and lowest weight value after replication.

Benefits

- Manage enterprise CAC topology from a single cluster.
- Alleviates location and link configuration overhead when clusters share a large number of common locations.
- Alleviates configuration mistakes in locations and links across clusters.
- Other clusters in the enterprise require the configuration only of locations needed for location-to-device and endpoint association.
- Provides a single cluster for monitoring of the global locations topology.

Figure 11-18 illustrates Cisco Unified Communications Manager Session Management Edition (SME) as a Location and Link Management Cluster for three leaf clusters.

Figure 11-18 Example of SME as a Location and Link Management Cluster



In Figure 11-18 there are three leaf clusters, each with devices in only a regional and remote locations. SME has the entire global topology configured with locations and links, and intercluster LBM replication is enabled between all four clusters. None of the clusters in this example share locations, although all of the locations are common locations because SME has configured the entire location and link topology. Note that Leaf 1, Leaf 2, and Leaf 3 configure only locations that they require to associate to devices and endpoints, while SME has the entire global topology configured. After intercluster replication, all clusters will have the global topology.

Intercluster Enhanced Locations CAC Design and Deployment Recommendations and Considerations

- Oversubscription in bandwidth reservations can be incurred since reservations are made locally and replicated out to the rest of the LBM replication network. To avoid QoS impacts during oversubscription, observe the following guidelines:
  - Oversubscription is transient and will correct itself as the calls, using the oversubscribed locations and links in the path, clear and relinquish the bandwidth.
  - Bandwidth overhead should be provisioned in the QoS network policy to accommodate oversubscription. Cisco recommends over-provisioning by a minimum of one call of the highest bandwidth value in each QoS class (audio and video) for applicable locations and links. For audio-only implementations this may be a single call at 24 kbps or 80 kbps. For video implementations this may be the bit rate of a single video call.
  - Locations and links where CAC limits are often reached during the busy hour or in general, are prime candidates for over-provisioning of the QoS bandwidth capacity.
  - In cases of very high busy hour call completions (BHCC) and long delays between Unified CM clusters, Cisco recommends monitoring the locations and links to determine the amount of oversubscription during the busy hour, then ensure that the network is over-provisioned with an equal amount of bandwidth.
- A cluster requires the location to be configured locally for location-to-device association.
- Each cluster should be configured with the immediately neighboring locations so that each cluster's topology can inter-connect. This does not apply to Location and Link Management Cluster deployments.
- Links need to be configured to establish points of interconnect between remote topologies. This does not apply to Location and Link Management Cluster deployments.
- Discrepancies of bandwidth limits and weights on common locations and links are resolved by using the lowest bandwidth and weight values.
- Naming locations consistently across clusters is critical. Follow the practice, "Same location, same name; different location, different name."
- The Hub_None location should be renamed to be unique in each cluster or else it will be a common (shared) location by other clusters.
- Cluster-ID should be unique on each cluster for serviceability reports to be usable.
- All LBM hubs are fully meshed between clusters.
- An LBM hub is responsible for communicating to hubs in remote clusters.
- An LBM spoke does not directly communicate with other remote clusters. LBM spokes receive and send messages to remote clusters through the Local LBM Hub.
- LBM Hub Groups
  - Used to assign LBMs to the Hub role
  - Used to define three remote hub members that replicate hub contact information for all of the hubs in the LBM hub replication network
  - An LBM is a hub when it is assigned to an LBM hub group.
  - An LBM is a spoke when it is not assigned to an LBM hub group.
- If a cluster has no LBM hub, or if the LBM hub is not running, the cluster will be isolated and will not participate in the intercluster LBM replication network.

Performance Guidelines

- Maximum of 2,000 locally configured locations. This limit of 2,000 locations also applies to the Location and Link Management Cluster.
- Maximum of 8,000 total replicated locations with intercluster CAC

Enhanced Locations CAC for TelePresence Immersive Video

Since TelePresence endpoints now provide a diverse range of collaborative experiences from the desktop to the conference room, Enhanced Locations CAC includes support to provide CAC for TelePresence immersive video calls. This section discusses the features in Enhanced Locations CAC that support TelePresence immersive video CAC.

Video Call Traffic Class

Video Call Traffic Class is a attribute that is assigned to all endpoints, and that can also be enabled on SIP trunks, to determine the video classification type of the endpoint or trunk. This enables Unified CM to classify various call flows as either immersive, desktop video, or both, and to deduct accordingly from the appropriate location and/or link bandwidth allocations of video bandwidth, immersive bandwidth, or both. For TelePresence endpoints there is a non-configurable Video Call Traffic Class of **immersive** assigned to the endpoint. SIP trunks can be configured through the SIP Profile as either desktop video, high definition immersive video, or a system that has both classifications of video endpoints, such as a Cisco TelePresence System Video Communications Server (VCS). All other endpoints and trunks have a non-configurable Video Call Traffic Class of **desktop video**.

TelePresence immersive endpoints mark their media with a DSCP value of CS4 by default, and desktop video endpoints mark their media with AF41 by default, as per recommended QoS settings. For Cisco endpoints this is accomplished through the configurable Unified CM QoS service parameters **DSCP for Video calls** and **DSCP for TelePresence calls**. When a Cisco TelePresence endpoint registers with Unified CM, it downloads a configuration file and applies the QoS setting of **DSCP for TelePresence calls**. When a Unified Communications video-capable endpoint registers with Unified CM, it downloads a configuration file and applies the QoS setting of **DSCP for Video calls**. All third-party video endpoints require manual configuration of the endpoints themselves and are statically configured, meaning they do not change QoS marking depending on the call type; therefore, it is important to match the Enhanced Locations CAC bandwidth allocation to the correct DSCP. Unified CM achieves this by deducting desktop video calls from the Video Bandwidth location and link allocation for devices that have a Video Call Traffic Class of **desktop**. End-to-end TelePresence immersive video calls are deducted from the Immersive Video Bandwidth location and link allocation for devices or trunks with the Video Call Traffic Class of **immersive**. This ensures that end-to-end desktop video and immersive video calls are marked correctly and counted correctly for call admission control. For calls between desktop devices and TelePresence immersive devices, bandwidth is deducted from both the Video Bandwidth and the Immersive Video Bandwidth location and link allocations.

TelePresence Endpoints

TelePresence endpoints have a fixed non-configurable Video Call Traffic Class of **immersive** and are identified by Unified CM as immersive.

Bandwidth reservations are determined by the classification of endpoints in a video call, and they deduct bandwidth from the locations and links bandwidth pools as listed in Table 11-4.

Table 11-4 Bandwidth Pool Usage per Endpoint Type

Endpoint A	Endpoint B	Locations and Links Pool Used
Immersive video	Immersive video	Immersive bandwidth
Immersive video	Desktop video	Immersive and video bandwidth
Desktop video	Desktop video	Video bandwidth
Audio-only call	Any	Audio bandwidth

SIP Trunks

A SIP trunk can also be classified as desktop, immersive, or mixed video in order to deduct bandwidth reservations of a SIP trunk call, and the classification is determined by the calling device type and Video Call Traffic Class of the SIP trunk. The SIP trunk can be configured through the SIP Profile trunk-specific information as:

- Immersive — High-definition immersive video
- Desktop — Standard desktop video
- Mixed — A mix of immersive and desktop video

A SIP trunk can be classified with any of these three classifications and is used primarily to classify Video or TelePresence Multipoint Control Units (MCUs), a video device at a fixed location, a Unified Communications system such as Unified CM prior to version 9.0, or a Cisco TelePresence System Video Communications Server (VCS).

Bandwidth reservations are determined by the classification of an endpoint and a SIP trunk in a video call, and they deduct bandwidth from the locations and links bandwidth pools as listed in Table 11-5.

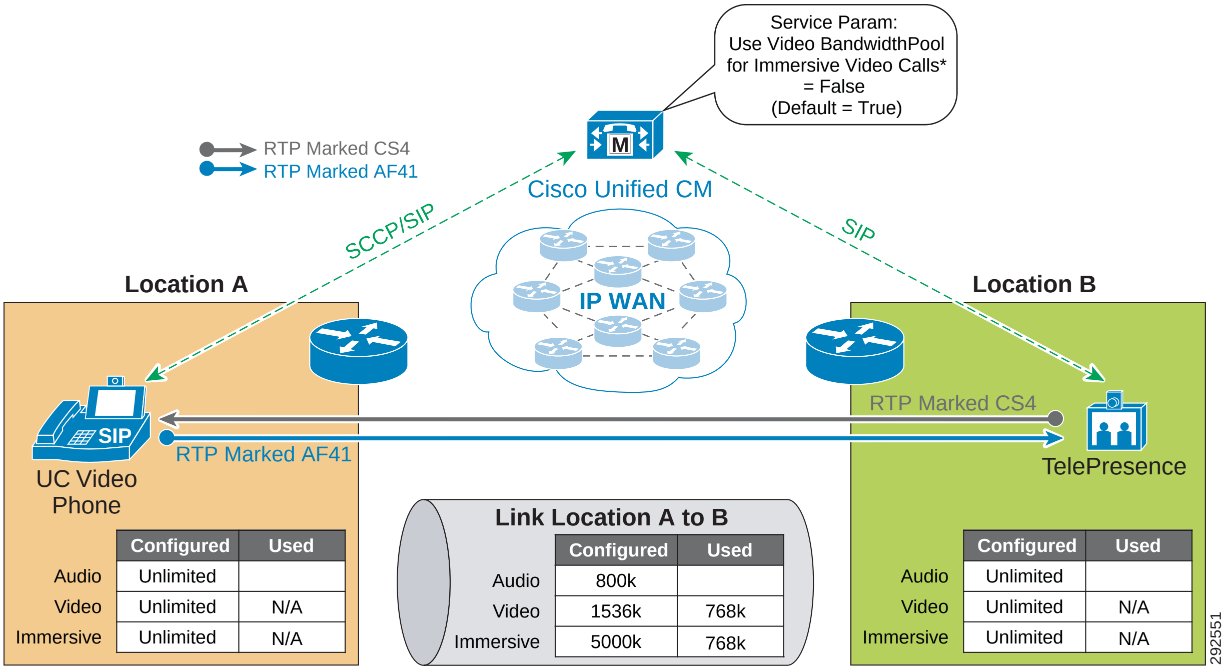
Table 11-5 Bandwidth Pool Usage per SIP Trunk and Endpoint Type

Endpoint	SIP Trunk	Locations and Links Pool Used
TelePresence endpoint	Immersive	Immersive bandwidth
TelePresence endpoint	Desktop	Immersive and video bandwidth
TelePresence endpoint	Mixed	Immersive and video bandwidth
Desktop endpoint	Immersive	Immersive and video bandwidth
Desktop endpoint	Desktop	Video bandwidth
Desktop endpoint	Mixed	Immersive and video bandwidth
Non-video endpoint	Any	Audio bandwidth

By default, all video calls from either immersive or desktop endpoints is deducted from the locations and links video bandwidth pool. To change this behavior, set the Unified CM service parameter **Use Video BandwidthPool for Immersive Video Calls** to **False**, and this will enable the immersive video bandwidth deductions.

As described earlier, a video call between a Unified Communications video endpoint (desktop Video Call Traffic Class) and a TelePresence endpoint (immersive Video Call Traffic Class) will mark their media asymmetrically and, when immersive video CAC is enabled, will deduct bandwidth from both video and immersive locations and links bandwidth pools. Figure 11-19 illustrates this.

Figure 11-19 Enhanced Locations CAC Bandwidth Deductions and Media Marking for a Multi-Site Deployment



Location A

	Configured	Used
Audio	Unlimited	
Video	Unlimited	N/A
Immersive	Unlimited	N/A

Link Location A to B

	Configured	Used
Audio	800k	
Video	1536k	768k
Immersive	5000k	768k

Location B

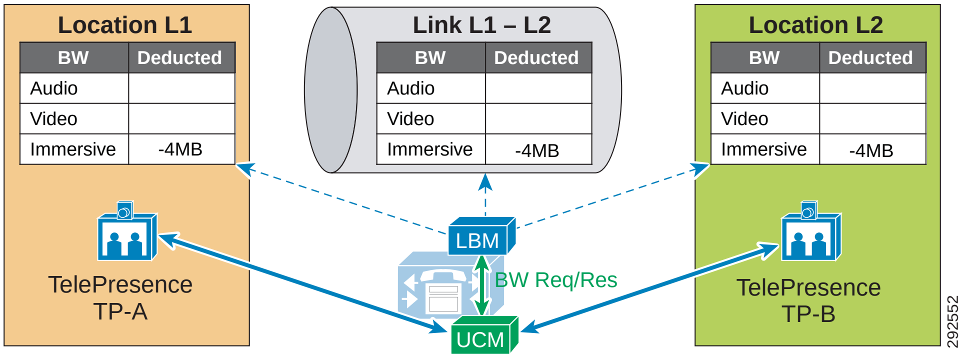
	Configured	Used
Audio	Unlimited	
Video	Unlimited	N/A
Immersive	Unlimited	N/A

Examples of Various Call Flows and Location and Link Bandwidth Pool Deductions

The following call flows depict the expected behavior of locations and links bandwidth deductions when the Unified CM service parameter **Use Video BandwidthPool for Immersive Video Calls** is set to **False**.

Figure 11-20 illustrates an end-to-end TelePresence immersive video call between TP-A in Location L1 and TP-B in Location L2. End-to-end immersive video endpoint calls deduct bandwidth from the immersive bandwidth pool of the locations and the links along the effective path.

Figure 11-20 Call Flow for End-to-End TelePresence Immersive Video



Location L1

BW	Deducted
Audio	
Video	
Immersive	-4MB

Link L1 – L2

BW	Deducted
Audio	
Video	
Immersive	-4MB

Location L2

BW	Deducted
Audio	
Video	
Immersive	-4MB

Figure 11-21 illustrates an end-to-end desktop video call between DP-A in Location L1 and DP-B in Location L2. End-to-end desktop video endpoint calls deduct bandwidth from the video bandwidth pool of the locations and the links along the effective path.

Figure 11-21 Call Flow for End-to-End Desktop Video

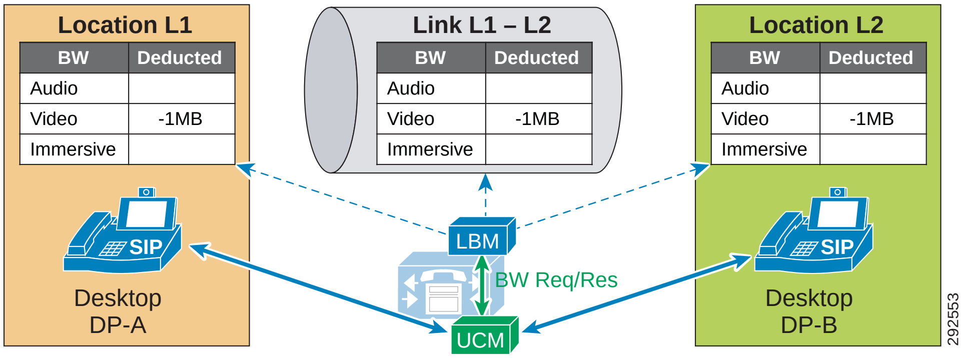
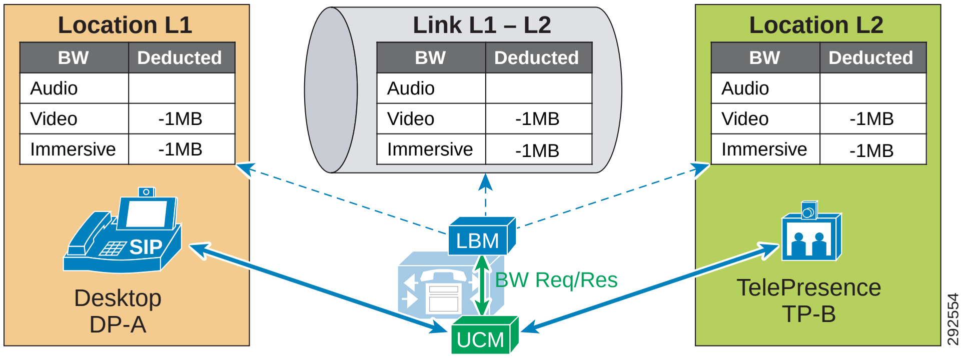


Figure 11-22 illustrates a video call between desktop video endpoint DP-A in Location L1 and TelePresence video endpoint TP-B in Location L2. Interoperating calls between desktop video endpoints and TelePresence video endpoints deduct bandwidth from both video and immersive locations and the links bandwidth pools along the effective path.

Figure 11-22 Call Flow for Desktop-to-TelePresence Video



In Figure 11-23, a desktop video endpoint and two TelePresence endpoints call a SIP trunk configured with a Video Traffic Class of **immersive** that points to a TelePresence MCU. Bandwidth is deducted along the effective path from the immersive locations and the links bandwidth pools for the calls that are end-to-end immersive and from both video and immersive locations and the links bandwidth pools for the call that is desktop-to-immersive.

Figure 11-23 Call Flow for a Video Conference with an MCU

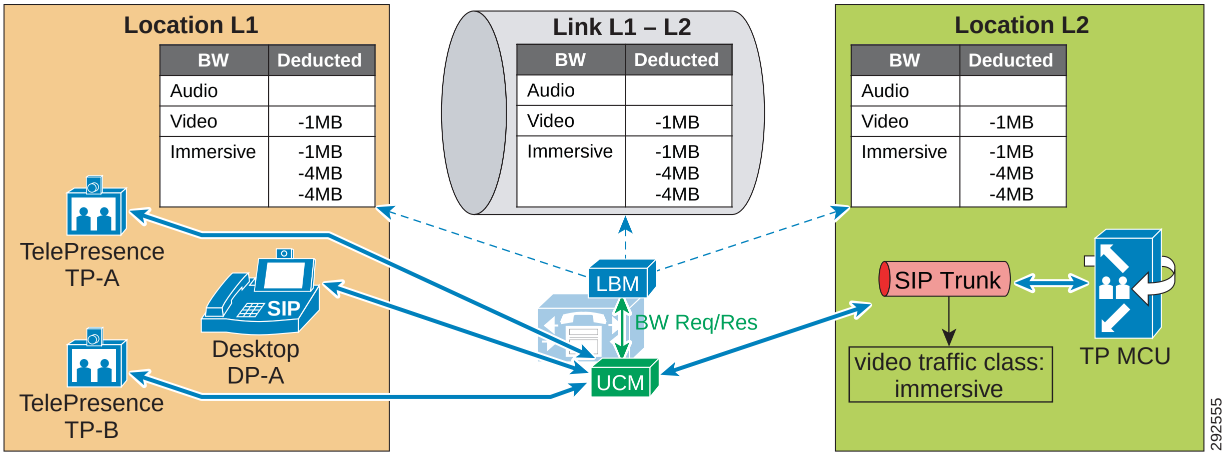


Figure 11-24 illustrates an end-to-end immersive video call across clusters, which deducts bandwidth from the immersive bandwidth pool of the locations and links along the effective path.

Figure 11-24 Call Flow for End-to-End TelePresence Immersive Video Across Clusters

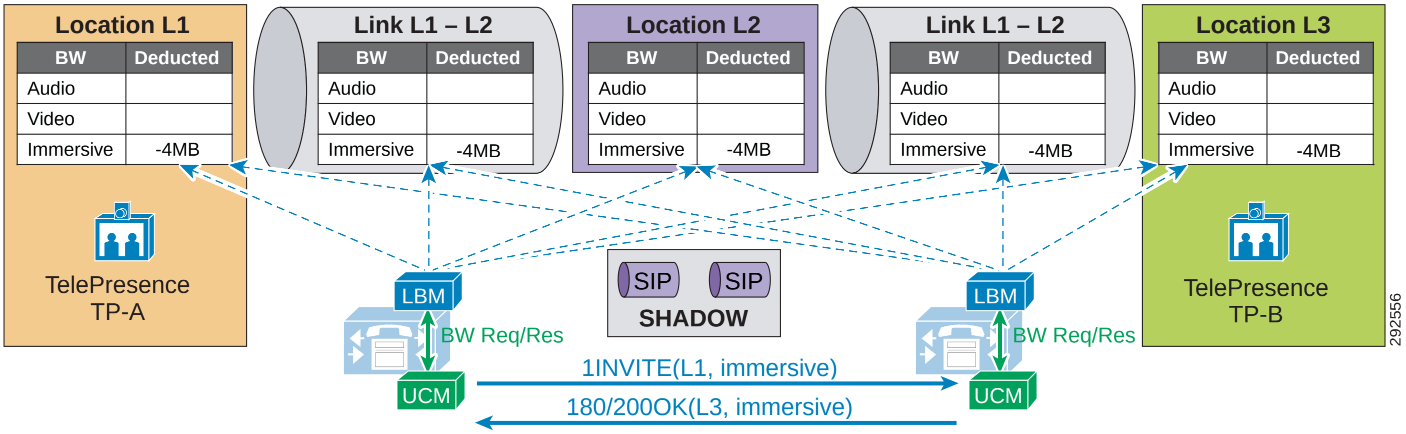
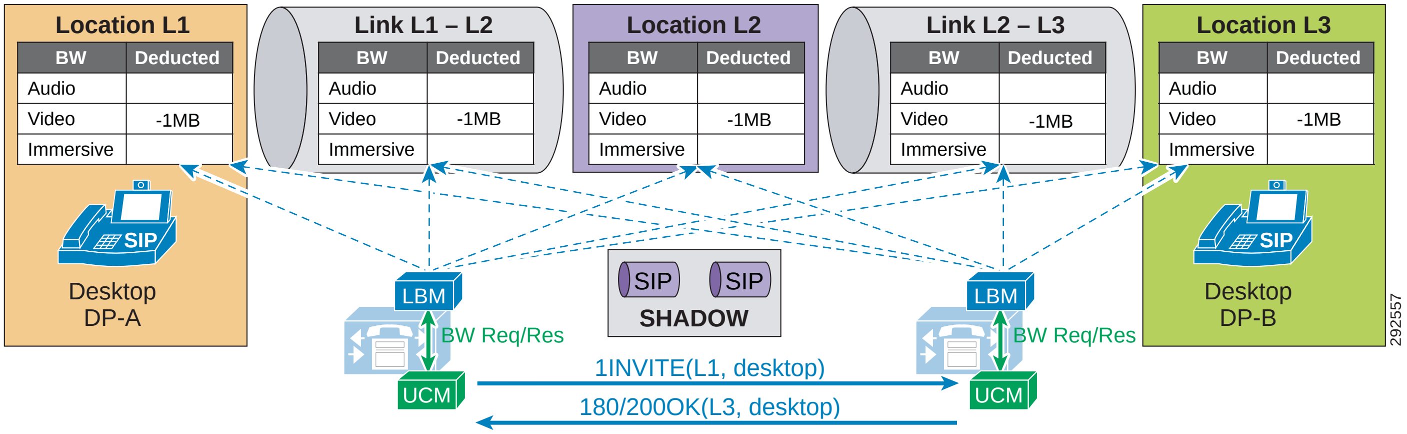


Figure 11-25 illustrates an end-to-end desktop video call across clusters, which deducts bandwidth from the video bandwidth pool of the locations and links along the effective path.

Figure 11-25 Call Flow for End-to-End Desktop Video Call Across Clusters



Location L1

BW	Deducted
Audio	
Video	-1MB
Immersive	

Link L1 – L2

BW	Deducted
Audio	
Video	-1MB
Immersive	

Location L2

BW	Deducted
Audio	
Video	-1MB
Immersive	

Link L2 – L3

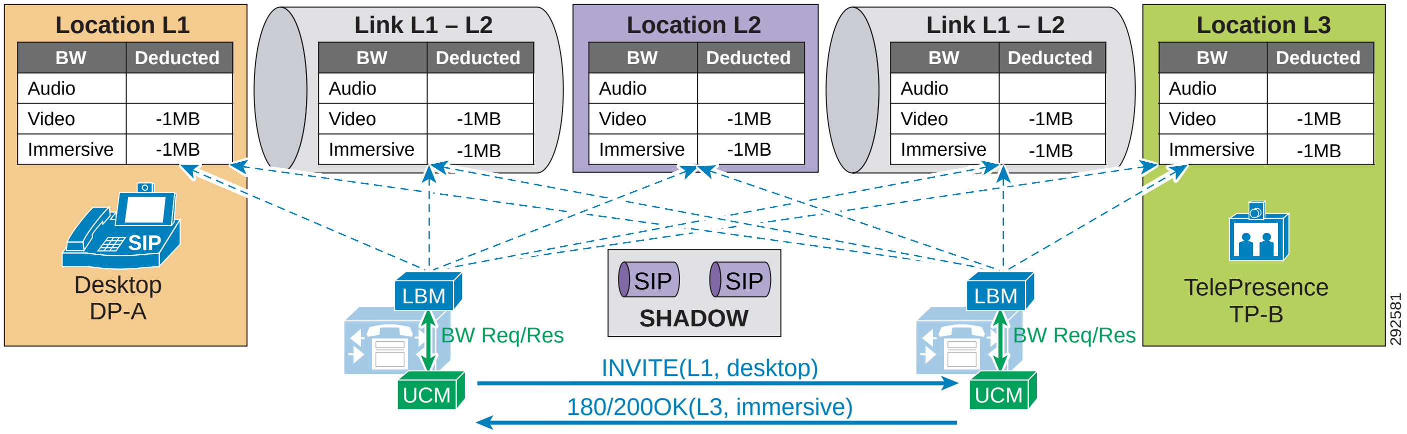
BW	Deducted
Audio	
Video	-1MB
Immersive	

Location L3

BW	Deducted
Audio	
Video	-1MB
Immersive	

Figure 11-26 illustrates a desktop video endpoint calling a TelePresence endpoint across clusters. the call deducts bandwidth from both video and immersive bandwidth pools of the locations and links along the effective path.

Figure 11-26 Call Flow for Desktop-to-TelePresence Video Across Clusters



Location L1

BW	Deducted
Audio	
Video	-1MB
Immersive	-1MB

Link L1 – L2

BW	Deducted
Audio	
Video	-1MB
Immersive	-1MB

Location L2

BW	Deducted
Audio	
Video	-1MB
Immersive	-1MB

Link L1 – L2

BW	Deducted
Audio	
Video	-1MB
Immersive	-1MB

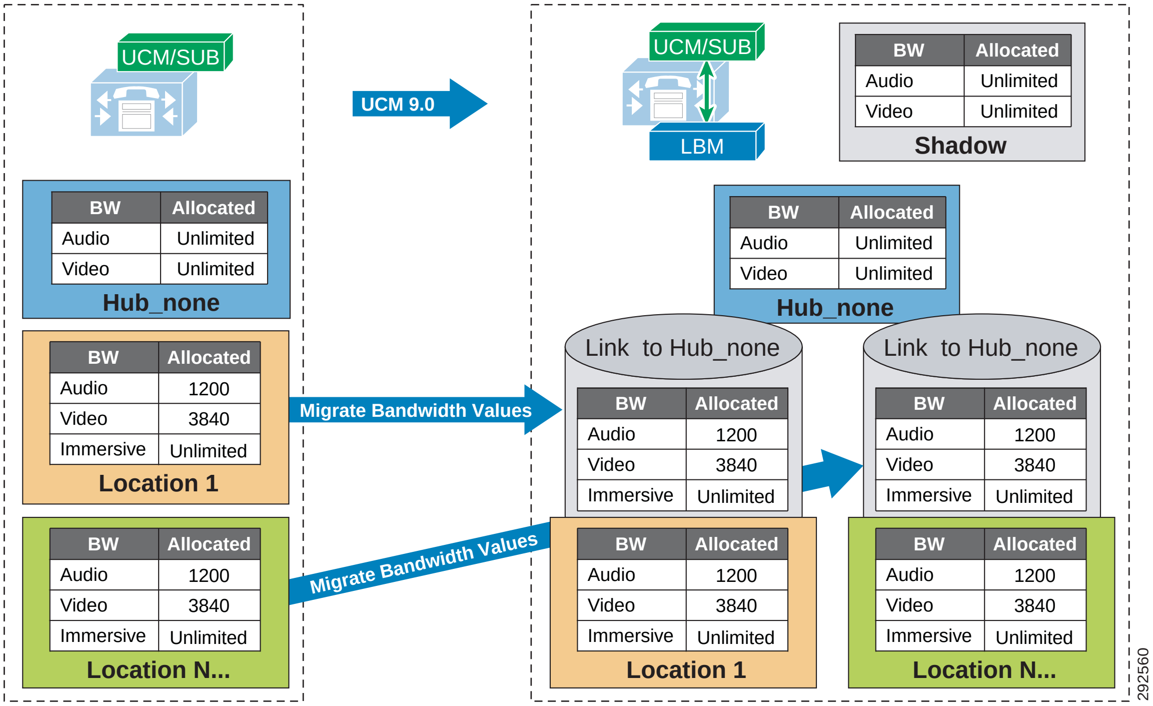
Location L3

BW	Deducted
Audio	
Video	-1MB
Immersive	-1MB

Upgrade and Migration from Locations CAC to Enhanced Locations CAC

Upgrading to Cisco Unified CM 9.*x* from a previous release will result in the migration of Locations CAC to Enhanced Locations CAC. The migration consists of taking all previously defined locations bandwidth limits of audio and video bandwidth and migrating them to a link between the user-defined location and Hub_None. This effectively recreates the hub-and-spoke model that previous versions of Unified CM Locations CAC supported. Figure 11-27 illustrates the migration of bandwidth information.

Figure 11-27 Migration from Locations CAC to Enhanced Locations CAC After Unified CM Upgrade



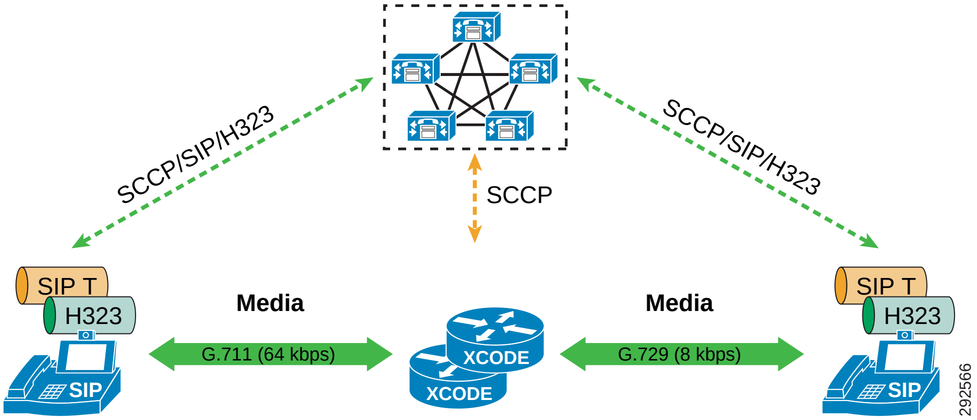
Settings after an upgrade to Cisco Unified CM 9.*x*:

- The LBM is activated on each Unified CM subscriber running the Cisco CallManager service.
- The Cisco CallManager service communicates directly with the local LBM.
- No LBM group or LBM hub group is created.
- All LBM services are fully meshed.
- Intercluster Enhanced Locations CAC is not enabled.
- All intra-location bandwidth values are set to unlimited.
- Bandwidth values assigned to locations are migrated to a link connecting the user-defined location and Hub_None.
- Immersive bandwidth is set to unlimited.
- A shadow location is created.

- Phantom and shadow locations have no links.
- Enhanced Locations CAC bandwidth adjustment for MTPs and transcoders:

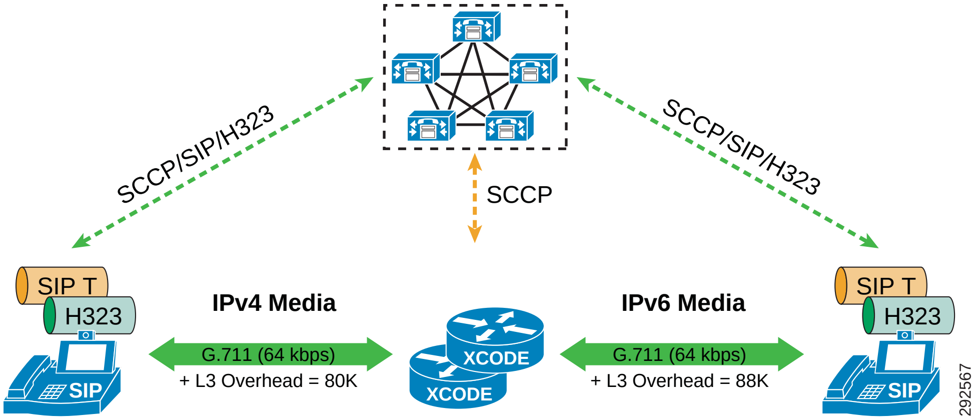
  For transcoding insertion, the bit rate is different on each leg of the connection. Figure 11-28 illustrates this.

Figure 11-28 Example of Different Bit Rate for Transcoding



For dual stack MTP insertion, the bit rate is different on each connection but the bandwidth is different due to IP header overhead. Figure 11-29 illustrates the difference in bandwidth used for IPv4 and IPv6 networks with dual stack MTP insertion.

Figure 11-29 Bandwidth Differences for Dual Stack MTP Insertion



Enhanced Locations CAC does not account for these differences in bandwidth between MTPs and transcoders. The service parameter **Locations Media Resource Audio Bit Rate Policy** determines whether the largest or smallest bandwidths should be used along the locations and links path. Lowest Bit Rate (default) or Highest Bit Rate can be used to manage these differences in bandwidth consumption.

Cisco IOS Gatekeeper Zones

A Cisco IOS gatekeeper can provide call routing and call admission control between devices such as Cisco Unified CM, Cisco Unified Communications Manager Express (Unified CME), or H.323 gateways connected to legacy PBXs. It uses the H.323 Registration Admission Status (RAS) protocol to communicate with these devices and route calls across the network.

Gatekeeper call admission control is a policy-based scheme requiring static configuration of available resources. The gatekeeper is not aware of the network topology, so it is limited to simple hub-and-spoke topologies.

For a listing of the available Cisco IOS gatekeeper platforms and the features supported on each platform, refer to the *Cisco IOS H323 Gatekeeper Data Sheet* at

http://www.cisco.com/en/US/prod/collateral/voicesw/ps6788/vcallcon/ps4139/data_sheet_c78_561921.html

The call admission control capabilities of a Cisco IOS gatekeeper are based on the concept of gatekeeper *zones*. A zone is a collection of H.323 devices, such as endpoints, gateways, or Multipoint Control Units (MCUs), that register with a gatekeeper. There can be only one active gatekeeper per zone, and you can define up to 100 local zones on a single gatekeeper. A local zone is a zone that is actively handled by that gatekeeper – that is, all H.323 devices assigned to that zone register with that gatekeeper.

When multiple gatekeepers are deployed in the same network, a zone is configured as a local zone on only one gatekeeper. On the other gatekeepers, that zone is configured as a remote zone. This configuration instructs the gatekeeper to forward calls destined for that zone to the gatekeeper that "owns it" (that is, the gatekeeper on which that zone is configured as a local zone).

For details on configuring the gatekeeper, refer to the *Cisco IOS H.323 Configuration Guide* at

http://www.cisco.com/en/US/docs/ios/voice/h323/configuration/guide/15_0/vh_15_0_book.html

The bandwidth value deducted by the gatekeeper for every active call is double the bit-rate of the call, excluding Layer 2, IP, and RTP overhead. For example, a G.711 audio call that uses 64 kbps would be denoted as 128 kbps in the gatekeeper, and a 384-kbps video call would be denoted as 768 kbps. Table 11-6 shows the bandwidth values used by the gatekeeper feature for some of the most popular call speeds.

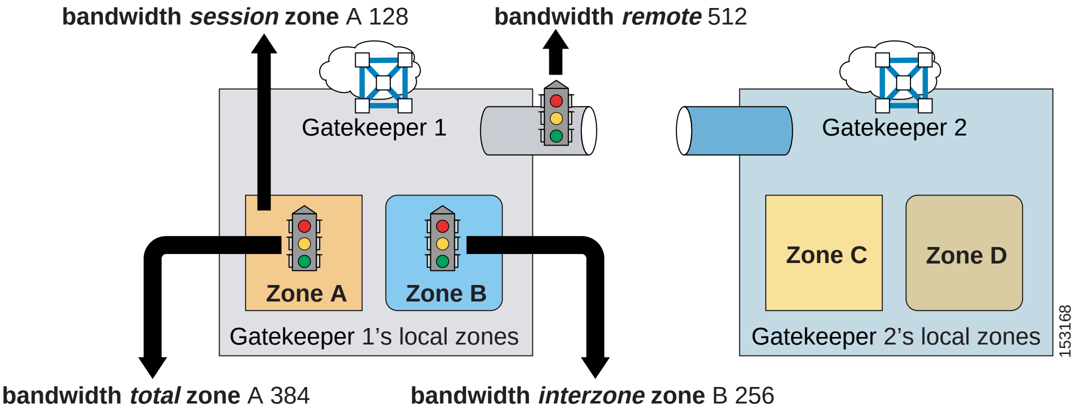
Table 11-6 Gatekeeper Bandwidth Settings for Various Call Speeds

Call Speed	Gatekeeper Bandwidth Value
G.711 audio call (64 kbps)	128 kbps
G.729 audio call (8 kbps)	16 kbps
128-kbps video call	256 kbps
384-kbps video call	768 kbps
512-kbps video call	1024 kbps
768-kbps video call	1536 kbps

Note Bandwidth calculations for the call Admission Request (ARQ) do not include compressed Real-Time Transport Protocol (cRTP) or any other transport overhead. See Bandwidth Provisioning, page 3-45, for details on how to provision interface queues.

To better understand the application of the **bandwidth** commands in a real network, consider the example shown in Figure 11-30.

Figure 11-30 Example of Cisco IOS Gatekeeper bandwidth Commands


bandwidth *session* zone A 128
bandwidth *remote* 512
Gatekeeper 1
Gatekeeper 2
Zone A
Zone B
Zone C
Zone D
Gatekeeper 1's local zones
Gatekeeper 2's local zones
153168
bandwidth *total* zone A 384
bandwidth *interzone* zone B 256

Assuming that all calls are voice-only calls using the G.711 codec, and given the configuration commands shown in Figure 11-30, the following statements hold true:

- The maximum amount of bandwidth requested by any device in zone A for a single call is 128 kbps, which means that calls trying to use codecs with a higher bit-rate than 64 kbps will be rejected.
- The maximum amount of bandwidth used by all calls involving devices in zone A (either within the zone or with other zones) is 384 kbps, which means that there can be at most three active calls involving devices in zone A.
- The maximum amount of bandwidth used by all calls between devices in zone B and devices in any other zone is 256 kbps, which means that there can be at most two active calls between devices in zone B and devices in zones A, C, and D.
- The maximum amount of bandwidth used by all calls between devices registered with gatekeeper GK 1 and devices registered with any other gatekeeper is 512 kbps, which means that there can be at most four active calls between devices in zones A and B and devices in zones C and D.

Unified Communications Architectures Using Resource Reservation Protocol (RSVP)

This section covers the various Unified Communications architectures that implement Resource Reservation Protocol (RSVP) as the call admission control mechanism. The section begins with an introduction to RSVP and an overview of the protocol architecture, concepts of RSVP and Quality of Service, Application ID, and a summary of the infrastructure design considerations and recommendations.

Next this section discusses Unified CM RSVP-enabled locations in a single-cluster Unified CM environment. The discussion covers the components involved as well as the provisioning of those components, Unified CM's use of RSVP policy and Application ID, and a recommended migration strategy from call admission control based on Unified CM Enhanced Locations.

This section then covers distributed call processing architectures, beginning with RSVP SIP Preconditions, with an overview of the feature and how it works to synchronize the RSVP layer and call control layer between the various call control applications such as Unified CM, Unified CME, and SIP-TDM Cisco IOS Gateways. Then each call control application is discussed in further detail with regard to RSVP SIP Preconditions, including feature notes and design recommendations and considerations.

Resource Reservation Protocol (RSVP)

The Resource Reservation Protocol (RSVP) is the first significant industry-standard protocol for dynamically setting up end-to-end QoS across a heterogeneous network. RSVP, which runs over IP, was first introduced by the IETF in RFC 2205, and it enables an application to reserve network bandwidth dynamically. Using RSVP, applications can request a certain level of QoS for a data flow across a network. Because of its distributed and dynamic nature, RSVP is capable of reserving bandwidth across any network topology, therefore it can be used to provide topology-aware call admission control for voice and video calls.

This section focuses on the RSVP protocol principles and its interactions with the WAN infrastructure, specifically the QoS aspects, while the motivation and the mechanisms for call admission control based on RSVP are described in other sections of this chapter.

This section covers the following specific topics:

- RSVP Principles, page 11-42
- RSVP in MPLS Networks, page 11-45
- RSVP and QoS in WAN Routers, page 11-48
- RSVP Application ID, page 11-52
- RSVP Design Best Practices, page 11-57

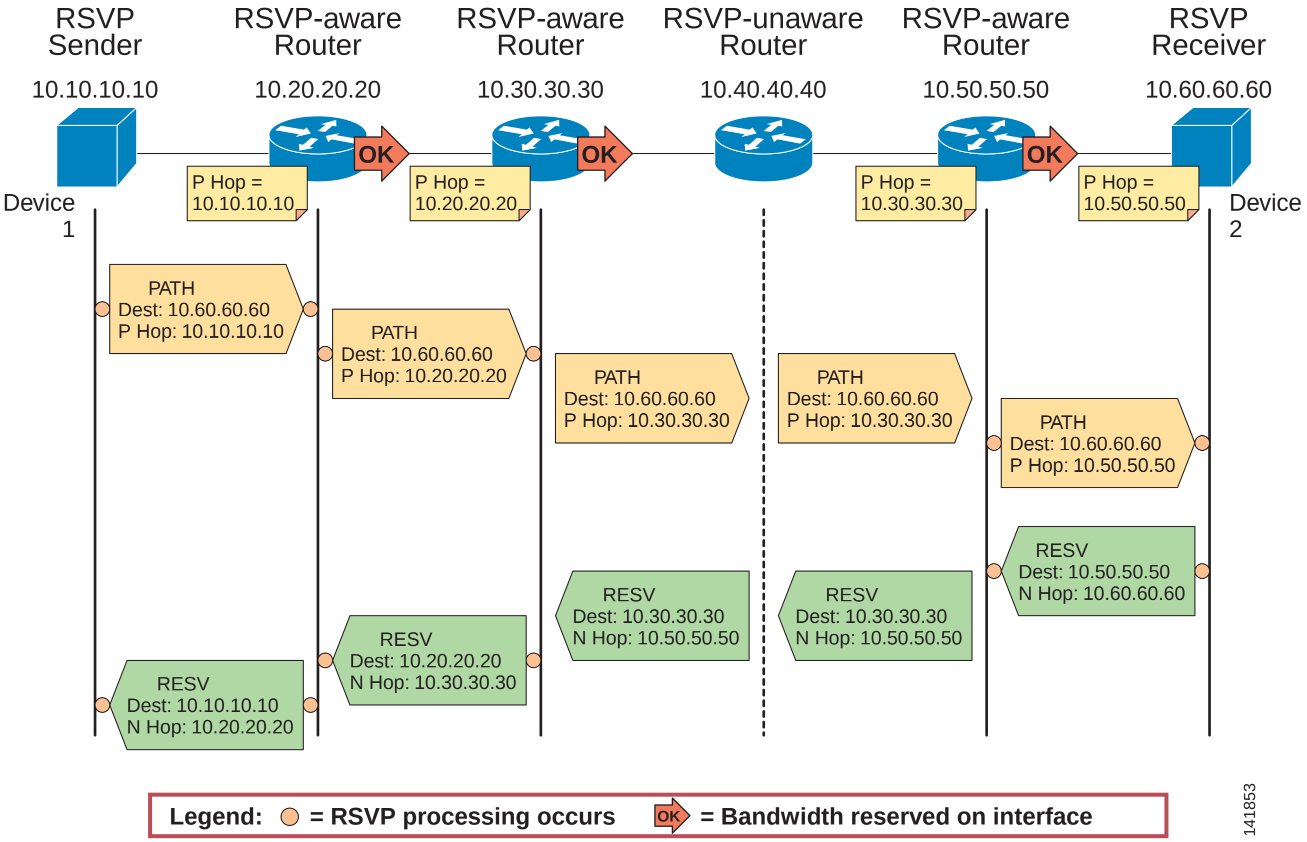
RSVP Principles

RSVP performs resource reservations for a given data flow across a network. RSVP reservations are unidirectional. Therefore, for a single audio call that contains two RTP streams, two RSVP reservations are generated, one for each RTP stream. The resource reservation is created by exchanging signaling messages between the source and destination devices for the data flow, and the messages are processed by intermediate routers along the path. The RSVP signaling messages are IP packets with the protocol number in the IP header set to 46, and they are routed through the network according to the existing routing protocols.

Not all routers on the path are required to support RSVP because the protocol is designed to operate transparently across RSVP-unaware nodes. On each RSVP-enabled router, the RSVP process intercepts the signaling messages and interacts with the QoS manager for the router's outbound interface involved in the data flow in order to "reserve" bandwidth resources. When the available resources are not sufficient for the data flow anywhere along the path, the routers signal the failure back to the application that originated the reservation request.

The principles of RSVP signaling can be explained by using the example shown in Figure 11-31. In this diagram, an application wishes to reserve network resources for a data stream flowing from Device 1, whose IP address is 10.10.10.10, to Device 2, whose IP address is 10.60.60.60.

Figure 11-31 Example of RSVP Path and Resv Message Flow


RSVP Sender
RSVP-aware Router
RSVP-aware Router
RSVP-unaware Router
RSVP-aware Router
RSVP Receiver
10.10.10.10
10.20.20.20
10.30.30.30
10.40.40.40
10.50.50.50
10.60.60.60
Device 1
Device 2
OK
P Hop = 10.10.10.10
OK
P Hop = 10.20.20.20
OK
P Hop = 10.30.30.30
P Hop = 10.50.50.50
PATH Dest: 10.60.60.60 P Hop: 10.10.10.10
PATH Dest: 10.60.60.60 P Hop: 10.20.20.20
PATH Dest: 10.60.60.60 P Hop: 10.30.30.30
PATH Dest: 10.60.60.60 P Hop: 10.30.30.30
PATH Dest: 10.60.60.60 P Hop: 10.50.50.50
RESV Dest: 10.50.50.50 N Hop: 10.60.60.60
RESV Dest: 10.30.30.30 N Hop: 10.50.50.50
RESV Dest: 10.30.30.30 N Hop: 10.50.50.50
RESV Dest: 10.20.20.20 N Hop: 10.30.30.30
RESV Dest: 10.10.10.10 N Hop: 10.20.20.20
Legend: ○ = RSVP processing occurs OK = Bandwidth reserved on interface
141853

The following steps describe the RSVP signaling process for as single data flow, as shown by the example in Figure 11-31:

1. The application residing on Device 1 originates an RSVP message called Path, which is sent to the same destination IP address as the data flow for which a reservation is requested (that is, 10.60.60.60) and is sent with the "router alert" option turned on in the IP header. The Path message contains, among other things, the following objects:
   - The "session" object, consisting of destination IP address, protocol number, and UDP/TCP port, which is used to identify the data flow in RSVP-enabled routers.
   - The "sender T-Spec" (traffic specification) object, which characterizes the data flow for which a reservation will be requested. The T-Spec basically defines the maximum IP bandwidth required for a call flow using a specific codec. The T-Spec is typically defined using values for the data flow's average bit rate, peak rate, and burst size. Details of the T-Spec are discussed later in this chapter.
   - The "P Hop" (or previous hop) object, which contains the IP address of the router interface that last processed the Path message. In this example, the P Hop is initially set to 10.10.10.10 by Device 1.
2. By means of the "router alert" option, the Path message is intercepted by the CPU of the RSVP-aware router identified as 10.20.20.20 in Figure 11-31, which sends it to the RSVP process. RSVP creates a path state for this data flow, storing the values of the session, sender Tspec, and

P Hop objects contained in the Path message. Then it forwards the message downstream, after having replaced the P Hop value with the IP address of its outgoing interface (10.20.20.20 in this example).

3. Similarly, the Path message is intercepted by the CPU of the following RSVP-aware router, identified as 10.30.30.30 in Figure 11-31. After creating the path state and changing the P Hop value to 10.30.30.30, this router also forwards the message downstream.
4. The Path message now arrives at the RSVP-unaware router identified as 10.40.40.40 in Figure 11-31. Because RSVP is not enabled on this router, it just routes this message according to the existing routing protocols like any other IP packet, without any additional processing and without changing the content of any of the message objects.
5. Therefore, the Path message gets to the RSVP-aware router identified as 10.50.50.50, which processes the message, creates the corresponding path state, and forwards the message downstream. Notice that the P Hop recorded by this router still contains the IP address of the last RSVP-aware router along the network path, or 10.30.30.30 in this example.
6. The RSVP Receiver at Device 2 receives the Path message with a P Hop value of 10.50.50.50, and it can now initiate the actual reservation by originating a message called Resv. For this reason, RSVP is known as a receiver-initiated protocol. The Resv message carries the reservation request hop-by-hop from the receiver to the sender, along the reverse paths of the data flow for the session. At each hop, the IP destination address of the Resv message is the IP address of the previous-hop node, obtained from the path state. Hence, in this case Device 2 sends the Resv message with a destination IP address of 10.50.50.50. The Resv message contains, among other things, the following objects:
   - The "session" object, which is used to identify the data flow.
   - The "N Hop" (or next hop) object, which contains the IP address of the node that generated the message. In this example, the N Hop is initially set to 10.60.60.60 by Device 2.
7. When RSVP-aware router 10.50.50.50 receives the Resv message for this data flow, it matches it against the path state information using the received session object, and it verifies if the reservation request can be accepted based on the following criteria:
   - Policy control — Is this user and/or application allowed to make this reservation request?
   - Admission control — Are there enough bandwidth resources available on the relevant outgoing interface to accommodate this reservation request?
8. In this case, we assume that both policy and admission control are successful on 10.50.50.50, which means that the bandwidth provided by the Tspec in the path state for this session is reserved on the outgoing interface (in the same direction as the data flow, that is from Device 1 to Device 2), and a corresponding "reservation state" is created. Now router 10.50.50.50 can send a Resv message upstream by sending it as a unicast IP packet to the destination IP address stored in the P Hop for this session, which was 10.30.30.30. The N Hop object is also updated with the value of 10.50.50.50.
9. The Resv message now transits through the RSVP-unaware router identified as 10.40.40.40, which will route it toward its destination of 10.30.30.30 like any other IP packet. This mechanism allows RSVP signaling to work across a heterogeneous network where some nodes are not RSVP-enabled.
10. The RSVP-aware router identified as 10.30.30.30 receives the Resv message and processes it according to the mechanisms described in steps 7 and 8. Assuming policy and admission control are successful also at this hop, the bandwidth is reserved on the outgoing interface and a Resv message is sent to the previous hop, or 10.20.20.20 in this example.
11. After a similar process within the router identified as 10.20.20.20, the Resv finally reaches the RSVP sender, Device 1. This indicates to the requesting application that an end-to-end reservation has been established and that bandwidth has been set aside for this data flow in all RSVP-enabled routers across the network.

This example shows how the two main RSVP signaling messages, Path and Resv, travel across the network to establish reservations. Several other messages are defined in the RSVP standard to address error situations, reservation failures, and release of resources. In particular, the ResvErr message is used to signal failure to reserve the requested resources due to either policy control or admission control somewhere along the network. If, for example, admission control had failed at node 10.50.50.50 in Figure 11-31, this node would have sent a ResvErr message back to Device 2, specifying the cause of the failure, and the application would have been notified.

Another important aspect of the RSVP protocol is that it adopts a soft-state approach, which means that for each session both the path state and the reservation state along the network need to be refreshed periodically by the application by sending identical Path and Resv messages. If a router does not receive refresh messages for a given session for a certain period of time, it deletes the corresponding state and releases the resources reserved. This allows RSVP to react dynamically to network topology changes or routing changes due to link failures. The reservations simply start flowing along the new routes based on the routing protocol decisions, and the reservations along the old routes time-out and are eventually deleted.

RSVP in MPLS Networks

In some MPLS service-provider networks, the IP addresses used on the links between the customer edge (CE) and the provider edge (PE) are not distributed to the rest of the MPLS network, thus ensuring that the subnets stay local to the PE and are not advertised beyond the PE (because they are not unique and are being reused elsewhere). This creates a situation where RSVP is not able to forward RSVP messages because the P Hop (Previous Hop) value of the RSVP message is unknown in the network. Figure 11-32 illustrates this type of situation.

Figure 11-32 RSVP Over MPLS Without P Hop Overwrite

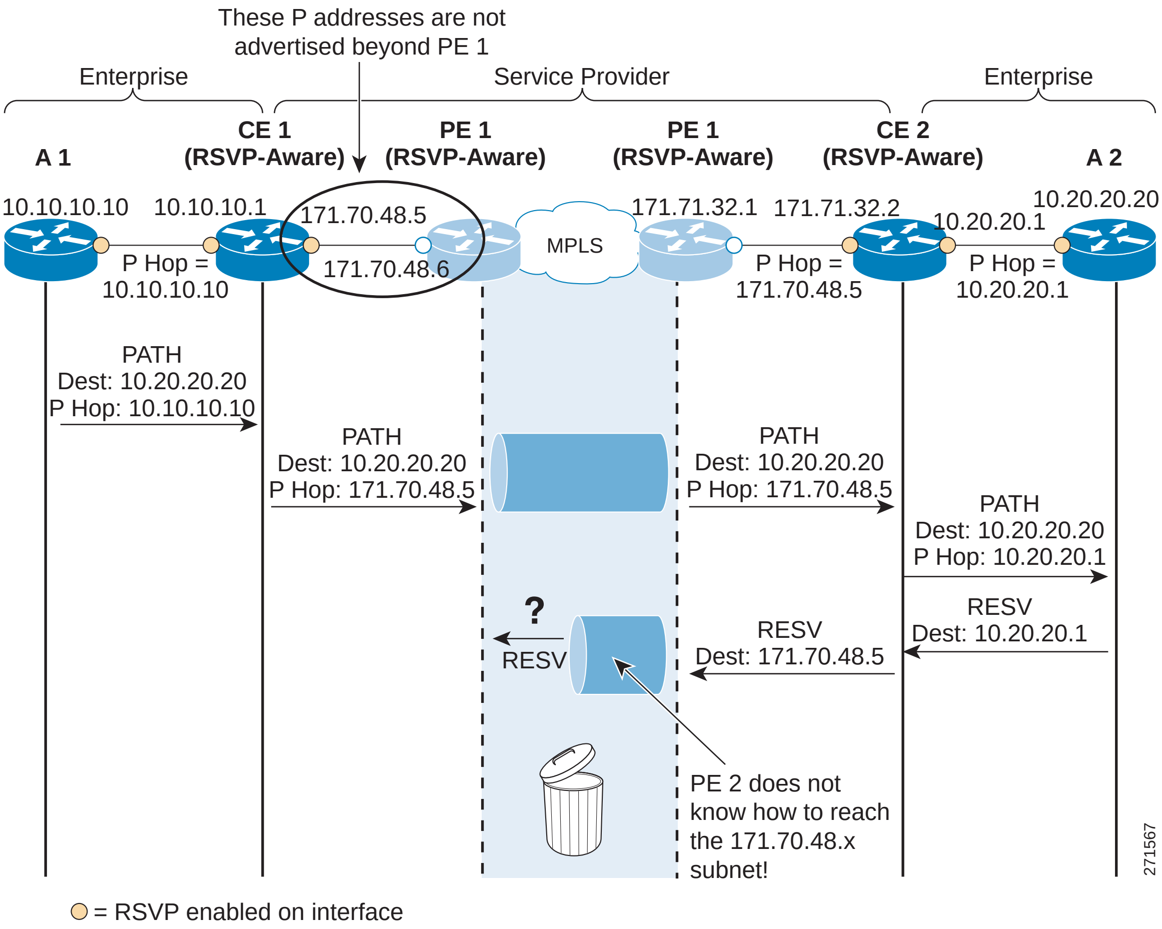


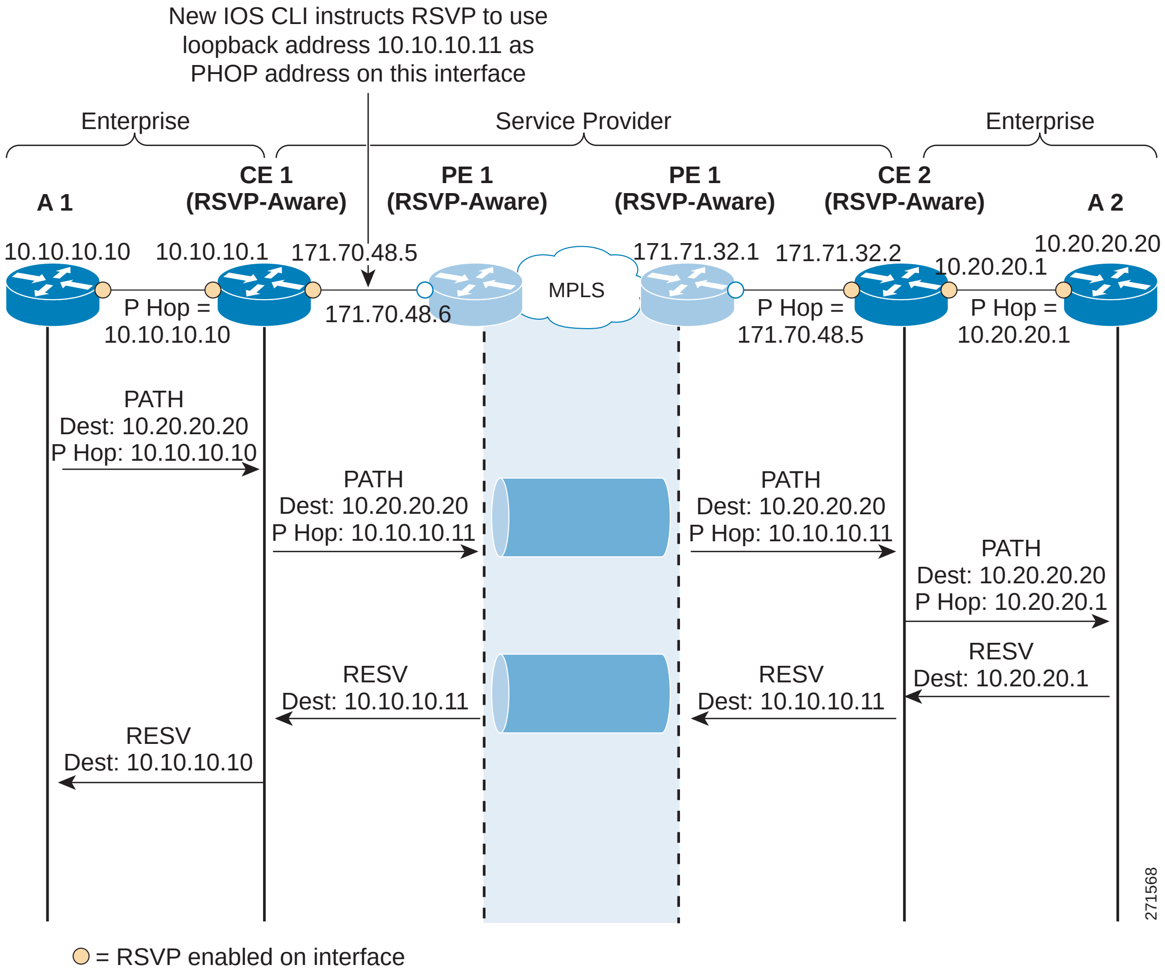
Figure 11-32 shows an enterprise network and a service provider MPLS network. CE1 and CE2 are RSVP-aware, and PE1 and PE2 are RSVP-unaware. The RSVP Path message contains a P Hop object. This object is rewritten at every RSVP hop. Its purpose is to enable an RSVP router (for example, CE1) to send a Path message to the next RSVP router (for example, CE2) to indicate that it (CE1) is the previous RSVP hop (or P Hop). This information is used by CE2 to forward the corresponding Resv message upstream hop-by-hop toward the sender.

In Cisco IOS, the RSVP Router always sets the P Hop address to the IP address of the egress interface onto which it transmits the Path message. There are situations where, although some IP addresses of CE1 are reachable, the IP address of its egress interface is not reachable from a remote RSVP Router CE2. The result is that the corresponding Resv message generated by CE2 never reaches CE1, thus the reservation is never established.

When a call is made from A1 to A2, A1 tries to set up an RSVP session and starts by sending a Path message to CE1. A1 will populate the P Hop object in the Path message of its outgoing interface IP (in this case, 10.10.10.10). CE1 will then receive the Path message, process it, create the corresponding path state, update the P Hop field of the message with its egress interface IP address (171.70.48.5), which is not a routable IP address, and forward the Path message downstream. This Path message will be tunneled across the service provider network and will be processed by CE2. Upon reception of the Path message, CE2 records the IP address of the P Hop object (CE1's egress interface IP address) and forwards the Path message downstream to A1. A1 will record and process the Path message and initiate an RSVP message to CE2. CE2 will process the RSVP message and send it's own RSVP message upstream to CE1. However, when CE2 replies with this Resv message, it will try to send it to the IP address that it had recorded earlier from the Path message received from CE1. Since this IP address (171.70.48.5) is not routable from CE2, the Resv message will fail, thus causing the reservation attempt to fail.

To resolve this behavior, a feature called Previous Hop Overwrite has been introduced in Cisco IOS Release 12.4.(20)T. P Hop Overwrite is a mechanism whereby the CE populates the Hop object in the Path message with an IP address from another interface on the router that is reachable in the customer VPN. In this way, the Resv message can find its way back toward the sender and reservations can be established. The P Hop Overwrite mechanism is illustrated in Figure 11-33.

Figure 11-33 RSVP P Hop Overwrite Feature in Cisco IOS 12.4(20)T



Describing Data Flow Characteristics in RSVP (TSpec)

RSVP was designed to support requesting Quality of Service (QoS) for any traffic flow, not just voice or video, across a wide range of Layer 2 technologies. To accomplish this, RSVP must be able to describe in detail the traffic flow for which it is requesting QoS, so that the intermediate routers can make admittance decisions correctly.

The bandwidth requirements for data flows for an RSVP session are characterized by senders in the TSpec (traffic specification) contained in Path messages and are mirrored in the RSpec (reservation specification) sent by receivers in Resv messages. The TSpec gets transported through the network to all intermediary routers and to the destination endpoint. The intermediate routers do not change this object, and the object gets delivered unchanged to the ultimate receiver(s).

The TSpec object contains the following elements:

- AverageBitRate (kbps)
- BurstSize (bytes)
- PeakRate (kbps)

Audio TSpec

For audio flows, the TSpec calculations specify the following measurements:

- AverageBitRate (kbps) — Including IP overhead
- BurstSize (bytes) — This value is calculated as the size of the packet times the number of packets in a burst. For audio flows, the burst usually specifies 1 to 2.
- PeakRate (bytes) — The peak rate, in bytes, refers to the maximum bytes per second that the endpoint transmits at any given time. If the burst is small, as is the case in audio streams, the peak rate can be calculated as 1.1 (or 1.2) times the tokenRate.

To avoid adjusting the bandwidth reservation upward when the call gets answered, Cisco Unified CM reserves the maximum bandwidth for each region codec at call setup time. Unified CM then modifies or adjusts the bandwidth based on the media capability of the connected parties when the call gets answered.

For more information on RSVP for Unified Communications, refer to the *Cisco Unified Communications Manager System Guide*, available at

http://www.cisco.com/en/US/products/sw/voicesw/ps556/prod_maintenance_guides_list.html

Note This section focuses on providing an overview of RSVP principles and mechanisms. For more information on protocol behavior and extensions, complete message formats, and interactions with other protocols, refer to the numerous RFC documents related to RSVP, available at http://www.ietf.org.

RSVP and QoS in WAN Routers

RSVP has been supported in Cisco routers for many years, however most configurations recommended in this document are based on the RSVP Scalability Enhancements feature, which was first introduced in Cisco IOS Release 12.2(2)T.

By issuing the following Cisco IOS command in interface configuration mode on each Cisco IOS router interface, you can enable RSVP and define the maximum amount of bandwidth that it can control:

```
ip rsvp bandwidth [interface-kbps] [single-flow-kbps]
```

The *interface-kbps* parameter specifies the upper limit of bandwidth that RSVP can reserve on the given interface, while the *single-flow-kbps* parameter provides an upper bandwidth limit for each individual reservation (so that flows with higher bandwidth requests will be rejected even if there is bandwidth available on the interface).

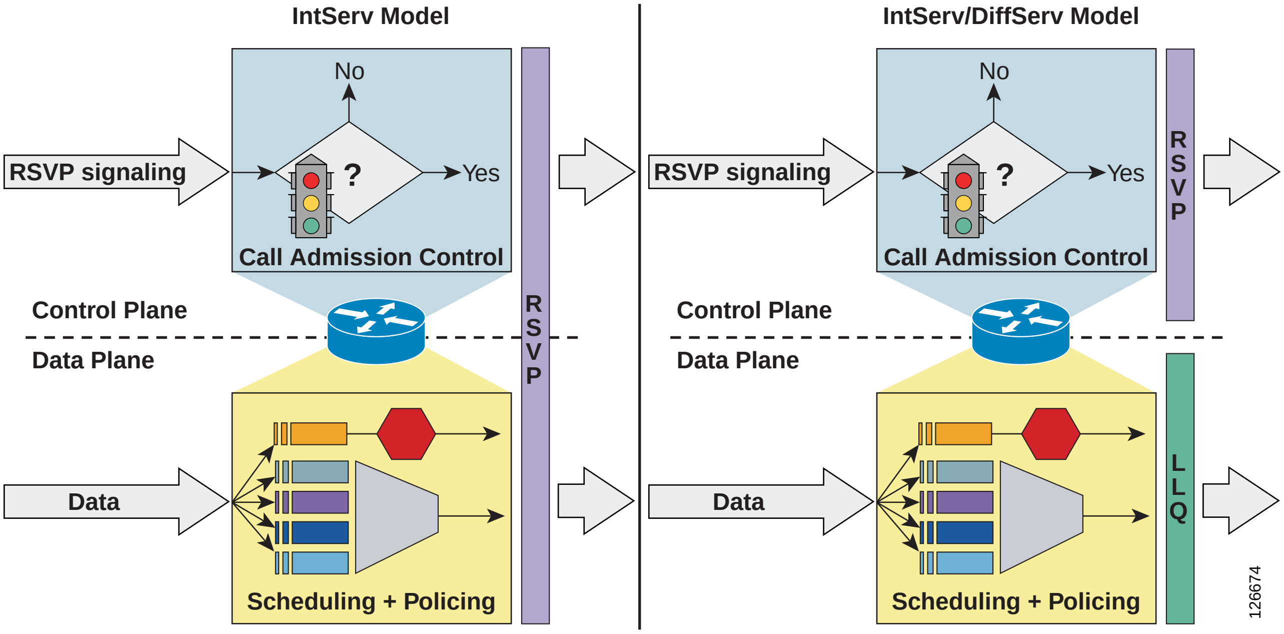
Note When RSVP is enabled on a router interface, all other interfaces in the router will drop RSVP messages unless they are also enabled for RSVP. To avoid dropping RSVP messages, enable RSVP on all interfaces through which you expect RSVP signaling to transit. If call admission control is not desired on an interface, set the bandwidth value to 75% of the interface bandwidth.

Within Cisco IOS, RSVP can be configured to operate according to two different models: the Integrated Services (IntServ) model, described in RFC 2210, or the Integrated Services/Differentiated Services (IntServ/DiffServ) model, described in RFC 2998. Both RFC documents are available on the IETF website at

http://www.ietf.org

Figure 11-34 shows the difference between these two approaches from the perspective of a Cisco IOS router.

Figure 11-34 The Two RSVP Operation Models: IntServ and IntServ/DiffServ

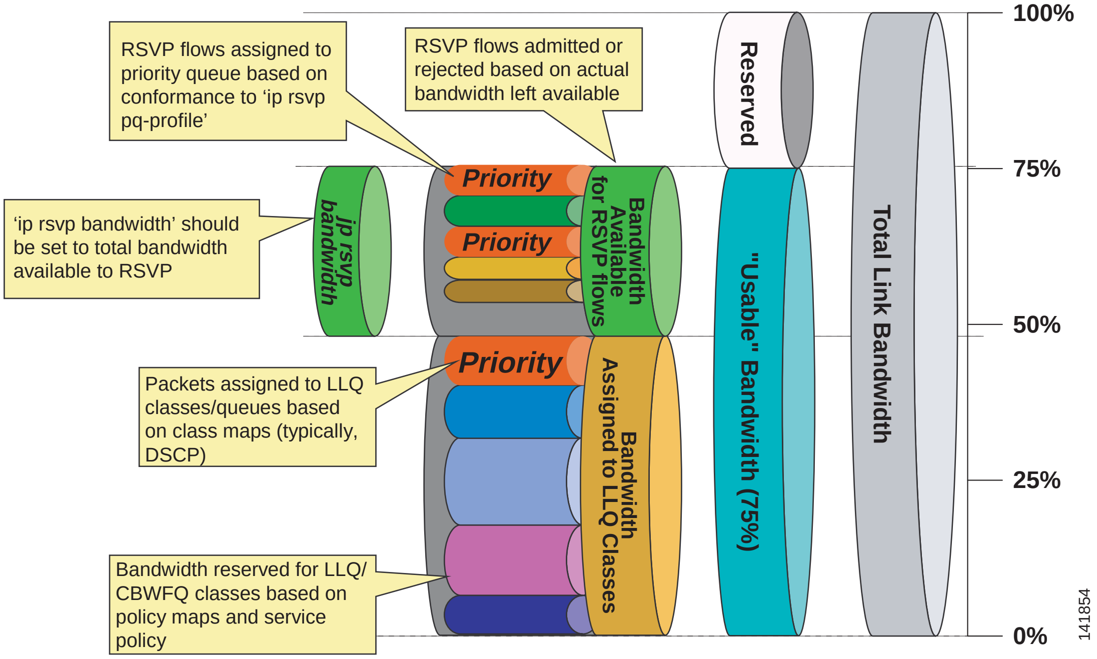


The IntServ Model

As shown on the left side of Figure 11-34, RSVP in the IntServ model involves both the control plane and the data plane. In the control plane, RSVP admits or denies the reservation request. In the data plane, it classifies the data packets, polices them based on the traffic description contained in the RSVP messages, and queues them in the appropriate queue. The classification that RSVP performs is based on the 5-tuple consisting of the source IP address, source port, destination IP address, destination port, and protocol number. In this model, all data packets transiting through the router must be intercepted by RSVP so that RSVP can inspect the 5-tuple and look for a match among the established reservations. If a match is found, the packets are scheduled and policed by RSVP according to the reservation's traffic specification.

As shown in Figure 11-35, when you combine the IntServ model with Low Latency Queuing (LLQ), the usable bandwidth is divided between RSVP and the predefined LLQ queues. RSVP controls the entrance criteria to the RSVP reserved bandwidth, while policy maps control the entrance criteria for the predefined queues.

Figure 11-35 Combining the IntServ Model with LLQ


RSVP flows assigned to priority queue based on conformance to 'ip rsvp pq-profile'
RSVP flows admitted or rejected based on actual bandwidth left available
'ip rsvp bandwidth' should be set to total bandwidth available to RSVP
ip rsvp bandwidth
Priority
Priority
Bandwidth Available for RSVP flows
Packets assigned to LLQ classes/queues based on class maps (typically, DSCP)
Priority
Bandwidth Assigned to LLQ Classes
Bandwidth reserved for LLQ/ CBWFQ classes based on policy maps and service policy
Reserved
"Usable" Bandwidth (75%)
Total Link Bandwidth
100%
75%
50%
25%
0%
141854

To use the IntServ operation model on a Cisco IOS router, use the following commands in interface configuration mode:

```
ip rsvp resource-provider wfq [interface | pvc]
no ip rsvp data-packet classification
```

When these commands are active, RSVP admits or rejects new reservations, not only based on the upper bandwidth limit defined within the **ip rsvp bandwidth** command, but also based on the actual bandwidth resources available. For example, if there are LLQ classes with bandwidth statements, these amounts are deducted from the bandwidth pool that can be assigned to RSVP reservations. While LLQ classes statically allocate bandwidth at configuration time, RSVP does not allocate any amount until a reservation request is received. Therefore, it is important to ensure that an appropriate percentage of the available interface bandwidth is *not* allocated to LLQ classes, so that it can be used by RSVP as reservation requests are received.

Because the total maximum bandwidth that can be assigned to QoS mechanisms on a link is equal to 75% of the link speed, if you want to reserve 33% of the link bandwidth for RSVP-admitted flows, you have to make sure that the bandwidth assigned to LLQ classes does not exceed (75 - 33) = 42% of the link bandwidth.

Because RSVP is in control of assigning packets to the various queues within this model, it is possible to define a mechanism for RSVP to know whether or not to place flows in the Priority Queue (PQ) based on the data flow's T-Spec values by using the following Cisco IOS command in interface configuration mode:

```
ip rsvp pq-profile [r [b [p-to-r]]]
```

Cisco IOS RSVP uses the RSVP TSpec parameters *r*, *b*, and *p-to-r* to determine if the flow being signaled for is a voice flow that needs PQ treatment. These parameters represent the following values:

- *r* = the average traffic rate in bytes per second
- *b* = the maximum burst of a flow in bytes
- *p-to-r* = the ratio of peak rate to average rate, expressed as a percentage

If the traffic characteristics specified by the RSVP TSpec messages for a certain flow are less than or equal to the parameters in the Cisco IOS command, then RSVP will direct the flow into the PQ. If no parameters are provided with the command, the following values, representing the largest of the commonly used voice codecs (G.711), are used as default:

- *r* = 12288 bytes per second
- *b* = 592 bytes
- *p-to-r* = 110%

The IntServ/DiffServ Model

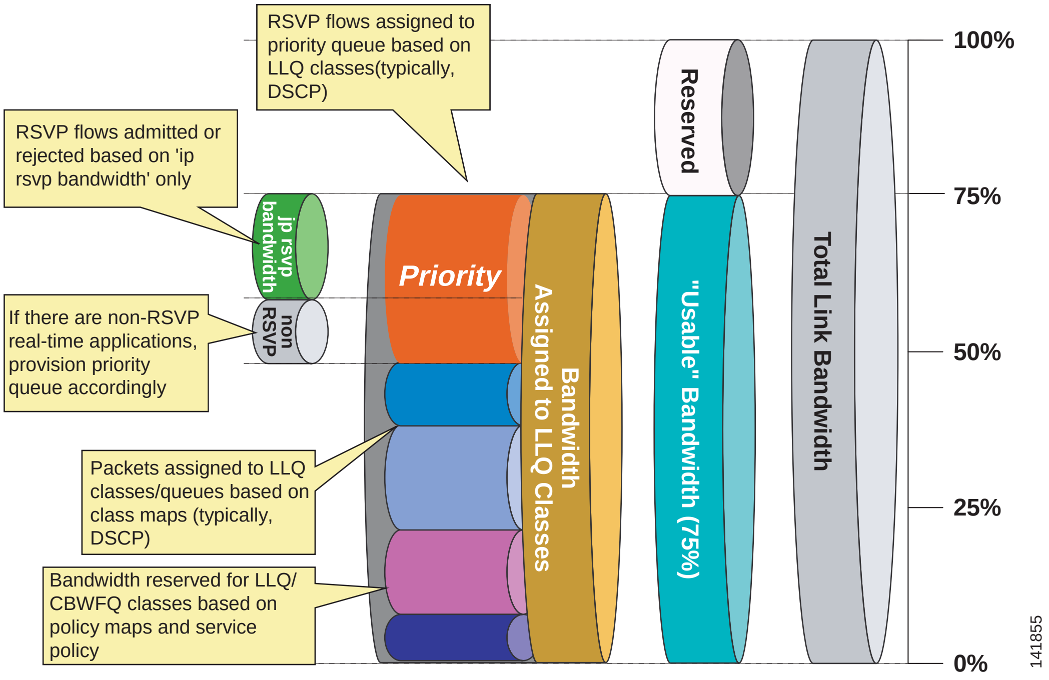
As shown on the right side of Figure 11-34, RSVP in the IntServ/DiffServ model involves only the control plane performing admission control but does not involve the data plane. This means that the call admission control function is separate from the scheduling and policing functions, which can be performed by the Low Latency Queuing (LLQ) algorithm according to predefined class maps, policy maps, and service policies.

With the IntServ/DiffServ model, it is therefore possible to add RSVP call admission control to a network that is already using a Differentiated Services approach to QoS. RSVP admits or rejects calls based on a preconfigured bandwidth amount, but the actual scheduling is based on the pre-existing LLQ criteria such as the DSCP value of each packet.

The entire usable bandwidth (75% of the link speed) can be assigned to LLQ classes, as shown in Figure 11-36, as it normally is today. The policy maps define the traffic that is admitted into each queue. RSVP is typically configured to admit flows up to the amount of bandwidth defined for priority traffic, but keep in mind that RSVP in this model does not adjust the scheduling, so any traffic admitted by RSVP in excess of the predefined priority queue may be dropped or remapped to other lower-priority queues.

If all applications that send priority traffic are RSVP-enabled, you may configure the RSVP bandwidth to match the size of the priority queue. If, on the other hand, there are non-RSVP applications that also need to send priority traffic (such as Unified CM locations-based CAC or a gatekeeper), as shown in Figure 11-36, the priority queue is divided into priority traffic that is controlled by non-RSVP mechanisms and priority traffic that is controlled by RSVP. The combined non-RSVP and RSVP admission control mechanisms must not use more bandwidth than is allocated to ensure that the priority queue is never over-subscribed.

Figure 11-36 LLQ Bandwidth Allocation with RSVP


RSVP flows assigned to priority queue based on LLQ classes(typically, DSCP)
RSVP flows admitted or rejected based on 'ip rsvp bandwidth' only
If there are non-RSVP real-time applications, provision priority queue accordingly
Packets assigned to LLQ classes/queues based on class maps (typically, DSCP)
Bandwidth reserved for LLQ/ CBWFQ classes based on policy maps and service policy
ip rsvp bandwidth
non RSVP
Priority
Bandwidth Assigned to LLQ Classes
Reserved
"Usable" Bandwidth (75%)
Total Link Bandwidth
100%
75%
50%
25%
0%
141855

To use the IntServ/DiffServ operation model on a Cisco IOS router, use the following commands in interface configuration mode:

```
ip rsvp resource-provider none
ip rsvp data-packet classification none
```

When these commands are active, RSVP admits or rejects new reservations uniquely based on the upper bandwidth limits defined within the **ip rsvp bandwidth** command, independently from the actual bandwidth resources available on the interface. Once admitted, the RSVP flows are subject to the same scheduling rules as all other non-RSVP traffic (for example, LLQ class and policy maps). Therefore, it is important to ensure that the RSVP-enabled traffic is marked with the appropriate DSCP value and that the bandwidth of the corresponding PQ or CBWFQ queues is provisioned to accommodate both RSVP-enabled traffic and all other traffic.

In this operating model, the **ip rsvp pq-profile** command is inactive because RSVP does not control the scheduling function.

RSVP Application ID

An application identity (app-id) is an RSVP object that can be inserted into the policy element of an RSVP message. This object is described in RFC 2872. This policy object helps to identify the application and associate it with the RSVP reservation request, thus allowing routers along the path to make appropriate decisions based on the application information.

The need for an app-id arises because RSVP is used to support multiple applications such as voice and video.

Without using an app-id, there is only one bandwidth value that is configurable per interface in RSVP. RSVP will admit requests until this bandwidth limit is reached. It does not differentiate between the requests and is not aware of the type of application for which the bandwidth is requested. As a result of this, it is quite possible for RSVP to exhaust the allowed bandwidth by admitting requests representing just one type of application, thus causing all subsequent requests to be rejected due to unavailable bandwidth. In this way, a few video calls could prevent all or most of the voice calls from being admitted. For example, if an organization allocates 1000 units to RSVP, RSVP might exhaust a majority of this amount by admitting two 384-kbps video calls, thus leaving very little bandwidth for voice calls.

The solution to this problem is to configure separate bandwidth limits for individual applications or classes of traffic. Limiting bandwidth per application requires that an RSVP local policy matching the application bandwidth limit be applied to the router interface and that each reservation request flag the application to which it belongs so that it may be admitted against the appropriate bandwidth limit.

The app-id is not a single piece of information but multiple variable-length strings. As is described in RFC 2872, the object may include the following attributes:

- An identifier of the application (APP). This attribute is required.
- Global unique identifier (GUID). Optional.
- Version number of the application (VER). This attribute is required.
- Sub-application identifier (SAPP). An arbitrary number of sub-application elements can be included. Optional.

For example:

- APP = AudioStream
- GUID = CiscoSystems
- VER = 5.0.1.0
- SAPP = (not specified)

For more information on how Unified CM uses the Application ID, see RSVP Application ID and Unified CM, page 11-72.

Cisco IOS Features

This section describes new Cisco IOS features that apply to the design of deployments using Cisco Unified CM 8.5 and later releases. These features are new in Cisco IOS Release 15.1(3)T, and it is important to use the correct Cisco IOS version to obtain these features.

RSVP Ingress Call Admission Control

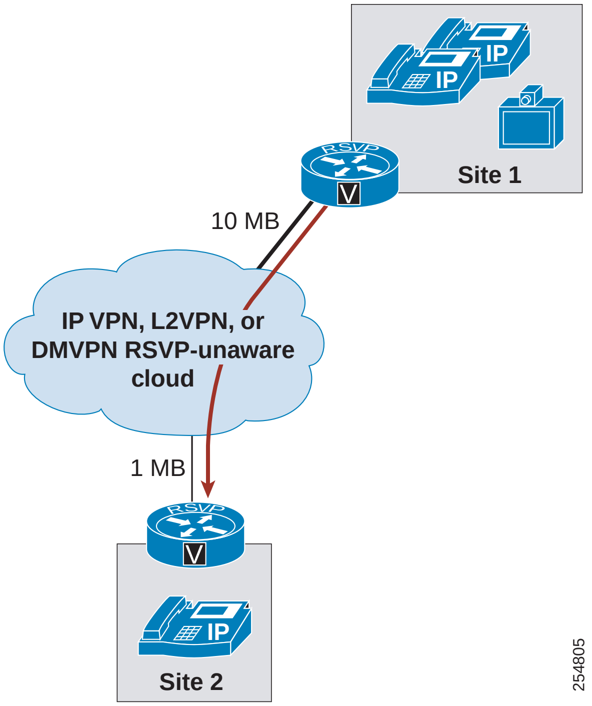
As indicated in the section on RSVP Principles, page 11-42, the RSVP protocol reserves the resources required by the source device to communicate with the destination device. The reservation is unidirectional. The source device signals the path message, which advertises the resource being requested. Upon receiving the path message, the destination device responds with the reservation message. The RESV message traverses the intermediate nodes between the source device and the destination device, hop-by-hop (only RSVP-aware hops), and determines if these nodes can allocate resources for the flow being requested. The reservation is checked against the resources on the outgoing interface only (egress with regard to the direction of the stream) while going downstream in the direction of the destination device.

In the following scenarios, the RESV message is not an indicator for guaranteed communication between the source device and destination device against the resource reserved:

Asymmetric Link Between Two RSVP-Aware Routers

In Figure 11-37, the path message enters the RSVP-unaware cloud through a 10 MB link and goes out of the RSVP-unaware cloud into a 1 MB link. For the stream flowing from Site 1 to Site 2, only the egress interface of the RSVP-aware router at Site 1 is taken into consideration. Downstream, the 1 MB link at Site 2 is not accounted for while making the reservation. In most scenarios this is not an issue because every call has two streams, and a stream in the opposite direction (from Site 2 to Site 1) will reserve the bandwidth on the Site 2 RSVP-aware router.

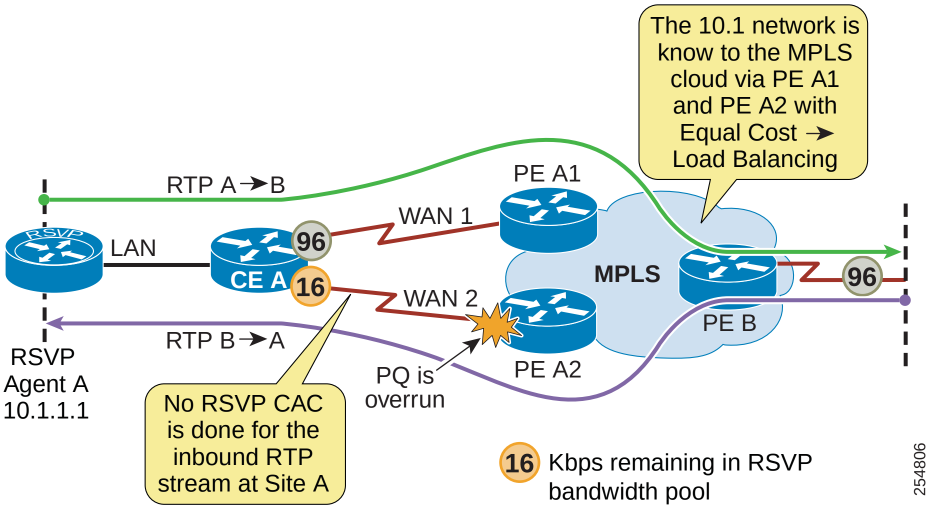
Figure 11-37 Asymmetric Link Between RSVP-Aware Routers



Asymmetric Routing Path Such as a Dual-Attached Customer Equipment (CE) with Load Balancing

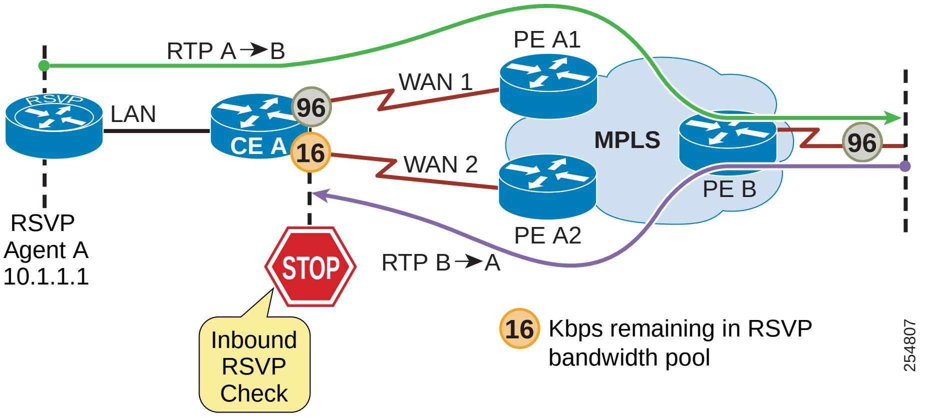
In Figure 11-38, an audio call (two streams, one in each direction) is made between RSVP Agent A and RSVP Agent B. One stream in direction A to B flows over WAN 1, and the other stream in direction B to A flows over WAN 2. This is referred to as an asymmetrically routed call and is common in dual-attached networks with load balancing from a service provider. When RSVP is accounting on only the egress interface into a WAN network that is RSVP unaware, as is the case with service provider clouds for example, there is the potential to overrun the WANs provisioned bandwidth when traffic is load balanced.

Figure 11-38 Asymmetric Routing for a Dual-Attached Customer Equipment (CE) with Load Balancing


The 10.1 network is know to the MPLS cloud via PE A1 and PE A2 with Equal Cost → Load Balancing
RTP A → B
PE A1
WAN 1
RSVP
LAN
96
CE A
16
MPLS
96
WAN 2
PE B
RTP B → A
RSVP Agent A 10.1.1.1
PQ is overrun
PE A2
No RSVP CAC is done for the inbound RTP stream at Site A
16 Kbps remaining in RSVP bandwidth pool
254806

To overcome the limitations posed by the above scenarios and to conform with RFC 2205, use the Ingress Call Admission Control feature, which is supported in Cisco IOS Release 15.1(3)T. Ingress Call Admission Control allows the reservation of an RSVP request to be validated against a bandwidth pool on ingress into the router instead of upon egress only. (See Figure 11-39.) Note that egress bandwidth validation will continue to function as usual.

Figure 11-39 RSVP Ingress Call Admission Control



RSVP VPN Tunnel Support

Dynamic Multipoint Virtual Private Network (DMVPN) allows users to scale large and small IPsec VPNs by combining GRE tunnels, IPsec encryption, and Next Hop Resolution Protocol (NHRP). The RSVP VPN Tunnel feature supports the following types of configurations:

- RSVP over manually configured generic routing encapsulation (GRE) and multipoint generic routing encapsulation (mGRE) tunnels
- RSVP over manually configured GRE and mGRE tunnels in an IPsec protected mode
- RSVP over GRE and mGRE tunnels (IPsec protected and IPsec unprotected) in a DMVPN environment

For more information on DMVPN and the RSVP VPN Tunnel feature, refer to the *Cisco IOS Quality of Service Solutions Configuration Guide, Release 15.1*, available at

http://www.cisco.com/en/US/docs/ios/qos/configuration/guide/15_1/qos_15_1_book.html

RSVP for Flexible Bandwidth Interfaces

As indicated in the section on RSVP Principles, page 11-42, when RSVP bandwidth is configured on an interface, the bandwidth value for that interface is fixed. This causes an issue with flexible interfaces (otherwise known as bundled interfaces) such as Multi-Link PPP, ATM IMA, FRF12, Gigabit EtherChannel (GEC), and so forth. The problem is that, when you configure a static RSVP bandwidth amount on a flexible bandwidth interface that contains bundles of links, if one or more of the links fail and the total bandwidth is reduced, the RSVP bandwidth remains fixed. This means that the ratio between the RSVP bandwidth and the total flexible interface bandwidth is no longer equal to the configured value, and this could cause oversubscription of that flexible bandwidth interface.

Note that Low Latency Queuing (LLQ) already allows for the Priority Queue and Class-Based Weighted Fair Queues to implement percentages; therefore, when applied to flexible bandwidth interfaces, LLQ parameters change in conjunction with the interface on which they are configured.

The Flexible Bandwidth Interfaces feature enhances the **ip rsvp bandwidth** command to allow for the configuration of a percentage of the interface bandwidth. This allows the RSVP bandwidth to change in parallel with the interface bandwidth, and it is applicable to interfaces that consist of a number of physical links that are bundled into one link.

For enterprise customers with sites that leverage bundled WAN interfaces either within their network or to a service provider, this feature allows them to fully maximize the bandwidth utilization for RSVP call admission control during complete up-time while also allowing them to use the same percentage of bandwidth on the bundle dynamically during link failures.

Figure 11-40 and Figure 11-41 illustrate the use of RSVP with flexible bandwidth interfaces.

Figure 11-40 Flexible Bandwidth Interfaces with 10 MB Total Bandwidth

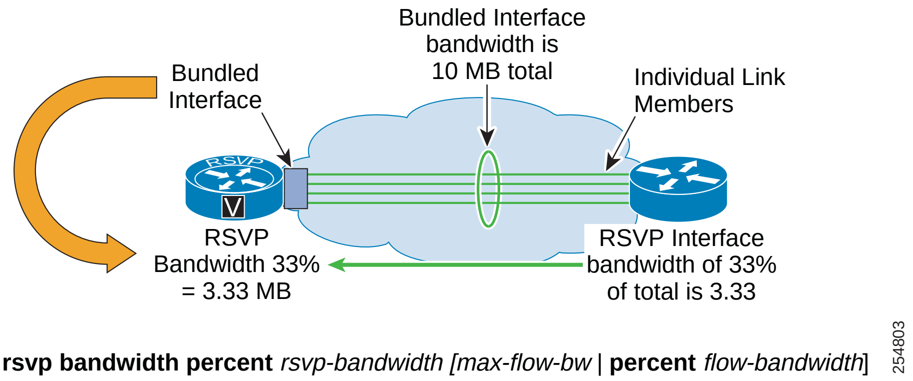
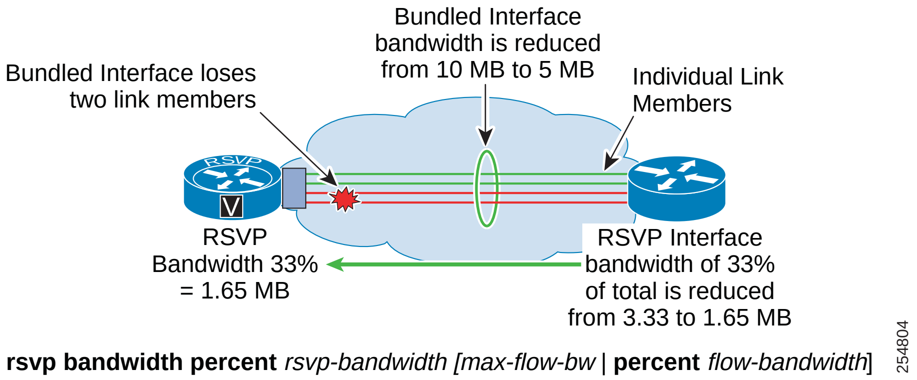

Bundled Interface bandwidth is 10 MB total
Bundled Interface
Individual Link Members
RSVP
V
RSVP Bandwidth 33% = 3.33 MB
RSVP Interface bandwidth of 33% of total is 3.33
ip rsvp bandwidth percent *rsvp-bandwidth [max-flow-bw* | **percent** *flow-bandwidth*]
254803

Figure 11-41 Flexible Bandwidth Interfaces with Total Bandwidth Reduced to 5 MB


Bundled Interface bandwidth is reduced from 10 MB to 5 MB
Bundled Interface loses two link members
Individual Link Members
RSVP
V
RSVP Bandwidth 33% = 1.65 MB
RSVP Interface bandwidth of 33% of total is reduced from 3.33 to 1.65 MB
ip rsvp bandwidth percent *rsvp-bandwidth [max-flow-bw* | **percent** *flow-bandwidth*]
254804

For more information on the use of RSVP with flexible bandwidth interfaces, refer to the *Cisco IOS Quality of Service Solutions Configuration Guide, Release 15.1*, available at

http://www.cisco.com/en/US/docs/ios/qos/configuration/guide/15_1/qos_15_1_book.html

RSVP Design Best Practices

When deploying RSVP in the IP WAN in conjunction with Unified CM, observe the following design best practices:

- Cisco recommends that you use the IntServ/DiffServ model if either of the following statements is true:
  - The only traffic destined for the Priority Queue (PQ) in the IP WAN interfaces is RSVP-enabled traffic.
  - All the non-RSVP traffic destined for the PQ can be deterministically limited to a certain amount by an out-of-band call admission control mechanism (such as Unified CM locations or a Cisco IOS gatekeeper).
- If all the PQ traffic is RSVP-enabled, the value specified in the **ip rsvp bandwidth** command and the **priority** command should match once Layer 2 overhead of the priority queue bandwidth has been taken into account.

- If RSVP is enabled on one or more interfaces of a router, all interfaces through which you expect RSVP signaling to transit should also be enabled for RSVP to ensure that RSVP messages do not get dropped. If call admission control is not desired on an interface, set the bandwidth value to 75% of the interface bandwidth.
- If some PQ traffic is not RSVP-enabled, you must ensure that the sum of the values specified in the **ip rsvp bandwidth** command and in the out-of-band call admission control mechanism do not exceed the bandwidth value specified in the **priority** command.
- Enable RSVP Application ID support if you need to limit the maximum amount of bandwidth used by voice and video calls. Application ID Support is introduced in Cisco IOS Release 12.4(6)T.
- Enable RSVP at the edge of the network, including the router WAN interfaces on both sides of the WAN link.
- Enable RSVP at all possible WAN congestion points, including redundant links of different speeds.
- If you do not have symmetric routing on load-balanced MPLS WAN links, ensure that ingress call admission control is configured (see RSVP Ingress Call Admission Control, page 11-53).
- RSVP is currently not available on most Catalyst Switching Platforms.

Bandwidth Provisioning for RSVP

This section discusses bandwidth provisioning as it relates to RSVP only. For a more general and complete discussion on bandwidth provisioning, see Bandwidth Provisioning, page 3-45.

Calculating RSVP Bandwidth Values for Use with Unified CM

At the time Unified CM instructs the Cisco RSVP Agent to make the initial reservation for the call flow, the endpoints that are involved in the call have not fully exchanged their codec capabilities. Without this information, Unified CM must rely on the region settings to determine how to describe the traffic flow. The size of the traffic flow is a function of two things, the codec bit-rate and the sampling rate (or packets per second). The region settings contain the maximum codec bit rate but do not describe the sampling rate. The preferred sampling rates for audio codecs are defined in the following cluster-wide service parameters:

- Preferred G722 millisecond packet size: 20 ms by default
- Preferred G711 millisecond packet size: 20 ms by default
- Preferred G729 millisecond packet size: 20 ms by default

However, the codec type and codec sampling rate are negotiated for every call and might not be the preferred settings because they are not supported on one or more of the endpoints. To avoid having to increase the reservation size once the capabilities are fully exchanged, possibly causing a post-ring failure, this initial reservation is for the worst-case scenario (the largest codec bit rate using the smallest packet size) for that codec. Once media capabilities have been exchanged between the endpoints, then the reservation is revised to the correct bandwidth allocation. In most cases, the default sampling rate is used, resulting in the reservation being reduced.

Note Unified CM does not include the SRTP overhead or the Layer 2 overhead in the RSVP Reservation. When compared to the RSVP T Spec bandwidth value, the Layer 3 IP RSVP bandwidth statement must take into account any SRTP traffic, and the Layer 2 priority queue value must also be over-provisioned if SRTP traffic is present. (See Table 3-10 and Table 3-11.)

Voice Bearer Traffic

Inter-region call with audio codec maximum set to G729, connecting using G.729:

- Initial request: 40 kbps using a 10 ms worst-case scenario
- Updated request: 24 kbps using the preferred sample size of 20 ms

Inter-region call with audio codec maximum set to G.728/iLBC, connecting using iLBC:

- Initial request: 48 kbps using a G.728 10 ms worst-case scenario
- Updated request: 31.2 kbps using iLBC with a preferred sample size of 20 ms

Inter-region call with audio codec set to G711, connecting using G.711:

- Initial request: 96 kbps using a 10 ms worst-case scenario
- Updated request: 80 kbps using the preferred sample size of 20 ms

Video Bearer Traffic

As with the audio stream, the initial reservation for the video stream will rely on the region settings because the endpoint codec capabilities will not be fully negotiated at the time of the reservation. The region settings for video calls include the bandwidth for the audio stream. (See IP Video Telephony, page 12-1, for more information.) Because the audio stream has its own reservation, the final reservation for the video stream will be the region setting minus the audio codec bit-rate. However, because these codecs have not been fully negotiated, the video stream reservation will be for the worst-case scenario, which assumes no audio stream. Once media capabilities have been exchanged between the endpoints, then the reservation will be revised to the correct bandwidth allocation.

Because video is inherently bursty, it is necessary to add some overhead to the stream requirements. (See Voice Bearer Traffic, page 3-46, for more information.) Unified CM uses the stream bandwidth to determine how to calculate the overhead, as follows:

- If the stream is < 256 kbps, then the overhead will be 20%
- If the stream is >= 256 kbps, then the overhead will be 7%

Inter-region video call, with G.729 audio codec and video setting of 384 kbps:

- Initial request: 384 * 1.07 = 410 kbps
- Updated request: (384 - 8) * 1.07 = 402 kbps

Inter-region video call, with G.711 audio codec and video setting of 384 kbps:

- Initial request: 384 * 1.07 = 410 kbps
- Updated request: (384 - 64) * 1.07 = 342 kbps

Configuration Recommendation

Because the initial reservation will be larger than the actual packet flow, over-provisioning the RSVP and LLQ bandwidth is required to ensure that the desired number of calls can complete.

When provisioning the RSVP bandwidth value for N calls, Cisco recommends that the Nth value be the worst-case bandwidth to ensure that the Nth call gets admitted.

For example:

- To provision four G.729 streams:

  (3 * 24) + 40 = 112 kbps

- To provision four G.711 streams:

  (3 ∗ 80) + 96 = 336 kbps
- To provision four 384 kbps video streams (G.729 audio)

  (3 ∗ (384 - 8) + 384) ∗ 1.07 = 1618 kbps
- To provision four 384 kbps video streams (G.711 audio)

  (3 ∗ (384 - 64) + 384) ∗ 1.07 = 1438 kbps

Configuring Cisco IOS Application ID Support

RSVP Application ID feature support was introduced in Cisco IOS Release 12.4(6)T, and that is the minimum release required for the following examples.

Combined Priority Queue

To utilize the functionality allowed in Unified CM's implementation of Application ID support (that is, allowing voice calls to consume all the bandwidth available in the priority queue), we must modify the previous recommendations that voice and video priority queues be kept separate. (See RSVP Application ID and Unified CM, page 11-72.) To use this functionality, you should combine both the voice and video match criteria into one class-map. Because the requirements are to match either voice traffic or video traffic, be sure to make the class-map match criteria **match-any** instead of **match-all**, as follows:

```
class-map match-any IPC-RTP
 match ip dscp ef
 match ip dscp af41  af42
```

Configure the priority queue to support both the voice and video traffic. The following example configuration allocates 33% of the link bandwidth to the priority queue:

```
policy-map Voice-Policy
 class IPC-RTP
  priority percent 33
```

Mapping Application ID to RSVP Policy Identities

The RSVP Local Policy provides the mechanism for controlling a reservation based on an Application ID. Application IDs are mapped to RSVP Local Policies through the **ip rsvp policy identity** command. RSVP Local Policy identities are defined globally and are available to each interface for policy enforcement. Each identity can have one policy locator defined to match an Application ID.

To give the user as much flexibility as possible in matching application policy locators to local policies, the RSVP local policy command line interface (CLI) accepts application ID match criteria in the form of Unix-style regular expressions for the policy locator. Regular expressions are already used in the CLI for existing Cisco IOS components such as Border Gateway Protocol (BGP). Refer to the follow documentation for more information on how regular expressions are used in Cisco IOS:

- *Access and Communication Servers Command Reference*

  http://www.cisco.com/en/US/docs/ios/11_0/access/command/reference/arbook.html
- *Using Regular Expressions in BGP*

  http://www.cisco.com/en/US/tech/tk365/technologies_tech_note09186a0080094a92.shtml
- *Regex Engine Performance Enhancement*

  http://www.cisco.com/en/US/docs/ios/12_3t/12_3t4/feature/guide/gt_rexpe.html

RSVP Policy Identities for Matching the Default Unified CM Application IDs

```
ip rsvp policy identity rsvp-video policy-locator .*VideoStream.*
ip rsvp policy identity rsvp-voice policy-locator .*AudioStream.*
```

Interface RSVP Local Policies

Whether configuring Application ID support or not, for an interface to support RSVP, you must configure the **ip rsvp bandwidth** <*value*> command on that interface. This value cannot be exceeded by any one RSVP reservation or the sum of RSVP reservations on that interface, regardless of Application ID support. In fact, if a reservation passes the local policy check, it still must pass the interface RSVP bandwidth check before it is reserved.

Local policies based on Application ID are applied to an interface using the **ip rsvp policy local identity** command.

For reservations that match its policy locator value, a local policy has the ability to perform the following functions:

- Define the maximum amount of bandwidth the reservations can reserve as a group or as a single sender
- Forward or not forward RSVP messages
- Accept or not accept RSVP messages
- Define the maximum bandwidth the group or sender can reserve

For example, to limit the amount of video bandwidth to 384 kbps on a Serial T1, use the following commands:

```
interface Serial0/0/1:0
 ip rsvp bandwidth 506
 ip rsvp policy local identity rsvp-video
  maximum bandwidth group 384
  forward all
```

There is also a catch-all local policy called the default local policy. This local policy will match any RSVP reservation that did not match the other RSVP local policies configured on the link. The default local policy can be used to match reservations that are not tagged with an Application ID or reservations that are tagged with an Application ID that you want to treat as untagged traffic.

Example

The following example supports both voice and video calls using the model discussed in How Unified CM Uses the Application ID, page 11-72. The voice calls are guaranteed 352 kbps of bandwidth while video calls are limited to 154 kbps of bandwidth. Voice calls can use all of the available RSVP bandwidth.

```
interface Serial0/0/1:0
 ip address 10.2.101.5 255.255.255.252
 service-policy output Voice-Policy
 ip rsvp bandwidth 506
 ip rsvp data-packet classification none
 ip rsvp resource-provider none
 ip rsvp policy local identity rsvp-voice
  maximum bandwidth group 506
  forward all
 ip rsvp policy local identity rsvp-video
  maximum bandwidth group 154
  forward all
 ip rsvp policy local default
  no accept all ! Will not show in the configuration
  no forward all! Will not show in the configuration
```

In this example, if an RSVP reservation is received that does not have an Application ID or its Application ID does not match the two configured options, the reservation will fail. This configuration works if RSVP traffic originates only from Cisco RSVP Agents controlled by Unified CM. However, if there is intercluster RSVP traffic via an IP-IP gateway or if RSVP messages from a controller other than Unified CM are traversing this link, then the default local policy should be configured to accept and forward the reservations and a maximum bandwidth value should be configured on the policy. Note that it is possible to oversubscribe the RSVP bandwidth via the use of multiple RSVP local policies (if the sum of the policies is greater than the RSVP interface bandwidth), but reservations then become first-come, first-serve.

Provisioning for Call Control Traffic with RSVP and Centralized Call Processing

This section discusses bandwidth provisioning for call control traffic when RSVP is used as the call admission control mechanism in a centralized call processing deployment. For a more general discussion of bandwidth provisioning when RSVP is not used, see .Provisioning for Call Control Traffic with Centralized Call Processing, page 3-49

Considerations for Calls Using RSVP

In systems where call admission control uses RSVP, there is additional SCCP call control traffic between Unified CM and the Cisco RSVP Agents located at the branch when IP calls are placed across the WAN. To compute the associated bandwidth, use the following equation:

Bandwidth (bps) = (21 ∗ CHW) ∗ (Number of IP phones and gateways in the branch)

Where CHW represents the number of calls placed across the IP WAN per hour per phone, including calls between IP phones at different branches as well as calls made through gateways located in a different site. For example, in a site where 20 phones each make 10 calls per hour, if 20% of the calls are placed across the IP WAN, then CHW = 2. The equation thus yields: (21∗2)∗20 = 840 bps.

The bandwidth calculated by this equation should be added to the required bandwidth for phone call control.

Unified CM RSVP-Enabled Locations

Cisco Unified CM provides a topology-aware call admission control mechanism based on the Resource Reservation Protocol (RSVP), which is applicable to any network topology and which eases the restriction of a traditional hub-and-spoke topology. The Cisco RSVP Agent is a Cisco IOS feature that enables Unified CM to perform the RSVP-based call admission control. For information on which Cisco IOS platforms support the Cisco RSVP Agent, refer to the *Cisco RSVP Agent Data Sheet*, available at

http://www.cisco.com/en/US/partner/products/ps6832/products_data_sheets_list.html

The Cisco RSVP Agent registers with Unified CM as either a media termination point (MTP) or a transcoder device with RSVP support. When an endpoint device makes a call in need of a bandwidth reservation, Unified CM invokes a Cisco RSVP Agent to act as a proxy for the endpoint to make the bandwidth reservation.

Figure 11-42 shows the signaling protocols used between Unified CM and various other devices, as well as the associated RTP streams for calls across the WAN in a given location. For any calls across the WAN, Unified CM directs the endpoint devices to send the media streams to their local Cisco RSVP Agent, which originates another call leg synchronized with an RSVP reservation to the Cisco RSVP Agent at the remote location. Figure 11-42 illustrates the following signaling protocols:

- Cisco RSVP Agents register to Unified CM via Skinny Client Control Protocol (SCCP).
- IP phones register with Unified CM via SCCP or Session Initiation Protocol (SIP).
- PSTN gateways register with Unified CM via Media Gateway Control Protocol (MGCP), SIP, or H.323 protocol.

Figure 11-42 Protocol Flows for RSVP-Enabled Locations

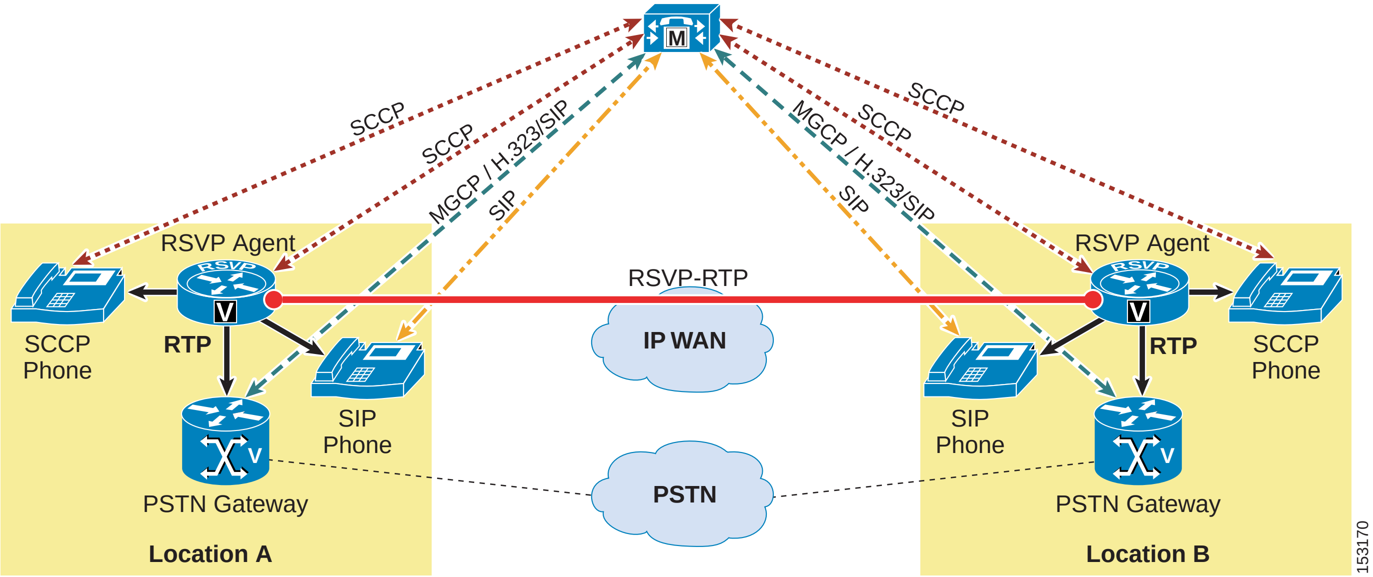
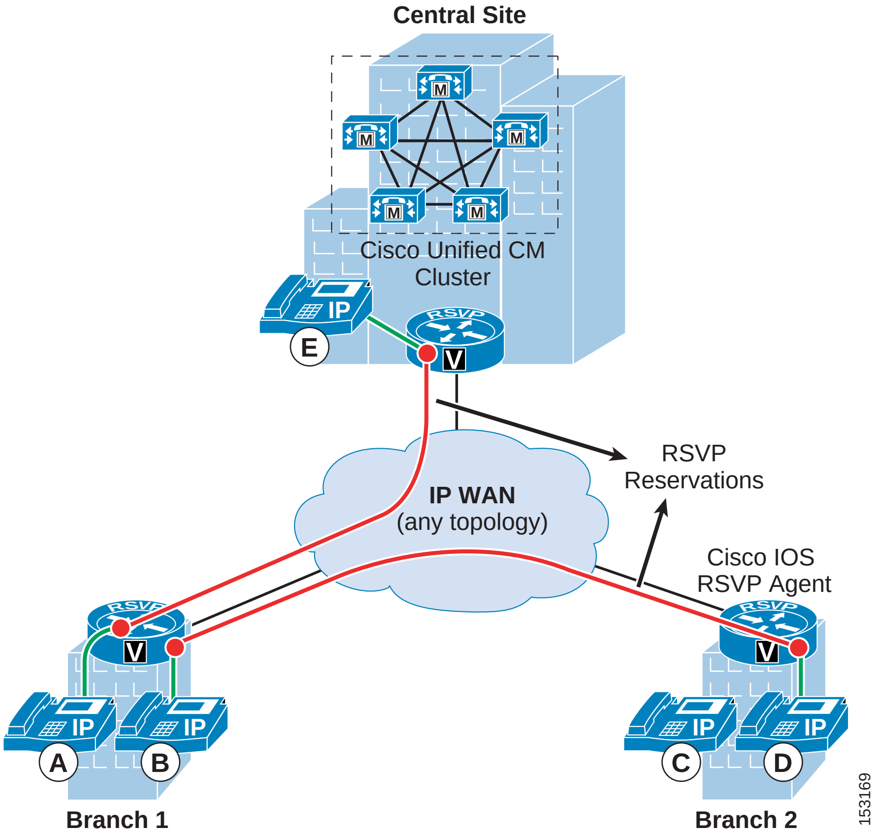


Figure 11-43 shows a typical Cisco RSVP Agent deployment within a Unified CM cluster, which includes three locations: Central Site, Branch 1, and Branch 2. The IP WAN connecting the three locations can be of any topology type and is not restricted to the hub-and-spoke topology. For any call between two locations that requires an RSVP reservation in the media path, a pair of Cisco RSVP Agents is invoked dynamically by Unified CM. The Cisco RSVP Agent acts as a proxy to make an RSVP reservation for the IP phone in the same location with the Cisco RSVP Agent. For example, when phone A in Branch 1 calls phone E in the Central Site, an RSVP reservation (illustrated as the red line in Figure 11-43) is established between Cisco RSVP Agents in the Branch 1 and Central Site locations.

There are three call legs for the media streams of this call. The first call leg is between phone A and the Branch 1 Cisco RSVP Agent, the second call leg is between the Branch 1 and Central Site Cisco RSVP Agents, and the third call leg is between the Central Site Cisco RSVP Agent and phone E. By the same token, when phone B in Branch 1 calls phone D in Branch 2, the RSVP reservation is established between the Branch 1 and Branch 2 Cisco RSVP Agents. Note that the media streams of a call between two branch locations are not sent through the Central Site in this case, which is different from a call made over the traditional hub-and-spoke topology.

Note While RSVP-enabled locations and the use of Cisco RSVP Agent introduce support for arbitrary WAN topologies, they are based on static assignment of devices to a location, which means that every time a device is moved from one physical site to another, its configuration in Unified CM needs to be updated. Device Mobility can be used to update site-specific device configuration information automatically when the device moves to a new physical site. For more information, see the section on Device Mobility, page 25-14.

Figure 11-43 Cisco RSVP Agent Concept



Cisco RSVP Agent Provisioning

The capacity of Cisco RSVP Agent in terms of simultaneous calls (also referred to as sessions) depends on the following factors:

- For software-based MTP functionality, the session capacity is determined by the router platform and the relative CPU load. (Refer to the *Cisco RSVP Agent Data Sheet*, available at http://www.cisco.com/en/US/products/ps6832/products_data_sheets_list.html.)
- For hardware-based MTP and transcoder functionality, the session capacity is limited by the number of DSPs available. (See Media Resources, page 17-1, for DSP sizing considerations.)

For more information on supported platforms, requirements, and capacities, refer to the *Cisco RSVP Agent Data Sheet*, available at:

http://www.cisco.com/en/US/products/ps6832/products_data_sheets_list.html

For software-based MTP functionality, the *Cisco RSVP Agent Data Sheet* provides guidelines for session capacity based on a router dedicated to the Cisco RSVP Agent and 75% CPU utilization. These numbers apply to specific Cisco IOS releases and should be considered as broad guidelines. Different combinations of specific services, configurations, traffic patterns, network topologies, routing tables, and other factors can significantly affect actual performance for a specific deployment and hence reduce the number of concurrent sessions supported. Cisco recommends careful planning and validation testing prior to deploying a multi-service router in a production environment.

Cisco RSVP Agent Registration

The Cisco RSVP Agent registers with Unified CM as an MTP or transcoder device with RSVP support. The Cisco RSVP Agent does not have transcoding capability when registering as an MTP device. To have transcoding capability, the Cisco RSVP Agent must register with Unified CM as a transcoder device.

Registration Switchover and Switchback

If the primary Unified CM fails, the Cisco RSVP Agent switches over to the secondary Unified CM. When the primary Unified CM recovers from the failure, the Cisco RSVP Agent switches its registration back to the primary Unified CM. Use the following commands to configure the Cisco RSVP Agent registration switchover and switchback:

```
sccp ccm group
 switchover method immediate
 switchback method guard timeout 7200
!
gateway
  timer receive-rtp 180
```

- The **switchover method immediate** command specifies the immediate registration switchover to the secondary Unified CM server after failure of the primary Unified CM server is detected. The available DSP resources become available immediately for new calls after the switchover has completed.
- The **switchback method guard timeout 7200** command specifies the registration switchback mechanism after the primary Unified CM recovers from its failure. With this command configured, the Cisco RSVP Agent starts to switch its registration gracefully back to the primary Unified CM after the last active call disconnects. If the graceful registration switchback has not initiated by the time the guard timer expires, the Cisco RSVP Agent will use the immediate switchback mechanism and register with the primary Unified CM right away. The default value of the guard timer is 7200 seconds, and it can be configured statically in the range of 60 to 172800 seconds.
- The **timer receiver-rtp** command in the gateway configuration mode defines the RTP clean-up timer for RSVP reservations. If a failure occurs, the RSVP reservation for the existing call will stay in place until the RTP clean-up timer expires. The default value of this timer is 1200 seconds. Cisco recommends that you configure this timer with its lowest allowed value, which is 180 seconds.

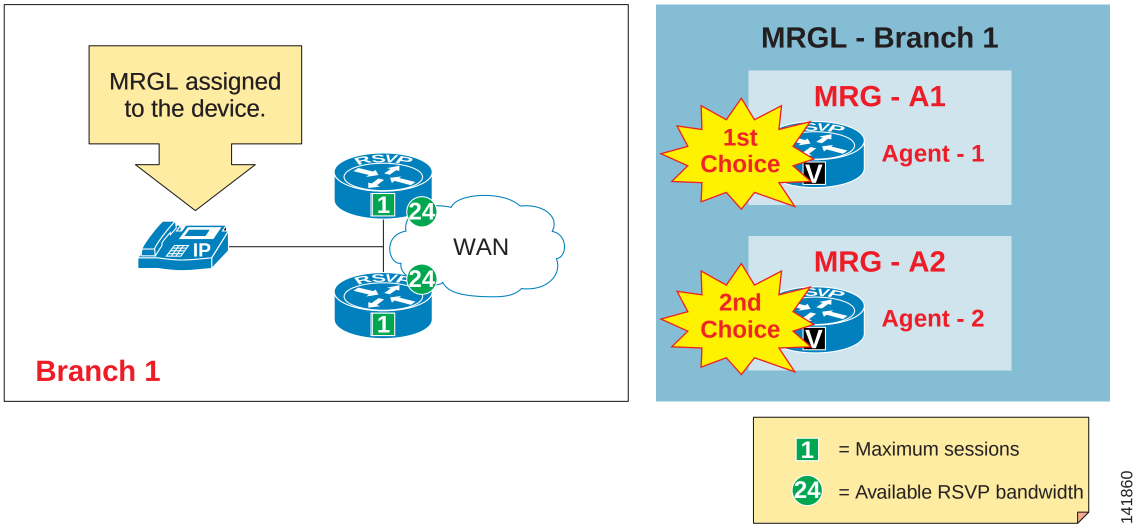
Maximum Sessions Support

The Cisco RSVP Agent supports a maximum number of calls or sessions, based on the software-based (CPU) and hardware-based (DSP) resources equipped on the Cisco RSVP Agent router. The **maximum sessions** command in the **dspfarm profile** configuration mode specifies the maximum number of calls that the Cisco RSVP Agent is able to handle. The Cisco RSVP Agent notifies Unified CM of its session capacity based on this configuration. The maximum number of sessions decreases by one for every call going through the Cisco RSVP Agent. When the counter reaches zero, the Cisco RSVP Agent is regarded as having no resources available, and Unified CM skips that Cisco RSVP Agent for any subsequent calls.

Figure 11-44 shows a branch site with dual Cisco RSVP Agents. The Cisco RSVP Agents are co-resident with the WAN routers, and Cisco RSVP Agent redundancy is achieved by assigning two Cisco RSVP Agents to different MRGs in the same MRGL. If Agent-1 in MRG-1 is unavailable or out of session capacity, Unified CM will try to allocate Agent-2 in MRG-2 for RSVP calls to or from Branch 1. To ensure that Agent-2 is selected when Agent-1's capacity is reached, Cisco recommends configuring the maximum number of sessions to match exactly the number of calls supported by the **ip rsvp bandwidth** configured on the WAN interface of the Cisco RSVP Agent. In this example, both Cisco RSVP Agents need to be configured with **maximum sessions 1**. This recommendation is made on the assumption that all calls going across the WAN will use the same type of codec so that an accurate calculation of the number of calls across the WAN can be derived, which is done by dividing the total available RSVP bandwidth by the bandwidth requested per call.

Note If the maximum number of sessions is higher than the number of calls supported by the **ip rsvp bandwidth** configuration, Unified CM will still send the call to the Cisco RSVP Agent but the RSVP reservation will fail because there is no bandwidth available, and Unified CM will follow the usual behavior for call admission control failure (that is, it will deny the call or invoke the AAR feature).

Figure 11-44 Configuring Maximum Sessions on the Cisco RSVP Agent



Pass-Through Codec

The pass-through codec enables a Cisco IOS Enhanced MTP device to terminate an RTP media stream received from an endpoint without knowing the media encoding of the stream. That is, the UDP packets of the media stream flow through the MTP without being decoded. This method enables the MTP to support every audio, video, and data codec that is defined in Unified CM. Because the MTP does not decode the media stream, the pass-through codec can also be used with encrypted (SRTP) media streams.

In fact, for video and SRTP media streams to use an MTP, it must support the pass-through codec. When configured with the pass-through codec, the Cisco RSVP Agent will substitute its own IP address for the source IP address in the IP/UDP headers of the packets and let them flow through.

The Cisco RSVP Agent will use the pass-through codec only if all of the following conditions are met:

- The two endpoint devices involved in the call have matching audio codec capability, and the region configuration permits the matching codec to be used for the call. In other words, no transcoder device needs to be inserted in the call.
- **MTP Required** is not configured for either endpoint device.
- All intermediate resource devices support the pass-through codec.

Note If the Cisco RSVP Agent registers as an MTP device and a transcoder device needs to be inserted in the call, the codec configured in the Cisco RSVP Agent dspfarm MTP profile must match the inter-region bit rate configured in Unified CM Administration. For example, if 8 kbps (G.729) is the inter-region bit rate configured in Unified CM Administration, then the G.729 codec must also be configured in the dspfarm MTP profile.

The following example shows an Cisco RSVP Agent configuration on a Cisco IOS 2900 Series platform:

```
interface Loopback0
 ip address 10.11.1.100 255.255.255.255
!
sccp local Loopback0
sccp ccm 20.11.1.50 identifier 1 priority 1 version 7.0+
sccp ccm 20.11.1.51 identifier 2 priority 2 version 7.0+
sccp
!
sccp ccm group 1
 associate ccm 1 priority 1
 associate ccm 2 priority 2
 associate profile 1 register RSVPAgent
 switchover method immediate
 switchback method guard timeout 7200
!
dspfarm profile 1 mtp
 codec pass-through
 codec g729ar8
 rsvp
 maximum sessions software 100
 associate application SCCP
```

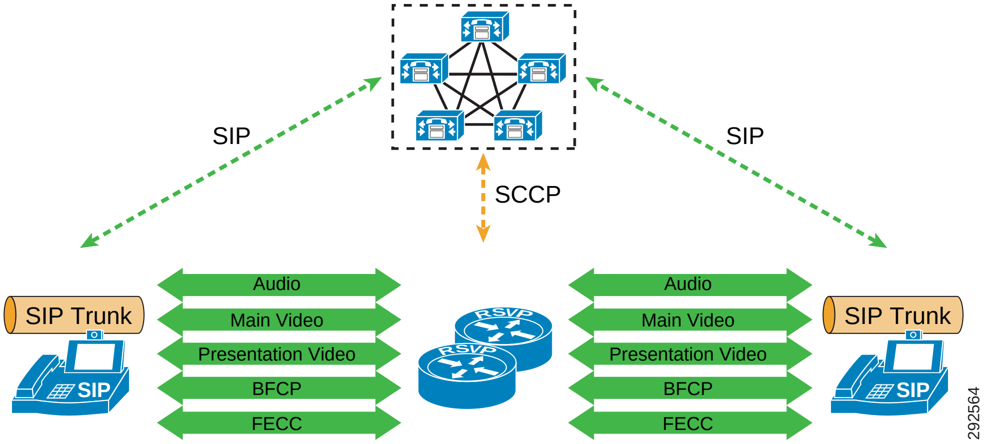
RSVP Agent Support for RTCP, BFCP and FECC Negotiation

As mentioned, RSVP Agent supports a pass-through codec that does not decode the media but, as the name implies, passes the media through yet terminating and re-originating Layer 3 headers. This allows RSVP Agent to support any codec defined or used in Unified CM. With Cisco Unified CM 9.x and Cisco IOS Release 15.2.1T and later releases, Unified CM supports RTCP, BFCP, and FECC negotiation and pass-through as described in the following sections.

BFCP and FECC Pass-Through

RSVP Agent (Cisco IOS MTP and transcoder) and Unified CM support Binary Flow Control Protocol (BFCP) and Far End Camera Control (FECC) pass-through. Previously this was not possible due to lack of media port support by the RSVP Agent that was limited to three media ports. With more media port support, BFCP and FECC negotiation now works end-to-end through the RSVP Agent. Figure 11-45 illustrates BFCP and FECC support in the RSVP Agent.

Figure 11-45 RSVP Agent Support for BFCP and FECC Pass-Through



When Unified CM negotiates a video call with presentation sharing using BFCP and FECC, the RSVP Agent passes through the BFCP control channel, the FECC channel if negotiated, and the a second video channel associated to the presentation sharing controlled through BFCP. However, RSVP Agent reserves the bandwidth only for the bit rate of the main video channel. Endpoints using BFCP down-speed the main video to allow the presentation video, so that the main video and presentation video do not use more bandwidth than requested for the main video. If there is no main video channel, then Unified CM reserves bandwidth only for the presentation video negotiated through BFCP. The latter scenario is quite rare; typically there will be a main video channel negotiated, and thus the bandwidth reservation made by RSVP Agent is associated to that.

When BFCP and FECC are negotiated with RSVP Agent, Unified CM requests ports from the RSVP Agent as they are available. This port request has a priority order to it that is hard-coded in the Unified CM request logic. The order is as follows:

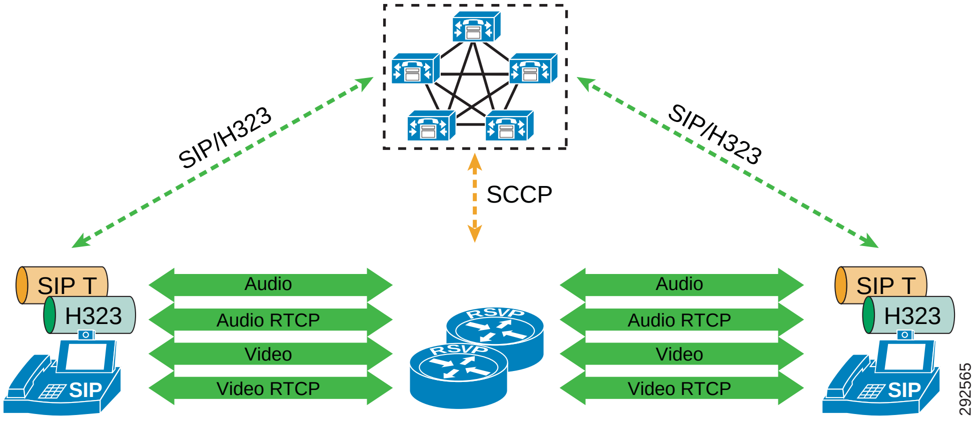
1. Audio
2. Main video
3. BFCP control channel
4. Second video or presentation video channel
5. FECC channel

An example of where this priority comes into play is in scenarios where the RSVP Agent can provide fewer ports than requested, in which case certain functions will be excluded from the negotiation. An example of this is if the RSVP Agent can provide only four ports for a video call request with presentation sharing and FECC. In this case FECC is last in the priority, so it will not get a channel because five channels are requested but only four are available.

RTCP Pass-Through

In the same way that RSVP Agent (Cisco IOS MTP and transcoder) supports BFCP and FECC, it also supports RTCP pass-through. RTCP is a highly utilized protocol negotiated between endpoints, and it can be critical for higher definition video calls to ensure audio and video synchronization. Figure 11-46 illustrates a video call with RTCP pass-through.

Figure 11-46 Video Call Using RTCP Pass-Through



When Unified CM negotiates a video call over a trunk or endpoint and both RTCP and RSVP are enabled, the RSVP Agent opens a new RTCP channel for each media stream. Figure 11-46 illustrates a video call where both audio and video have independent RTCP channels.

For more information on function, design, and deployment of Cisco IOS media resources (RSVP Agent, MTP, and transcoder) with BFCP, FECC, and RTCP support, see the chapter on Media Resources, page 17-1.

RSVP Policy

Unified CM can apply different RSVP policies to different location pairs. The RSVP policy can be configured in Unified CM Administration. The RSVP policy defines whether or not Unified CM will admit the call if the RSVP reservation attempt fails. The following RSVP policy settings can be configured between any two locations:

- No Reservation

  No RSVP reservation attempt is made and, if enabled, Enhanced Locations call admission control is performed by Unified CM.

- Mandatory

  Unified CM does not ring the terminating endpoint device until the RSVP reservation succeeds for the audio stream and, if the call is a video call, for the video stream as well.

- Mandatory (Video Desired)

  A video call can proceed as an audio-only call if a reservation for the video stream cannot be reserved but the reservation for the audio stream succeeds.

- Optional (Video Desired)

  A call can proceed as a best-effort audio-only call if it fails to obtain reservations for both its audio and video streams. The Cisco RSVP Agent re-marks the media packets as best-effort.

- Use System Default

  The RSVP policy for the location pair matches the clusterwide RSVP policy. The default clusterwide RSVP policy is No Reservation. To change the default RSVP policy in Unified CM Administration, select **System** > **Service Parameters** > **Cisco CallManager Service**.> **Default Inter-location RSVP Policy**

Note With the Optional (video desired) policy, IP WAN calls may proceed as best-effort not only if the RSVP reservation fails but also if the Cisco RSVP Agent is not available. In this case, Unified CM instructs SCCP and MGCP devices to re-mark their traffic as best-effort. However, this re-marking is not possible with H.323 and SIP devices, which will keep sending their traffic with the default QoS marking. To prevent over-subscribing the priority queue in the latter case, Cisco recommends configuring an access control list (ACL) on the IP WAN router to permit only packets marked with DSCP EF or AF41 if the source IP address is that of the Cisco RSVP Agent.

Cisco recommends configuring the RSVP policy as **Mandatory** or **Mandatory (Video Desired)** because those settings guarantee the bandwidth reservation and the voice quality of the call. The most efficient method for setting the clusterwide RSVP policy is to configure the **Default Inter-location RSVP Policy** in the RSVP clusterwide parameters of the Cisco CallManager Service Service Parameter Configuration, and leave the RSVP configuration in the location configuration set to **Use System Default**.

In the clusterwide RSVP parameters configuration, there is a service parameter named **Mandatory RSVP mid call error handle option**. If the RSVP policy is configured as **Mandatory** or **Mandatory (Video Desired)**, this parameter specifies how Unified CM will treat an existing RSVP call based on the failure of a mid-call RSVP reservation attempt. The mid-call RSVP reservation attempt can be triggered by (but is not limited to) a network convergence after a WAN failure or by an existing voice-only call becoming a video call. A network convergence makes the Cisco RSVP Agent not only start to send the media streams over the newly converged path but also to try to make a new RSVP reservation over the new path.

The default setting of the **Mandatory RSVP mid call error handle option** is **Call Becomes Best Effort**. With the default option configured, Unified CM will maintain the existing call even though the mid-call RSVP reservation attempt fails, but the RTP streams will be marked as best effort (DSCP 0). Cisco recommends configuring this parameter with the **Call Fails Following Retry Counter Exceeded** option. With this option configured, Unified CM will fail the call if the RSVP reservation attempt keeps failing after a certain number of retries. The default value of the retry counter is 1, which is defined by the **RSVP Mandatory mid-call retry counter** service parameter, and the default value of **RSVP retry timer** is 60 seconds. Cisco recommends having both the retry counter and the retry timer service parameters configured with their default values. With both set to their default values, Unified CM will wait for 60 seconds before it disconnects the call if the RSVP mid-call retry fails. During this period, users might experience degraded voice quality because no RSVP reservation is in place and the RTP streams are marked as best effort.

Migrating to RSVP Call Admission Control

To migrate to RSVP-based call admission control, Cisco recommends configuring and deploying RSVP in the network, configuring and deploying RSVP Agents in the branch locations and in Unified CM, and when all RSVP configurations are complete, using the Unified CM clusterwide RSVP service parameter **Default inter-location RSVP Policy** to switch all locations directly over to RSVP CAC. This method allows the administrator to completely deploy RSVP in both the network infrastructure and the Unified Communications infrastructure while continuing to use Enhanced Locations CAC until the switch is ready to be made. It also allows the administrator to easily switch back to Enhanced Locations CAC in the event of a misconfiguration.

Note that Unified CM first checks if RSVP is enabled and then checks locations and links through the LBM. This simultaneous functioning of CAC mechanisms allows for an easier migration and the ability to revert back to Enhanced Location CAC in the event that there is a misconfiguration.

The following is a short list of events that occur for an intra-cluster call when both RSVP and Enhanced Locations CAC are enabled:

1. Unified CM first checks the location pair policy or the clusterwide **Default inter-location RSVP Policy** of the locations of the devices in the call. If RSVP is enabled between the locations, Unified CM allocates RSVP Agents from the MRGL of each device in the call and makes an RSVP reservation request.
2. When RSVP Agent returns the reservation request result, Unified CM checks to see if Enhanced Locations CAC is enabled (LBM is active). If it is, Unified CM requests the bandwidth from LBM over the effective path (end-to-end location and link path).
3. LBM returns the results of the path request and, if the request is successful, allows the call.
4. If LBM is not enabled or available, Unified CM checks the Cisco CallManager service parameter **Call Treatment When No LBM Available**. If this parameter is set to allow the call, then the call will complete; if it is set to reject the call, then the call will fail.

Observe the following recommendations when planning an RSVP migration:

- Deploy RSVP in the WAN network infrastructure. See the section on RSVP and QoS in WAN Routers, page 11-48.
- Set up Cisco RSVP Agent in each branch location and assign each RSVP Agent to the MRG and MRGL associated to the IP phones and devices in each applicable branch. Ensure that the phones and devices in the branch use the RSVP Agent that is located in the same local branch.
- Ensure that each branch location RSVP setting between any pair of locations is configured with **Use System Default**.
- Ensure that the Cisco CallManager service parameter **Call Treatment When No LBM Available** is set to **allow call**.
- Once RSVP configuration and deployment is completed, change the Cisco CallManager clusterwide RSVP service parameter of **Default inter-location RSVP Policy** from **No Reservation** to the desired RSVP policy, such as **Mandatory** or **Mandatory (Video Desired)**.
- After ensuring that RSVP is being engaged for calls between locations, you can disable the Cisco Locations Bandwidth Manager (LBM).
- If inter-cluster RSVP support is required, enable RSVP over SIP Preconditions as outlined in the section on Migration from Enhanced Locations Call Admission Control to RSVP SIP Preconditions, page 11-78.
- If for any reason there is a need to return to Enhanced Locations CAC, enable LBM services on the Unified CM servers running the Cisco CallManager service and change the Cisco CallManager clusterwide RSVP service parameter **Default inter-location RSVP Policy** to **No Reservation**.

RSVP Application ID and Unified CM

The RSVP Application ID is a mechanism that enables Unified CM to add an identifier to both the voice and video traffic so that the Cisco RSVP Agent can set a separate bandwidth limit on either traffic based on the identifier received. To deploy the RSVP Application ID in the network, you must use a minimum version of Cisco IOS Release 12.4(6)T or higher on the Cisco RSVP Agent router. The RSVP Application ID strings can be configured via two service parameters in the clusterwide RSVP parameter configuration: **RSVP Audio Application ID** and **RSVP Video Application ID**.

Unified CM uses SCCP to convey the RSVP Application ID to the Cisco RSVP Agent. The Cisco RSVP Agent also inserts the RSVP Application ID into the RSVP signaling messages (such as the RSVP PATH and RESV messages) and sends those messages to the downstream or upstream RSVP routers.

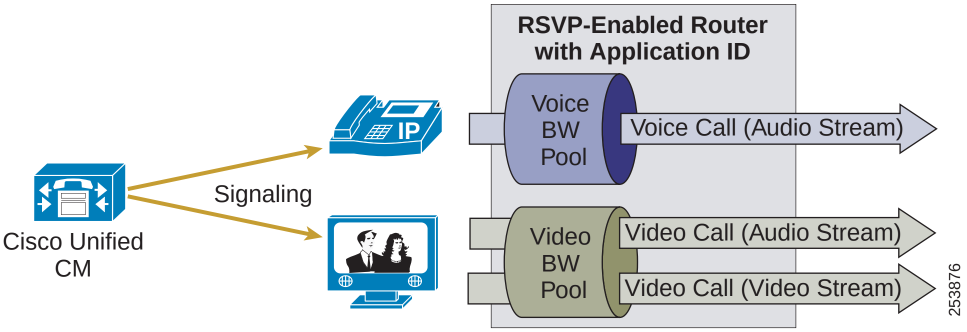
How Unified CM Uses the Application ID

Unified CM has two cluster-wide service parameters that define the Application ID used to tag audio and video call reservations using RSVP:

- RSVP Audio Application ID (Default = AudioStream)
- RSVP Video Application ID (Default = VideoStream)

Figure 11-47 shows how Unified CM tags voice and video calls with an Application ID in RSVP.

Figure 11-47 Unified CM and RSVP Application ID


RSVP-Enabled Router with Application ID
Voice BW Pool
Voice Call (Audio Stream)
Signaling
Cisco Unified CM
IP
Video BW Pool
Video Call (Audio Stream)
Video Call (Video Stream)
253876

How Voice Calls Are Tagged

When a voice call is made between locations with an RSVP policy, the resulting reservations for the audio stream will be tagged with the RSVP Audio Application ID.

How Video Calls Are Tagged

When a video call is made between locations with an RSVP policy, the resulting reservations for the audio and video streams will both be tagged with the RSVP Video Application ID. A video call has both audio and video associated to the Video Application ID.

RSVP Application ID Design Considerations and Best Practices

- The AudioStream Application ID is used for audio streams of audio-only calls.
- The VideoStream Application ID is used for both the audio and video streams of a video call.

- The Application ID does not currently differentiate between various types of video, such as telepresence video versus other video. All video in an RSVP session will be marked with the Video Application ID and the Video DSCP value.
- Unified CM currently has separate settings for both the Application ID and the DSCP values of the signaling and media streams. These are managed separately; however, Cisco recommends using the default values because they are configured to work in conjunction with one another.
- When video escalation occurs, the RSVP reservation for the audio stream is readmitted with the Video Application ID and configured DSCP value (PHB of AF41, by default). If the readmission for the audio stream fails due to insufficient bandwidth, the audio stream will continue as best-effort with a Video Application ID until the reservation into the Video Application ID pool succeeds.
- When video de-escalation occurs, the RSVP reservation for the audio stream is readmitted with the Audio Application ID and configured DSCP value (PHB of EF, by default). If the readmission for the audio stream fails due to insufficient bandwidth, the audio stream will continue as best-effort with an Audio Application ID until the reservation into the Audio Application ID pool succeeds.

Video Escalation Example with Application ID

An audio-only call is set up with the AudioStream Application ID, and the DSCP for the stream is set to a PHB value of EF. When the call is escalated video, the video streams are set up with the VideoStream Application ID. If the video stream reservation fails, the call will stay as an audio-only call with the AudioStream Application ID. However, if the video stream reservation succeeds, the audio stream will be readmitted from AudioStream Application ID to VideoStream Application ID. If the readmission succeeds, then both the video and audio streams will have the VideoStream Application ID set to a PHB value of AF41. If the readmission fails, then the video stream will have the VideoStream Application ID with a PHB value of AF41 while audio stream will have the VideoStream Application ID with a PHB set to 0 (video fail DSCP value).

See Unified CM Video Calls with RSVP SIP Preconditions, page 11-80, for information on video escalation and de-escalation in distributed Unified CM environments with RSVP SIP Preconditions.

RSVP SIP Preconditions

RSVP SIP Preconditions is based on SIP Preconditions as defined in RFC 3312 and RFC 4032, and it offers the ability for Cisco call processing products to negotiate a level of Quality of Service and perform call admission control using the RSVP protocol. The term RSVP SIP Preconditions is used to identify the passing of policy information elements, or preconditions, over SIP signaling to negotiate a Quality of Service (QoS) policy. The actual RSVP message is not signaled over the SIP trunk; only the policy-related information elements are. The RSVP messages are then carried over by the RSVP Agent or RSVP-capable router. This use of SIP preconditions extends the negotiation of RSVP Quality of Service policy across Unified CM clusters as well as to Unified CM Express and SIP-TDM Cisco IOS gateways to allow for synchronization of the RSVP layer and call control layer between these various call control applications.

Overview of SIP Preconditions

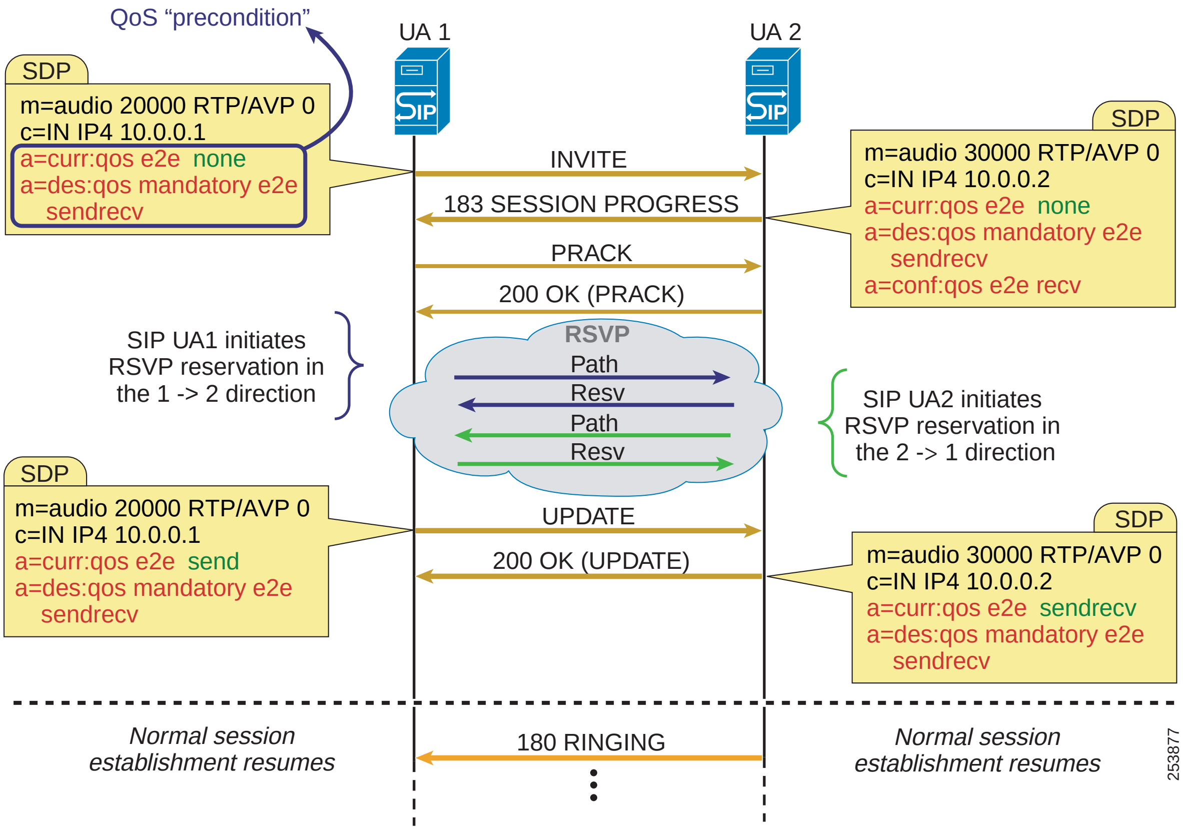
As mentioned, SIP Preconditions provide for the negotiation of RSVP policy information across call control applications, thus allowing synchronization between these call control applications of the RSVP Layer for resource reservation and the call control layer for call setup and establishment.

The concept of a precondition in SIP signaling also avoids the potential for what is referred to as "ghost rings" across independent call control applications. Ghost rings can occur at session establishment time if the called party is rung without having first reserved the resources necessary to establish the media between the callers. In order to minimize ghost rings, network resources for the session must be reserved before the called party is alerted. However, the reservation of network resources frequently requires learning the IP address, port, and session parameters from the called party. This information is obtained as a result of the initial offer and answer exchange carried in SIP. This exchange normally causes the phone to ring, thus introducing a loop dilemma: resources cannot be reserved without performing an initial offer and answer exchange, but the initial offer and answer exchange cannot be done without performing resource reservation.

RSVP SIP Preconditions solves this dilemma by setting SIP Preconditions or constraints about the session that are introduced in the offer. The recipient of the offer generates an answer but does not alert the user or otherwise proceed with session establishment. Proceeding occurs only when the preconditions are met. This knowledge can be obtained through a local event, such as a confirmation of a resource reservation, or through a new offer sent by the calling party.

Figure 11-48 illustrates how these SIP Preconditions work in a generic SIP signaling call flow.

Figure 11-48 SIP with RSVP Between Call Agents



In Figure 11-48, a SIP user agent (SIP UA 1) starts the call by sending a SIP Invite message. The preconditions are in the SIP Invite in the Session Description Protocol (SDP), where the calling party's IP address and port number are identified. The preconditions stipulate a current QoS policy (a=curr:qos) and a desired QoS policy (a=des:qos). In this example, SIP UA 1 sends an invite to SIP UA 2 with a current QoS policy for the audio line set to **none** and the desired QoS policy is set to **mandatory e2e sendrecv**. This tells the receiver that an RSVP reservation is mandatory before offering the call (ringing the end device). The SIP UA 2 receiving the Invite then responds with a 183 session progress message with SDP stipulating a response to the preconditions sent. In this example, SIP UA 2 sends a current QoS policy as **none**, a desired QoS policy of **mandatory e2e sendrecv**, and a configured QoS policy (a=conf:qos) of **e2e recv**, indicating that it has received the request and will initiate a reservation using RSVP. At this point both user agents negotiate RSVP to reserve bandwidth for the media as described in the SDP. If this reservation succeeds, the UAs update one another with the updated QoS policy preconditions and then proceed with the call by ringing the end user. In the example, SIP UA 2 then responds with a 180 Ringing message and the call can continue with normal establishment. If the reservation fails, then either SIP UA can terminate the call prior to ringing the called party. This avoids any "ghost ringing" condition.

Unified Communications Manager and RSVP SIP Preconditions

RSVP SIP Preconditions for Unified CM provides the functionality of intercluster call admission control in distributed Unified CM deployments. If you deploy RSVP SIP Preconditions in Unified CM, Cisco recommends having local RSVP-enabled locations-based call admission control fully functional prior to enabling RSVP SIP Preconditions. This approach is also recommended for migration purposes. For further details on enabling intra-cluster RSVP call admission control, see Unified CM RSVP-Enabled Locations, page 11-62.

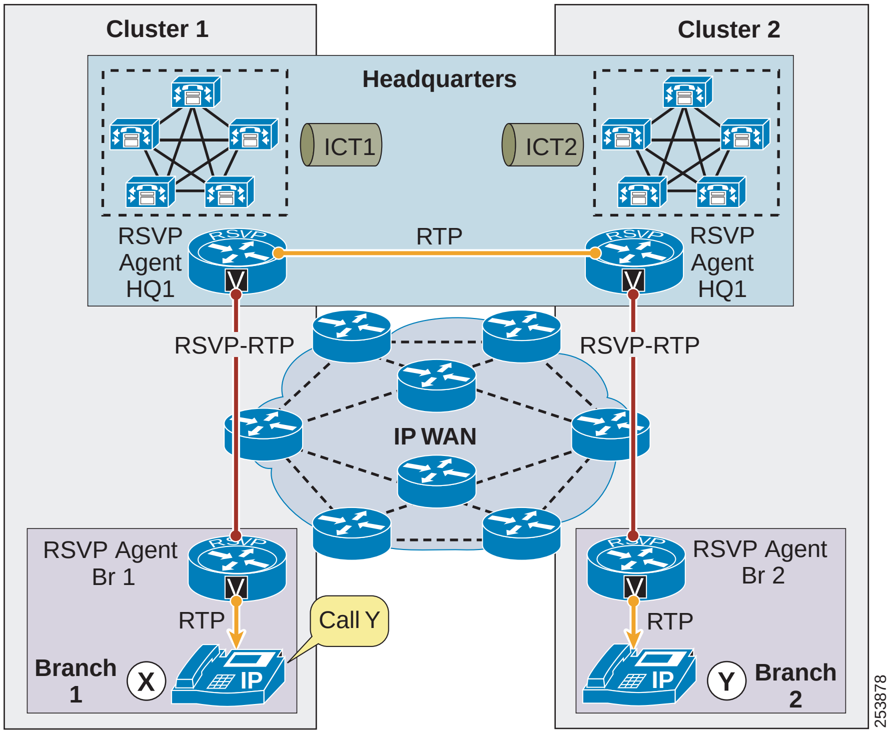
RSVP SIP Preconditions has two modes of configuration, local RSVP and end-to-end RSVP. These modes are configured on the SIP Trunk Profile in Unified CM administration pages.

Local RSVP

Local RSVP does not support reservations between two RSVP agents that are located in separate clusters. It uses two RSVP agents per cluster and does not use RSVP across the trunk that connects the clusters. This is the default configuration of the SIP Trunk Profile.

Figure 11-49 illustrates local RSVP in a distributed Unified CM deployment.

Figure 11-49 Local RSVP in a Distributed Unified CM Deployment



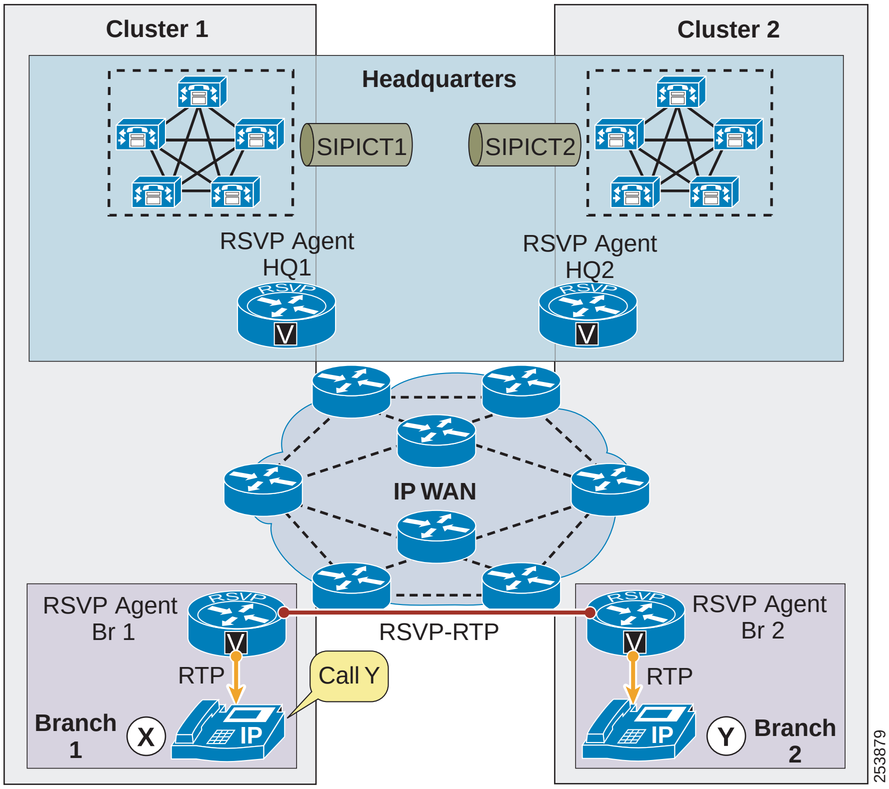
In Figure 11-49, X indicates an endpoint in cluster 1, Y indicates an endpoint in cluster 2, and ICT1 and ICT2 indicate the intercluster trunks configured in clusters 1 and 2 respectively. The RSVP agents associated with the respective devices are also indicated. In this scenario, Cisco Unified CM cluster 1 controls the reservation between AgentBr1 and AgentHQ1, and Cisco Unified CM cluster 2 controls the reservation between AgentBr2 and AgentHQ2.

End-to-End RSVP

End-to-end RSVP configuration is available if the clusters are connected by a SIP trunk. End-to-end RSVP uses RSVP on the entire connection between the RSVP agents, and it uses only one RSVP agent per cluster.

Figure 11-50 illustrates end-to-end RSVP in Unified CM.

Figure 11-50 End-to-End RSVP



In Figure 11-50, X indicates an endpoint in cluster 1, Y indicates an endpoint in cluster 2, and ICT1 and ICT2 indicate the intercluster trunks configured in clusters 1 and 2 respectively. The RSVP agents associated with the respective devices are also indicated. In this scenario, Cisco Unified CM establishes an end-to-end RSVP connection directly between AgentBr1 and AgentBr2.

RSVP SIP Preconditions and Fallback to Local RSVP

Unified CM can be configured to fall back from end-to-end RSVP to local RSVP by configuring **Fall back to local RSVP** on the SIP Trunk profile. This fallback occurs only when the terminating side of the SIP trunk returns a SIP 420 response (Bad Extension), which indicates that the terminating side does not understand the preconditions. Fallback does not occur when a response such as a SIP 580 response (Precondition Failed) is returned. In the case where an end-to-end RSVP SIP Preconditions failure occurs with a SIP 420 (Bad Extension) response during call establishment, Unified CM will invoke local RSVP. If this behavior is desired, a media resource group list with an RSVP Agent association must be assigned to the SIP intercluster trunk. If fallback to local RSVP is not configured, then Unified CM will continue down the route group and route list to another configured trunk or gateway (if configured), otherwise the call will fail.

This feature could be used in designs where a single SIP trunk could terminate to multiple destinations, where both SIP preconditions are supported and where they are not supported. An example might be with the Unified Proxy Server, where there is a single SIP trunk that is configured to a SIP proxy and the returned destination could be a terminating cluster that understands the SIP preconditions or a terminating cluster or SIP server that does not understand the SIP preconditions. Because there is only a single SIP trunk in this case, it would be enabled for RSVP SIP preconditions with fallback enabled. In cases where the terminating side does not understand the SIP preconditions, an RSVP agent can be associated to the SIP trunk in fallback mode so that, when a SIP 420 message (Bad Extension) is received and fallback occurs, a new SIP Invite will go out without the SIP preconditions. In cases where SIP preconditions are supported, the call will continue as explained in the Overview of SIP Preconditions, page 11-73.

Cisco does *not* recommend enabling local RSVP fallback. Instead, a different route should be configured to reach the destination. Cisco recommends using a function such as Local Route Group or a similar function to reroute calls that fail RSVP call admission control to a gateway that is local to the calling device in order to extend the call over the PSTN.

Migration from Enhanced Locations Call Admission Control to RSVP SIP Preconditions

When migrating from Enhanced Locations call admission control to RSVP SIP Preconditions, it is important to first follow the migration recommendations in the section on Migrating to RSVP Call Admission Control, page 11-70. Once migration of locations-based call admission control to local RSVP call admission control is complete, RSVP SIP Preconditions can be enabled on the SIP intercluster trunk.

The following steps are required to enable RSVP SIP Preconditions:

Step 1 Configure a SIP intercluster trunk in each cluster, and direct it to the other cluster.

Step 2 Place the SIP intercluster trunks into their own location. All devices must be in a separate location from the SIP intercluster trunk location and have an inter-location RSVP policy of **Mandatory** or **Mandatory (Video Desired)**. The inter-location policy determines the RSVP policy that is sent over the SIP trunk in the preconditions. (See Table 11-7, which lists the Unified CM inter-location policy that corresponds to the equivalent SIP audio and video precondition attributes.)

Step 3 Configure the intra-location RSVP policy of the SIP intercluster trunk to **Mandatory** or **Mandatory (Video Desired)**. Intra-location RSVP policy is accomplished by setting an inter-location RSVP policy of the specified location to itself. This is necessary for calls that are transferred back to the cluster on the same SIP intercluster trunk so that transfer does not fail.

Step 4 Configure the SIP profile of the SIP intercluster trunks on each Unified CM cluster by setting **RSVP Over SIP** to **E2E**, the **Fall back to local RSVP** field to your preference, and the **SIP Rel1XX Options** to **Send PRACK if 1XX contains SDP**.

Note For the SIP trunk configuration, IPv6 is not supported on RSVP SIP Preconditions. Therefore, ensure that the IPv6 enablement checkbox **Enable ANAT for early offer calls** is not checked because it is not supported with RSVP SIP Preconditions.

Note Unified CM ignores the **MTP Required** and **Use TRP** check boxes on the SIP trunk when it is configured for end-to-end RSVP.

As mentioned, the RSVP SIP Preconditions feature allows Unified CM endpoints to establish direct RSVP agent-to-agent reservations across clusters. Figure 11-51 shows the components involved in a call made with RSVP SIP Preconditions.

Figure 11-51 RSVP SIP Preconditions, Distributed Unified CM Deployment Dual Cluster Design

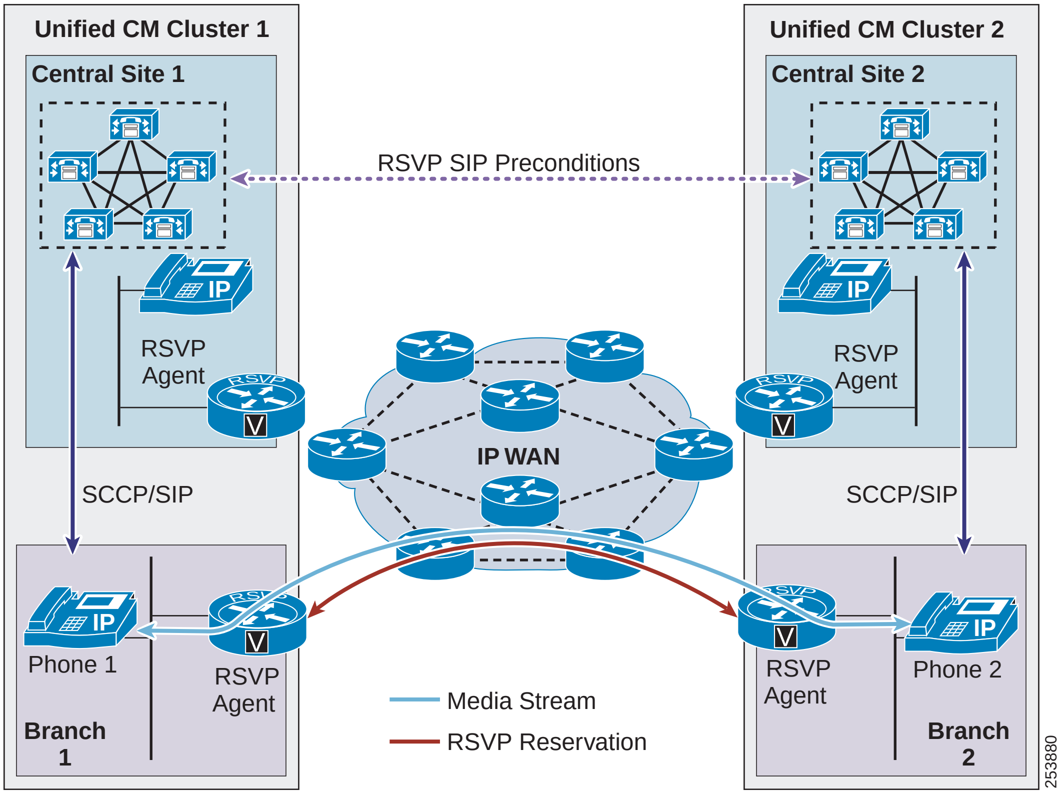


Figure 11-51 illustrates a typical dual cluster deployment with RSVP SIP Preconditions enabled. It includes four locations: Central Site 1, Branch 1, Central Site 2, and Branch 2. The IP WAN connecting the locations can be of any topology type and is not restricted to the hub-and-spoke topology. For any call between two clusters that requires an RSVP reservation in the media path, a Cisco RSVP Agent is invoked dynamically by each Unified CM cluster. The Cisco RSVP Agent acts as a proxy to make an RSVP reservation for the IP phone in the same location with the Cisco RSVP Agent. For example, when phone 1 in Branch 1 calls phone 2 in Branch 2, an RSVP reservation (illustrated as the red line in Figure 11-51) is established between Cisco RSVP Agents in the Branch 1 and Branch 2 locations. This is similar to the media stream setup of a single cluster RSVP-enabled locations solution. The difference here is that the SIP trunk is passing the RSVP policy negotiation between the two Unified CM clusters so that only a single RSVP Agent is allocated per cluster location associated with the respective phones.

There are three call legs for the media streams of this call. The first call leg is between phone 1 and the Branch 1 Cisco RSVP Agent, the second call leg is between the Branch 1 and Branch 2 Cisco RSVP Agents, and the third call leg is between the Branch 2 Cisco RSVP Agent and phone 2. Note that the media streams of a call between two branch locations are not sent through the central site in this case, which is different from a call made over the traditional hub-and-spoke topology using call admission control based on static locations.

There are five call legs for the signaling of this same call. The first call leg is between phone 1 and Unified CM Cluster 1; the second leg is between the Branch 1 Cisco RSVP Agent and Unified CM Cluster 1; the third call leg is between Unified CM Cluster 1 and Unified CM Cluster 2; the fourth is between Unified CM Cluster 2 and the Branch 2 Cisco RSVP Agent; and the fifth and last call leg is between Unified CM Cluster 2 and phone 2.

In Figure 11-51, Phone 1 in Cluster 1 Branch 1 calls a Phone 2 in Cluster 2 Branch 2. The call signaling between the phones and Unified CM could be SCCP or SIP, and the signaling between Unified CMs will be SIP with the RSVP SIP Preconditions feature enabled. When Phone 1 initiates a call to Phone 2, the Cluster 1 server allocates an RSVP Agent to Phone 1 based on the RSVP Agent located in Phone 1's media resource group and list, and it then extends the call over the SIP trunk to Cluster 2 with SIP preconditions (RSVP Policy). The preconditions that are advertised in the SIP INVITE to Cluster 2 are a derivative of the inter-location policy configured between the locations of Phone 1 and the SIP trunk in Cluster 1. So in this case, on Cluster 1 the inter-location policy between locations Branch 1 and HQ is set to Mandatory (Video Desired). For details about Unified CM policy, see RSVP Policy, page 11-69. This inter-location policy determines the policy set on the outbound SIP INVITE to Cluster 2. At this point, Cluster 2 receives a SIP INVITE from Cluster 1 with preconditions set to Mandatory. Cluster 2 then allocates an RSVP Agent to Phone 2 based on its media resource group and list, and also checks the configured locations between the SIP trunk on Cluster 2 and Phone 2 in Branch 2. If this policy is also Mandatory, then Cluster 2 responds with a 183 SESSION PROGRESS message (followed by a PRACK) and starts the RSVP negotiation between the two RSVP Agents in Branch 1 and Branch 2. Once the RSVP Agents have successfully negotiated a reservation for the call, they will inform their respective clusters and the SIP signaling will progress through to ringing stage.

Table 11-7 compares the Unified CM inter-location policy to the equivalent SIP audio and video precondition attribute. (For details about Unified CM RSVP Policy, see RSVP Policy, page 11-69.)

Table 11-7 Unified CM RSVP Policy and Equivalent SIP Preconditions

Unified CM RSVP Policy	SIP Precondition (Audio Call)	SIP Precondition (Video Call)

No Reservation	audio = none	audio = none
		video = none

Optional (Video desired)	audio = optional	audio = optional
		video = optional

Mandatory	audio = mandatory	audio = mandatory
		video = mandatory

Mandatory (Video desired)	audio = mandatory	audio = mandatory
		video = optional

Unified CM Video Calls with RSVP SIP Preconditions

Unified CM supports video escalation and de-escalation across Unified CM clusters with RSVP SIP Preconditions. Video escalation occurs when an ongoing audio-only call is escalated to video or when a video stream is added to the audio-only call. Conversely, de-escalation is the downgrading or de-escalating of a video call to an audio-only call.

In order to support video escalation and de-escalation across clusters with RSVP SIP Preconditions, Unified CM signals two media lines (or m-lines) in the SIP Preconditions within the SIP Session Description Protocol (SDP), one for the audio stream and one for the video stream. Having separate media lines for both audio and video allows each stream to have its own RSVP policy and status in SIP signaling. Because the audio and video streams have their own precondition attributes Unified CM, RSVP policies can be mapped easily into the preconditions. This function allows Unified CM to pass the successful status of an audio stream reservation while simultaneously passing the failed status of the video stream reservation, the potential of a Mandatory (Video Desired) policy, thus allowing the call to be downgraded from a video call to an audio-only call, without rejecting the call entirely.

The video bandwidth reserved for RSVP SIP Preconditions is set to the value configured between the region pair. In this case that would be the region of the endpoint and the SIP intercluster trunk region. Video bandwidth is then adjusted after the video channel is established. Cisco recommends ensuring a video bandwidth value that is greater than or equal to the expected negotiated bit rate of any two endpoints between the region pairs.

For mid-call video escalation, the video stream will be set up only after having video bandwidth reserved.

During hold/resume of a video call, video and audio bandwidth will continue to be reserved while connecting to music on hold.

For other supplementary services such as transfer, Unified CM triggers video reservation and video stream setup after the audio stream completes setup (this is also known as delayed video escalation).

Example 11-1 Delayed Video Escalation: Call transfer from audio-only to video call with RSVP SIP Preconditions

Video device A in cluster A calls audio device B in cluster B through a SIP trunk with RSVP SIP Preconditions enabled. The call is set up as an audio call whose audio streams are allocated to the audio pool with AudioStream Application ID and PHB (Per Hop Behavior) of EF.

Device B transfers the call to a video device C in cluster B. Audio streams between A and C are first established in the audio pool with AudioStream Application ID and PHB of EF.

Delayed video escalation happens between A and C only after the audio media connection is successful. The video streams are allocated to the video pool with VideoStream Application ID. If the video stream allocation fails, the call will stay as an audio-only call with AudioStream Application ID and PHB of EF. If the video stream reservation succeeds, the audio stream will be re-admitted from the audio pool to the video pool with the VideoStream Application ID. If the re-admission succeeds, then video and audio streams will have the VideoStream Application ID with a PHB of AF41. However, if the re-admission fails, then the video stream will have the VideoStream Application ID with a PHB of AF41 while the audio stream will have the VideoStream Application ID with PHB of 0 (video fail best effort value).

Unified CM and RSVP SIP Preconditions Best Practices and Design Considerations:

- The SIP trunk should always have both an inter-location and an intra-location RSVP policy. The inter-location policy ensures that the correct RSVP policy is set for inbound and outbound calls. The intra-location policy ensures that calls hair-pinned on the same trunk (due to intercluster forward and transfer operations) are ensured an end-to-end RSVP policy.
- Cisco recommends configuring a **Mandatory** or **Mandatory (Video Desired)** RSVP policy because those settings guarantee the bandwidth reservation and the voice quality of the call.
- Cisco recommends configuring the SIP trunk profile with the **SIP Rel1XX Options** field set to **Send PRACK if 1XX contains SDP**. A SIP PRACK message is required for RSVP SIP Preconditions operation, but only for 1XX messages containing SDP.
- Ensure the configuration of each cluster in an RSVP SIP Preconditions deployment is standardized across the solution so that the RSVP cluster service parameters, inter-location policies, and codecs used across the WAN as well as on the RSVP Agent are the same. It is important to ensure there is no mismatch in capabilities or configuration across the clusters when call establishment is being attempted.

- Unified CM with RSVP SIP Preconditions supports termination to shared lines across clusters, subject to the following guidelines and restrictions:
  - When setting up a call across clusters to a shared line, the RSVP reservation occurs between the calling device's RSVP Agent and the first RSVP Agent allocated for the shared line device. (This is not controllable by programming.) All other devices with this shared line in separate locations will only allocate an RSVP Agent and not establish a reservation.
  - One RSVP Agent is allocated to each location where one or more devices of the shared line exist.
  - If the device that had the first RSVP Agent allocated is the device that answers the call, then the call establishment will take place and the RSVP Agents that were allocated to other shared line devices in other locations will be released.
  - If a device that did not have a reservation established answers the call, then a new reservation will be initiated between the calling devices RSVP Agent and the one allocated for the answering device with an optional RSVP policy, and the RSVP Agents will continue to try the reservation for the duration of the call until a reservation is successful. During the time that the call is under an optional policy and the **Mandatory RSVP mid call error handle option** is set to **Call becomes best effort** (default), then the media stream between the two devices will be marked best-effort until a reservation succeeds, in which case the media will be re-marked to a PHB (Per Hop Behavior) value of EF (audio) or AF41 (video).
  - If the device that had the first RSVP Agent allocated fails the RSVP reservation with a Mandatory policy, then neither that device nor any device in that location will be rung. However, the shared line devices in all other location will be rung.
- Based on the above shared line limitations, Cisco recommends restricting a shared line to a group of devices within the same location.
- Unified CM with RSVP SIP Preconditions supports termination to Mobile Connect destinations (remote destinations), subject to the following guidelines and restrictions:
  - Local RSVP: For remote destinations to devices, gateways, or trunks that are remote to the calling device, gateway, or trunk, apply the same rules as explained above in the shared line support.
  - End-to-end RSVP: Remote destinations for any single line should not point to more than one RSVP SIP Preconditions destination. Unified CM supports only one RSVP SIP Preconditions call per line for remote destinations.
- If the MoH server is in the same location as the holding party (the party placing another party on hold), the initial reservation is reused and no new reservation is made.
- Hold/Resume functionality with RSVP SIP Preconditions will break the media streams across endpoints and RSVP agents, but the reservation will still be preserved.
- sRTP is supported with RSVP SIP Preconditions and is negotiated during media setup and after RSVP reservation. Unified CM does not signal RTP/SAVP and crypto attributes during the precondition phase.

- T.38 is supported with RSVP SIP Preconditions and negotiated from SIP, H.323, and MGCP endpoints supporting T.38 fax transmission. Unified CM will negotiate an initial reservation using the inter-region audio bandwidth (between the endpoint and SIP intercluster trunk). After call establishment and upon T.38 switchover, the bandwidth usage will be readjusted to 80 kbps if it is not already set.
  - Limitation: If the inter-region bit rate is set to less than 80 kbps, then after T.38 switchover occurs, the RSVP reservation will be readjusted to 80 kbps. This can cause failure if the new adjusted bandwidth cannot be reserved. In such cases, if the reservation fails after switchover, the call will continue because Unified CM will not signal this failure over the SIP intercluster trunk.
  - For the above reason, when deploying T.38 fax with RSVP SIP Preconditions, Cisco recommends using 80 kbps as the inter-region audio bit rate between T.38 endpoints and the intercluster trunk enabled for RSVP SIP Preconditions.
- To support RSVP SIP Preconditions for supplementary services such as hold/resume, transfer, and conference, media resources such as the music on hold servers, annunciators, and conference bridges must have a local RSVP agent assigned to their respective device pools' media resource group list (MRGL).

Note In various call flows for supplementary services such as hold/resume or transfer and conference, different media resources are brought into the RSVP SIP Preconditions call. Those media resources such as conference bridges, music on hold servers, and annunciators also require an RSVP Agent association, just as any other device would when invoked into a RSVP SIP Preconditions or RSVP-enabled locations call. These media resources obtain an RSVP resource from the media resource group list associated to the configured device pool.

Architecture and Considerations for Extension Mobility Cross Cluster

Extension Mobility Cross Cluster (EMCC) enables users in one cluster to log into IP phones of another cluster. For more detailed information about Extension Mobility Cross Cluster with regards to feature function, high availability, and scalability, see Extension Mobility Cross Cluster (EMCC), page 19-10. The rest of this section covers the EMCC feature as it applies to call admission control. It also assumes an understanding of the information covered in Extension Mobility Cross Cluster (EMCC), page 19-10.

EMCC and RSVP-Enabled Environments

In Unified CM RSVP-enabled deployments with either RSVP-enabled locations (for single cluster or intra-cluster) or RSVP SIP Preconditions (for distributed clusters or inter-cluster), a local RSVP Agent must be invoked into the call flow by Unified CM to accomplish the RSVP signaling on behalf of the IP phone. This is accomplished in EMCC environments by the passing of call control information between the Unified CM clusters to enable an EMCC user logged in remotely to make both intra-cluster and inter-cluster calls with RSVP.

In an EMCC deployment, there are always two clusters for any given login or registration interaction. From the EMCC user's perspective, this would be a home cluster and a visiting cluster (see Figure 11-52). The home cluster is the user's originating cluster, and it is where the user information is stored. The visiting cluster is the phone's originating cluster, where an EMCC user roaming between clusters would log in and where the device information is stored.

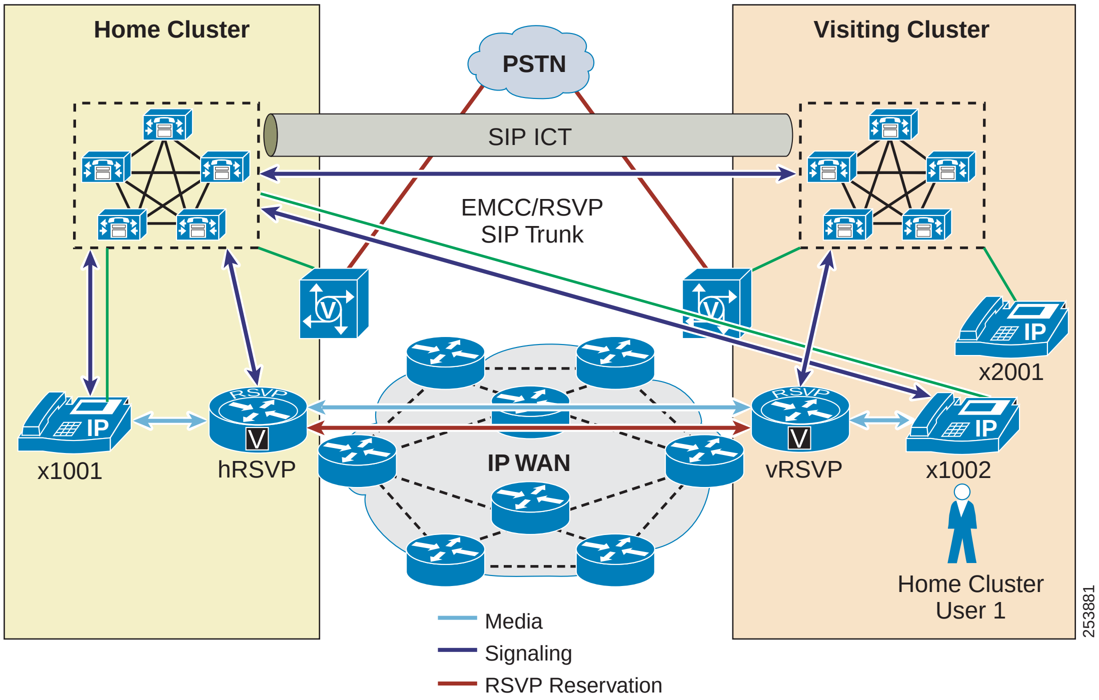
When a user logs into a visiting clusters phone (visiting phone), that phone in turn registers directly with the EMCC user's home cluster. All calls that are then made from that user and visiting phone are made from the call control of the home cluster. The home cluster thus manages the visiting phone and provides this visiting phone with an EMCC roaming device pool. (See Extension Mobility Cross Cluster (EMCC), page 19-10, for further information on EMCC roaming device pools.)

The home cluster makes requests to the visiting cluster for an RSVP Agent when required, and it does this over the EMCC-enabled SIP trunk between the two clusters (specifically in SIP REFER messages). An RSVP Agent is requested from the visiting cluster only when the home cluster has determined that the endpoint requires an RSVP Agent for a call. This is determined by the home cluster from the inter-location RSVP policy between the visiting phone's location (from the EMCC roaming device pool) and the called party's location (from the called party device, gateway, or trunk).

Once the RSVP policy is determined, an RSVP Agent is requested from the visiting cluster for the visiting phone. In this request for the RSVP Agent, the home cluster sends the device name (sep*xxxxxxxxxxxx*) so that the visiting cluster can do a look up on the device name to determine the RSVP Agent (derived from the MRGL on the device itself or on the device pool). Once the home cluster has the information for the RSVP Agent to associate to the visiting phone, it can start the procedures to establish a local RSVP call (RSVP-enabled locations within a cluster) or an RSVP SIP Preconditions call (between clusters).

Figure 11-52 illustrates the signaling and media connections between the various components involved in an EMCC call using RSVP over a SIP-enabled trunk.

Figure 11-52 EMCC Call Using RSVP over a SIP-Enabled Trunk



Best Practices

- Set the location policy between the EMCC roaming device pool location and all other locations to **Mandatory (Video Desired)**.

- RSVP-enabled locations-based call admission control should be functioning prior to enabling EMCC in conjunction with RSVP SIP Preconditions.
- Any IP phone to which an EMCC user can log in must have a local RSVP Agent associated to it.

Unified CM Interoperability and Feature Considerations

This section discusses interoperability considerations between Unified CM and Cisco IOS Gateways and Unified CME.

Cisco IOS Gateway and Unified CME

Both Cisco IOS SIP/TDM gateways and Cisco Unified Communication Manager Express (Unified CME) support RSVP SIP Preconditions. This support enables audio-only calls to be established between Unified CM and the Cisco IOS SIP/TDM gateway or Unified CME signaling RSVP policy over SIP signaling.

For further information regarding SIP RSVP features in Cisco IOS and restrictions for the RSVP SIP Preconditions feature on SIP/TDM Cisco IOS gateways and Unified CME, refer to the *Cisco IOS SIP Configuration Guide*, available at

http://www.cisco.com/en/US/docs/ios/voice/sip/configuration/guide/12_4t/sip_12_4t_book.html

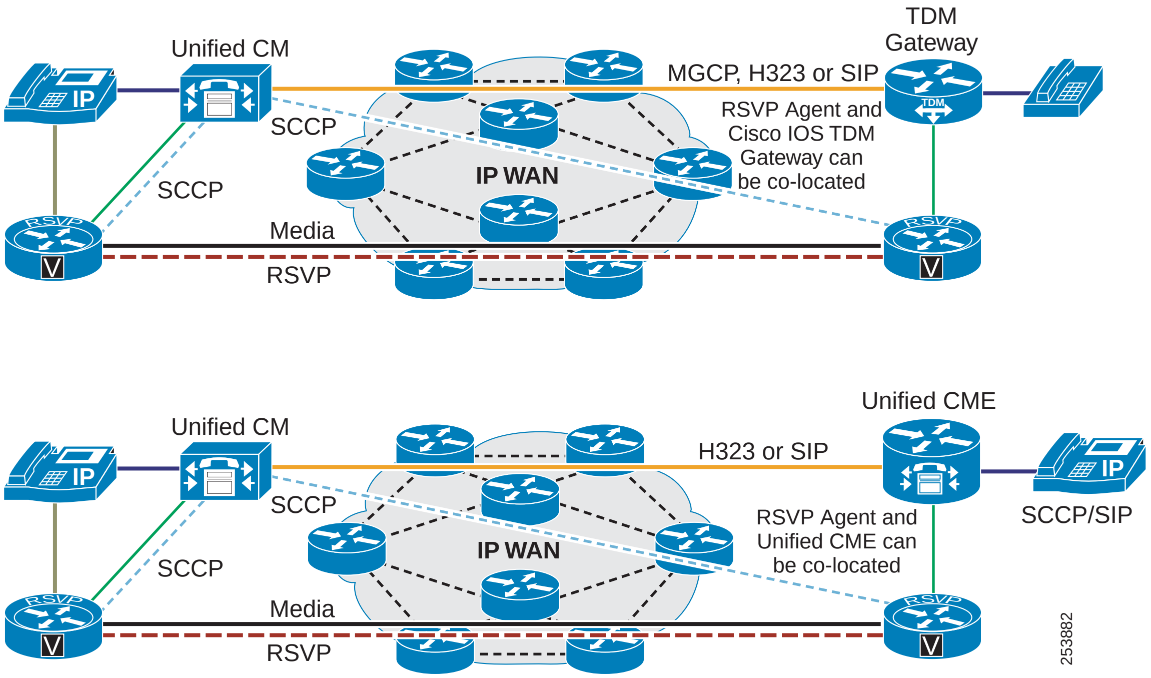
Unified CM has two modes of configuration when interoperating with Cisco IOS gateways and Unified CME in RSVP deployments: local RSVP supporting MGCP, H.323, and SIP call signaling, and end-to-end RSVP supporting SIP signaling only. When interoperating with Cisco IOS gateways and Unified CME, Unified CM can support both of these methods of operation. There are, however, implications when using one method or the other, as described in the following sections.

Unified CM and Local RSVP with Cisco IOS Gateways and Unified CME

In local RSVP mode, Unified CM supports interoperating with Cisco IOS TDM gateways over MGCP, H.323, or SIP call signaling protocols and with Unified CME over H.323 or SIP. In this mode, Unified CM allocates an RSVP agent to the Cisco IOS TDM gateway for calls established to or from the gateway, and it does not signal preconditions or RSVP policy to the Cisco IOS TDM gateway. This is the default configuration for MGCP, H.323, and SIP in Unified CM.

Figure 11-53 illustrates local RSVP integration of Unified CM with a Cisco IOS TDM gateway and with Unified CME.

Figure 11-53 Local RSVP Integration of Unified CM with a Cisco IOS TDM Gateway and Unified CME



Advantages

This model provides the following advantages:

- Support for a wide variety of Cisco IOS gateway signaling protocols (MGCP, H.323, SIP).
- Support for both SIP and SCCP Unified CME endpoints.
- Centralized administration of RSVP policy and Application ID from Unified CM.
- With MGCP for a Cisco IOS TDM gateway, the media path is optimized in call transfer and forward supplementary service scenarios. For calls transferred or forwarded from the local system (Cisco IOS TDM gateway), both media and signaling are torn down and re-established to the transferred or forwarded party.

Disadvantages

This model has the following disadvantages:

- It uses RSVP Agent sessions (software or hardware, depending on session requirements and functionality such as session transcoding).
- With H.323 and SIP integrations for Cisco IOS TDM gateways and Unified CME, the media path is not optimized in transfer and forward supplementary service scenarios. This means that calls transferred or forwarded from the local system (Cisco IOS TDM gateway or Unified CME) hairpin both media and signaling on the local system, which results in double bandwidth consumption.

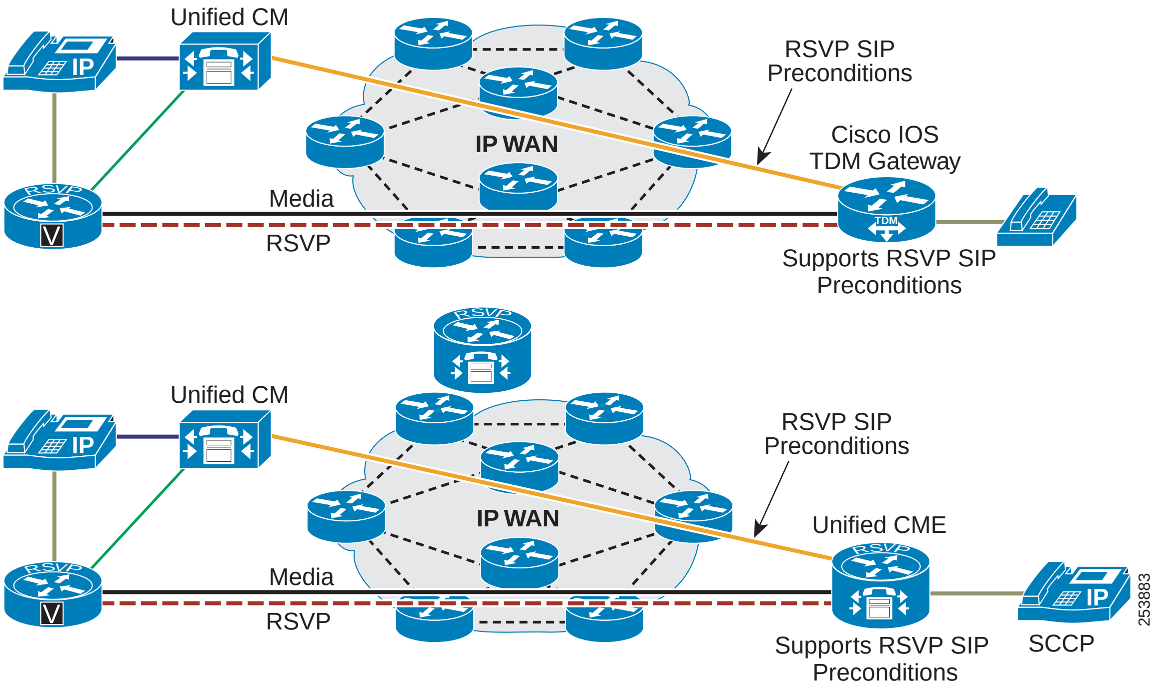
In this method, intra-cluster call admission control functions as explained in the section on Unified CM RSVP-Enabled Locations, page 11-62.

Unified CM and End-to-End RSVP or RSVP SIP Preconditions with Cisco IOS Gateway and Unified CME

In end-to-end RSVP mode, Unified CM supports interoperating with Cisco IOS gateways and Unified CME using RSVP SIP Preconditions signaling. In this mode, Unified CM does not allocate an RSVP agent. The Cisco IOS gateway or Unified CME natively supports RSVP. This method reduces the usage of RSVP agent software sessions on a Cisco Integrated Services Router (ISR).

Figure 11-54 illustrates an RSVP SIP Preconditions integration between Unified CM and a Cisco IOS TDM gateway or Unified CME.

Figure 11-54 RSVP SIP Preconditions Integration Between Unified CM and Cisco IOS Gateway or Unified CME



Advantages

This model provides the following advantages:

- Support for RSVP SIP Preconditions.
- Does not use RSVP Agent resources for SIP Cisco IOS TDM gateways or Unified CME.

Disadvantages

This model has the following disadvantages:

- It supports only SCCP Unified CME endpoints.
- It supports only SIP trunk implementations.
- The media path is not optimized in transfer and forward supplementary service scenarios. This means that calls that are transferred from the local system (SIP Cisco IOS TDM gateway or Unified CME) hairpin both media and signaling on the local system, which results in double bandwidth consumption.

Design Considerations for Unified CM Interoperability with SIP Cisco IOS TDM Gateway and Unified CME

When choosing between local RSVP and end-to-end RSVP deployments, determine the best option based on following criteria:

- The desired call signaling protocol (H.323, MGCP, or SIP). This could be based on many requirements outside of the scope of call admission control, such as dial-plan, PBX interoperability, and call signaling features, to name a few.
- Required supplementary services of call transfer and forward to destinations remote to the local system (SIP Cisco IOS TDM gateway and Unified CME). For example, these services might be required for forwarding of calls over the WAN to centralized voice messaging environments.
- Administration of the solution. Decide between centralized or distributed management of RSVP policy and application ID.
- Resource utilization. Consider the utilization of RSVP Agent sessions versus native RSVP. In some cases the number of sessions might require a dedicated platform and thus cannot reside on the SIP Cisco IOS TDM gateway or Unified CME.
- The SIP trunk configured on Unified CM pointing to the SIP Cisco IOS TDM gateway or Unified CME supporting RSVP SIP Preconditions should always have both an inter-location and an intra-location RSVP policy. The inter-location policy ensures that the correct RSVP policy is set for inbound and outbound calls. The intra-location policy ensures that calls hair-pinned on the same trunk (due to forward and transfer operations) are ensured an end-to-end RSVP policy.
- In Unified CM, Cisco recommends configuring a single separate location that can be applied to all Cisco IOS TDM gateways and Unified CMEs configured on a single cluster. That location should have an inter-location RSVP policy set to **Mandatory** or **Mandatory (Video Desired)** with all other locations including itself. An RSVP policy is required for the correct functioning of RSVP SIP Preconditions in these environments.

Note Even IP phones that are in the same physical LAN as the SIP Cisco IOS TDM gateway require an RSVP policy between their location and the location on the SIP trunk. This will utilize RSVP Agent resources for the IP phones but will not deduct bandwidth over the WAN because the RTP stream remains local.

- Ensure that the RSVP policy configured on Unified CM matches the policy configured on the Cisco IOS TDM gateway. Use the following options under the **dial-peer** configuration when enabling RSVP reservations for the SIP Cisco IOS TDM gateway or Unified CME:

  ```
  req-qos guaranteed-delay audio
  acc-qos guaranteed-delay audio
  ```

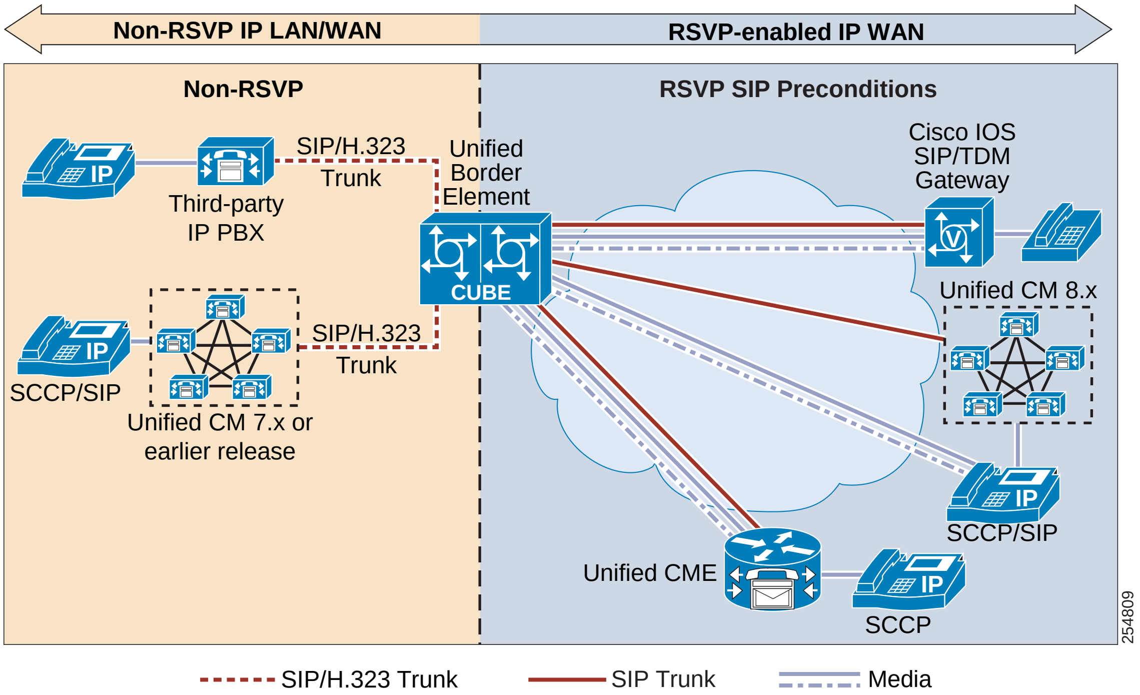
  This configuration ensures that, for each voice call, the SIP Cisco IOS TDM gateway will request an RSVP reservation using the guaranteed delay service. The fact that both the requested QoS and the acceptable QoS specify this RSVP service means that the RSVP reservation is mandatory for the call to succeed (that is, if the reservation cannot be established, the call will fail).
- Ensure that the Application ID configured in Unified CM matches the Application ID configured on the Cisco IOS TDM gateway and Unified CME.
- Ensure that inbound and outbound dial peers are correctly matched to ensure that the appropriate dial peers configured with SIP preconditions are utilized. For further information, refer to the *Cisco IOS SIP Configuration Guide*, available at

  http://www.cisco.com/en/US/docs/ios/voice/sip/configuration/guide/12_4t/sip_12_4t_book.html

Cisco Unified Border Element and RSVP SIP Preconditions

With Cisco Unified Communications System 8.5 and later releases, Cisco Unified Border Element offers audio-only call support for RSVP SIP Preconditions. This support enables enterprises to use the Unified Border Element to integrate non-RSVP call control applications into an RSVP SIP Preconditions infrastructure. On the non-RSVP side of the call control, the Unified Border Element supports integrations with both H.323 and SIP. On the RSVP side of the call, SIP can be used with RSVP Preconditions to integrate with RSVP SIP Preconditions call control such as Unified CM, Unified CME, and SIP-TDM Cisco IOS Gateways. Figure 11-55 illustrates this type of interworking.

Figure 11-55 Cisco Unified Border Element with RSVP over SIP Trunks



For SIP integrations on the non-RSVP side of call control, the Unified Border Element provides support for either Early Offer or Delayed Offer; and for H.323 integrations, either Fast Start or Slow Start is also supported. On the RSVP SIP Preconditions side of call control, SIP Early Offer is always sent in order to provide the preconditions required to negotiate the RSVP policy across call control applications; therefore, RSVP SIP Preconditions is always Early Offer.

For more information, refer to the *Cisco IOS SIP Configuration Guide*, available at

http://www.cisco.com/en/US/docs/ios/voice/sip/configuration/guide/12_4t/sip_12_4t_book.html

Service Advertisement Framework (SAF) and Call Control Discovery (CCD) Considerations

The Cisco Service Advertisement Framework (SAF) enables networking applications to advertise and discover information about networked services within an IP network. Call Control Discovery (CCD) uses SAF to distribute and maintain information about the availability of internal directory numbers (DNs) hosted by call control agents such as Unified CM and Unified CME. CCD also distributes the corresponding number prefixes that allow these internal directory numbers to be reached via the PSTN ("To PSTN" prefixes).

This section discusses SAF CCD as it relates to RSVP SIP Preconditions deployments. For more information on the Service Advertisement Framework and Call Control Discovery, refer to the chapters on Network Infrastructure, page 3-1, Unified Communications Deployment Models, page 5-1, and Dial Plan, page 9-1.

SAF CCD and RSVP SIP Preconditions work together to provide an easy to manage dynamic dial plan for moves, adds, and changes and a topology-aware call admission control method for complex, multi-homed, multi-tiered networks, thus providing a dynamic replacement for a static gatekeeper infrastructure for both dial plan resolution and call admission control that reacts to changes in the network.

SAF CCD works together with RSVP SIP Preconditions call admission control to ensure that there is an alternate route to reach the destination in case of reservation failure. This function is referred to as Call Control Discovery automatic PSTN failover.

Call Control Discovery Automatic PSTN Failover

SAF CCD is different than standard call routing in that only a single IP route can be chosen for any given call; whereas with standard call routing, multiple IP paths may be defined and consecutively attempted for a single call by using route lists and route groups. For any call made using SAF learned routes, the following options exist:

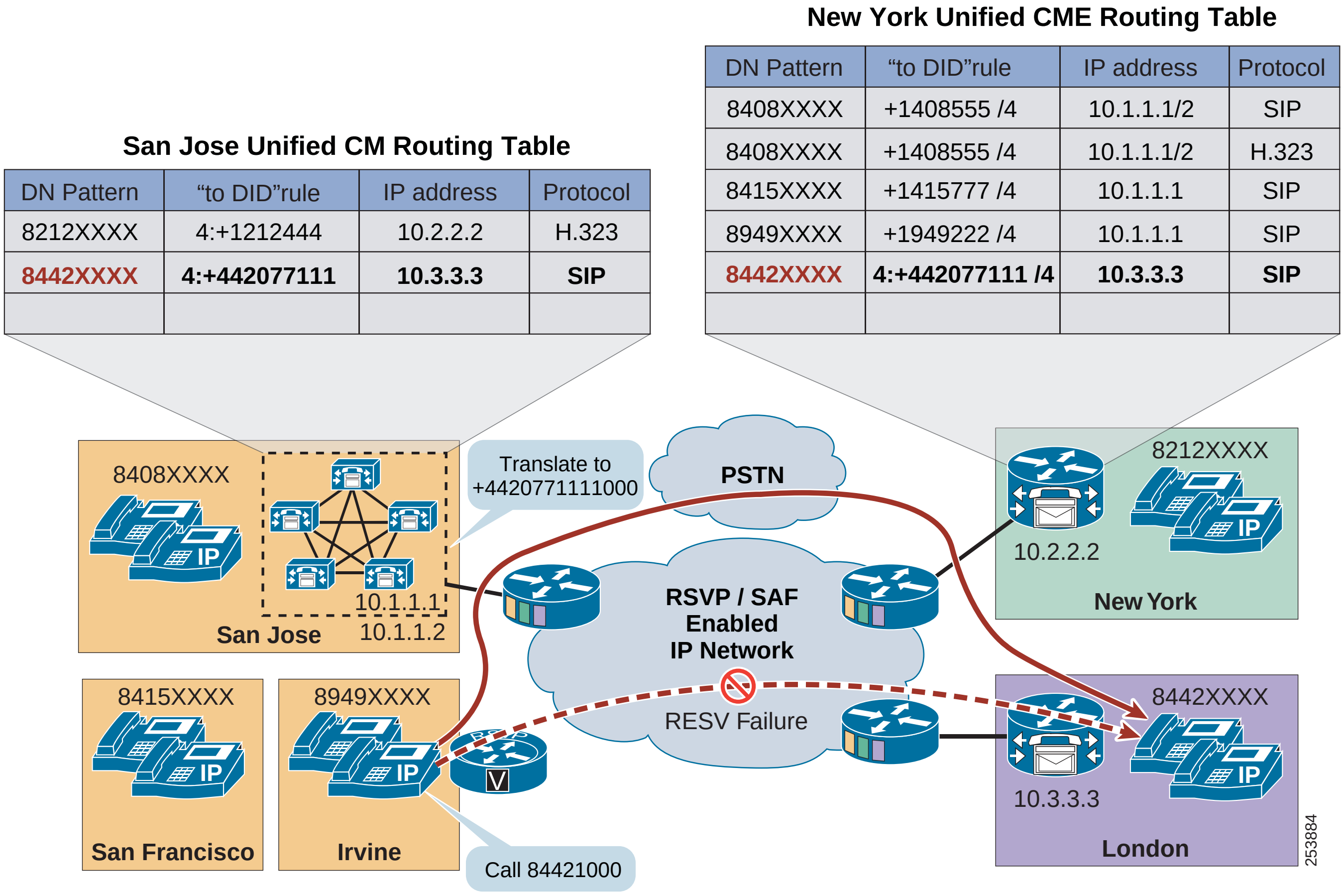
- Take the selected IP path to reach the called number.
- If the IP path is not available, use the PSTN prefix to modify the called number as well as the automatic alternate routing (AAR) calling search space (CSS) from the calling device and route the call via the PSTN.

When RSVP SIP Preconditions is used on a SAF learned route, and if the reservation succeeds, then the call will be established according to normal call establishment on the IP path with RSVP. However, if the reservation fails over the IP route with a returned call termination cause code of Precondition Failure or Reservation Failure (SIP message 580) or Precondition Unsupported (SIP message 420), CCD automatic PSTN failover will occur. (It can also occur on other call termination cause codes, as described below.) CCD automatic PSTN failover is similar to automated alternate routing (see Automated Alternate Routing, page 9-117) in that, when a SAF CCD IP route fails due to a call admission control failure, CCD automatic PSTN failover occurs much like AAR would occur in an intracluster call admission control failure scenario. However, CCD automatic PSTN failover is different from AAR in that it can also occur for other routing failures apart from call admission control. CCD automatic PSTN failover will occur on any call to a learned pattern that fails prior to the alerting phase of the call and that has a call termination cause code other than normal call clearing, user busy, destination out of order, unallocated number, or geo-location mismatch.

CCD automatic PSTN failover uses the AAR CSS as well as the "To PSTN" prefixes (distributed by CCD) to reroute the call. This allows the administrator to leverage the same Class of Service used for AAR call rerouting for local call admission control as for CCD automatic PSTN failover. A key difference is that CCD automatic PSTN failover uses a prefix provided by SAF CCD distribution ("To PSTN" prefix) and not the AAR prefixes.

Figure 11-56 illustrates CCD automatic PSTN failover after RSVP SIP Preconditions call admission control failure.

Figure 11-56 CCD Automatic PSTN Failover with RSVP SIP Preconditions



New York Unified CME Routing Table

DN Pattern	"to DID"rule	IP address	Protocol
8408XXXX	+1408555 /4	10.1.1.1/2	SIP
8408XXXX	+1408555 /4	10.1.1.1/2	H.323
8415XXXX	+1415777 /4	10.1.1.1	SIP
8949XXXX	+1949222 /4	10.1.1.1	SIP
8442XXXX	**4:+442077111 /4**	**10.3.3.3**	**SIP**

San Jose Unified CM Routing Table

DN Pattern	"to DID"rule	IP address	Protocol
8212XXXX	4:+1212444	10.2.2.2	H.323
8442XXXX	**4:+442077111**	**10.3.3.3**	**SIP**

There is also a difference in functionality between a CCD automatic PSTN failover from a SAF CCD learned route and a reroute by route list and route group functionality with a static route pattern. With a static route pattern pointing to a route group and route list, when a RSVP SIP Preconditions reservation failure occurs, Unified CM routes the call to the next trunk or gateway configured in the route group and list.

In either case (using SAF or a static route pattern), Cisco recommends ensuring that the next choice after call admission control failure is to route the call to a local route group. (See Local Route Group, page 9-13, for more information on local route groups.) This has to be done in the constructs of the CCD automatic PSTN failover function. It is important to ensure that the correct calling search space is configured in the AAR CSS of the calling party to ensure that, when the CCD automatic failover occurs, the call is directed to a route pattern that will engage the local route group function and route the call to the local gateway. This route pattern can be a catch-all pattern used specifically for all CCD automatic failover conditions to ensure the routing of calls out the gateway local of the calling party.

Cisco Unified SIP Proxy Considerations

The Cisco Unified SIP Proxy is a high-performance, highly available stateless Session Initiation Protocol (SIP) server for centralized routing and SIP signaling normalization. By forwarding requests between call control domains, the Cisco Unified SIP Proxy provides the means for routing sessions within enterprise and service provider networks. The main purpose for the Unified SIP Proxy in Cisco Unified Communications deployments is the aggregation of SIP signaling, SIP normalization, and dial plan centralization. For more information on the Cisco Unified SIP Proxy and its features and functions, refer to the documentation at

http://www.cisco.com/en/US/prod/collateral/modules/ps2797/data_sheet_c78-521390_ps2797_Products_Data_Sheet.html

In a RSVP SIP Preconditions environment, the Unified SIP Proxy simply passes along the preconditions that are contained the SDP portion of various SIP messages and does not modify the preconditions in any way.

Adaptive Security Appliance (ASA) Considerations

When deploying a Cisco ASA between any of the Cisco Unified Communications call processing applications such as Unified CM, Unified CME, Unified SIP Proxy, or Cisco IOS SIP/TDM gateway with RSVP SIP Preconditions, both of the following inspections are required:

- SIP Inspection

  The SIP inspection allows the SIP signaling from any Cisco SIP signaling product to traverse the ASA. The ASA subsequently opens the appropriate media pinholes that are recorded in the SDP of the various SIP messages. This is important when the ASA is in the media path between RSVP Agents that are reserving bandwidth and sourcing media flows for Unified Communications endpoints.

- IP Options Inspection

  The IP Options inspection allows the RSVP signaling from RSVP Agent to RSVP Agent to traverse the ASA. All RSVP messages have the IP Router Alert Option set in the IP header of every packet. The ASA drops these packets by default unless the IP Router Alert Option is allowed in the IP Options inspection so that these packets are allowed to flow through the ASA.

Support for both SIP inspection and IP Options inspection specific to RSVP SIP Preconditions implementation is provided in ASA Software Release 8.3. Therefore, for compatibility reasons you must use ASA 8.3 or later software release in any RSVP SIP Preconditions deployment where the ASA is required to inspect the SIP signaling and/or pass RSVP packets.

For information on configuring the ASA for the SIP and IP Options inspections, refer to the *Cisco ASA 5500 Series Configuration Guide*, available at

http://www.cisco.com/en/US/products/ps6120/products_installation_and_configuration_guides_list.html

Design Considerations for Call Admission Control

This section describes how to apply the call admission control mechanisms to various IP WAN topologies. With Unified CM Enhanced Locations CAC network modeling support, Unified CM is no longer limited to supporting simple hub-and-spoke or MPLS topologies but, together with intercluster enhanced locations, can now support most any network topology in any Unified CM deployment model. Enhanced Locations CAC is still a statically defined mechanism that does not query the network, and therefore the administrator still needs to provision Unified CM accordingly whenever network changes affect admission control. This is where a network-aware mechanism such as RSVP can fill that gap and provide support for dynamic changes in the network, such as when network failures occur and media streams take different paths in the network. This is often the case in designs with load-balanced dual or multi-homed WAN uplinks or unequally sized primary and backup WAN uplinks.

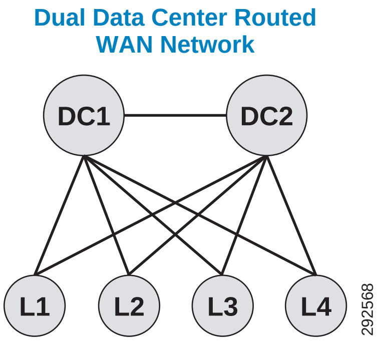
To learn how Enhanced Locations CAC functions and how to design and deploy Enhanced location CAC, see the section on Unified CM Enhanced Locations Call Admission Control, page 11-12.

In this section explores a few typical topologies and explains how Enhanced Locations CAC can be designed to manage them.

Dual Data Center Design

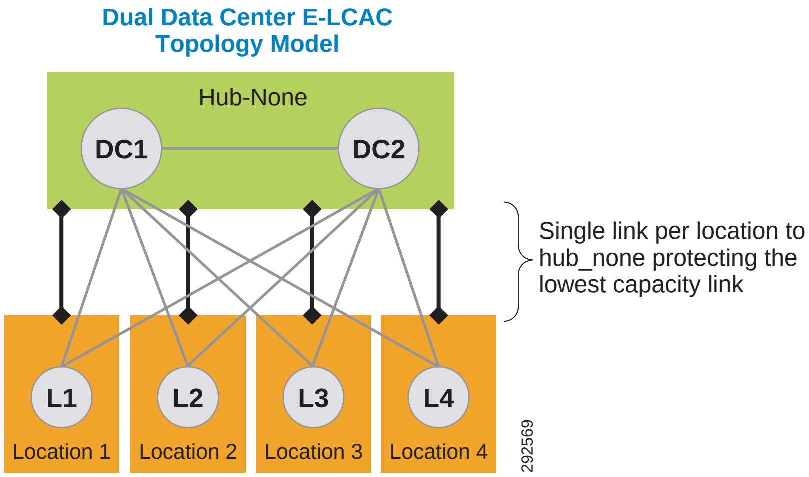
Figure 11-57 illustrates a simple dual data center WAN network design where each remote site has a single WAN uplink to each data center. The data centers are interconnected by a high-speed WAN connection that is over-provisioned for data traffic.

Figure 11-57 Dual Data Center WAN Network



Typically these WAN uplinks from the remote sites to the data centers are load-balanced or in a primary/backup configuration, and there are limited ways for a static CAC mechanism to handle these scenarios. Although you could configure this multi-path topology in Enhanced Locations CAC, only one path would be calculated as the effective path and would remain statically so until the weight metric was changed. A better way to support this type of network topology is to configure the two data centers as one data center or hub location in Enhanced Locations CAC and configure a single link to each remote site location. Figure 11-58 illustrates an Enhanced Locations (E-L) CAC locations and links overlay.

Figure 11-58 Enhanced Locations CAC Topology Model for Dual Data Centers



Design Recommendations

The following design recommendations for dual data centers with remote dual or more links to remote locations apply to both load-balanced and primary/backup WAN designs:

- A single location (Hub_None) represents both data centers.
- A single link between the remote locations and Hub_None protects the remote site uplinks from over-subscription during normal conditions or failure of the highest bandwidth capacity links.
- The capacity of link bandwidth allocation between the remote site and Hub_None should be equal to the lowest bandwidth capacity for the applicable Unified Communications media for a single link. For example, if each WAN uplink can support 2 Mbps of audio traffic marked EF, then the link audio bandwidth value should be no more than 2 Mbps to support a failure condition or equal-cost path routing.

MPLS Clouds

When designing for Multiprotocol Label Switching (MPLS) any-to-any connectivity type clouds in the E-L CAC network model, a single location can serve as the MPLS cloud. This location will not have any devices associated to it, but all of the sites that have uplinks to this cloud will have links configured to the location. In this way the MPLS cloud serves as a transit location for interconnecting multiple variable-sized bandwidth WAN uplinks to other remote locations. The illustrations in this section depict a number of different MPLS networks and their equivalent locations and links model.

In Figure 11-59, Hub_None represents the MPLS cloud serving as a transit location interconnecting the campus location where servers, endpoints, and devices are located, with remote locations where only endpoints and devices are located. Each link to Hub_None from the remote location may be sized according to the WAN uplink bandwidth allocated for audio, video, and immersive media.

Figure 11-59 Single MPLS Cloud

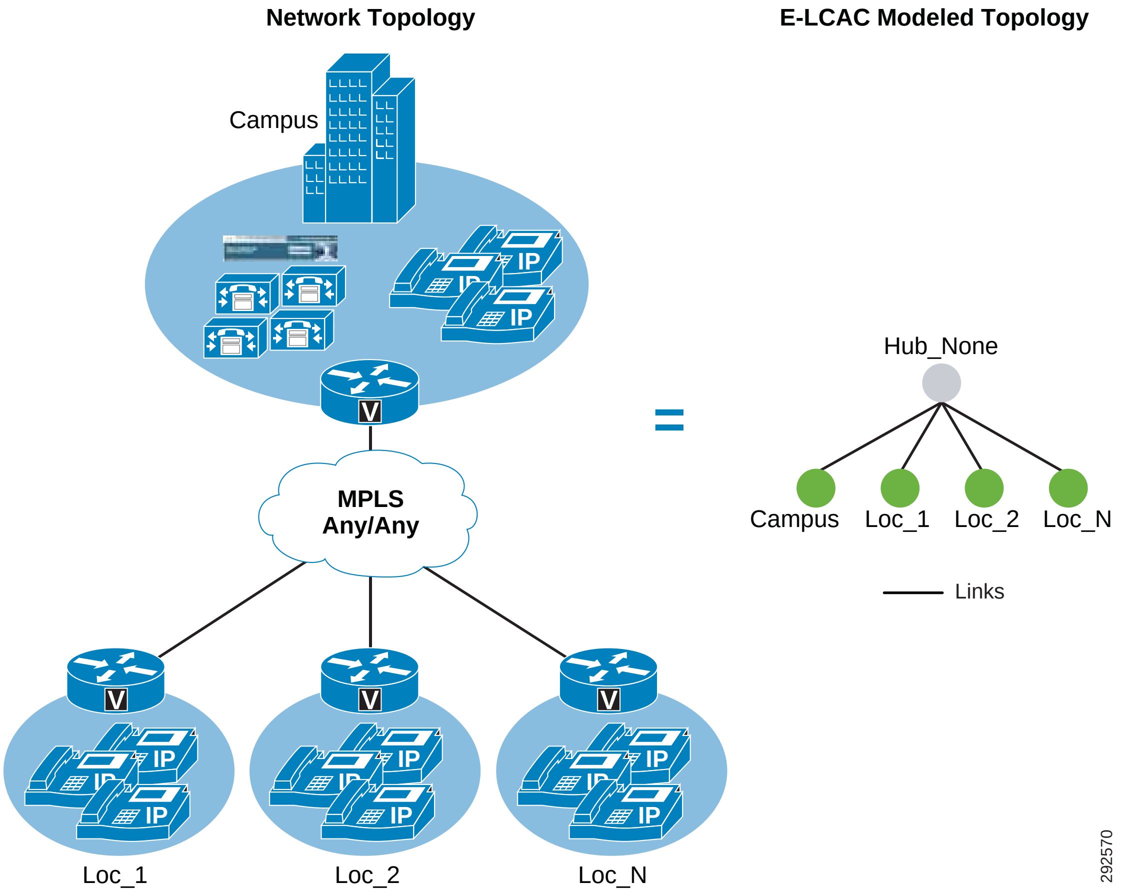


Figure 11-60 shows two MPLS clouds that serve as transit locations interconnecting the campus location where servers, endpoints, and devices are located, with remote locations where only endpoints and devices are located. The campus also connects to both clouds. Each link to the MPLS cloud from the remote location may be sized according to the WAN uplink bandwidth allocated for audio, vide, and immersive media. This design is typical in enterprises that span continents, with a separate MPLS cloud from different providers in each geographical location.

Figure 11-60 Separate MPLS Clouds

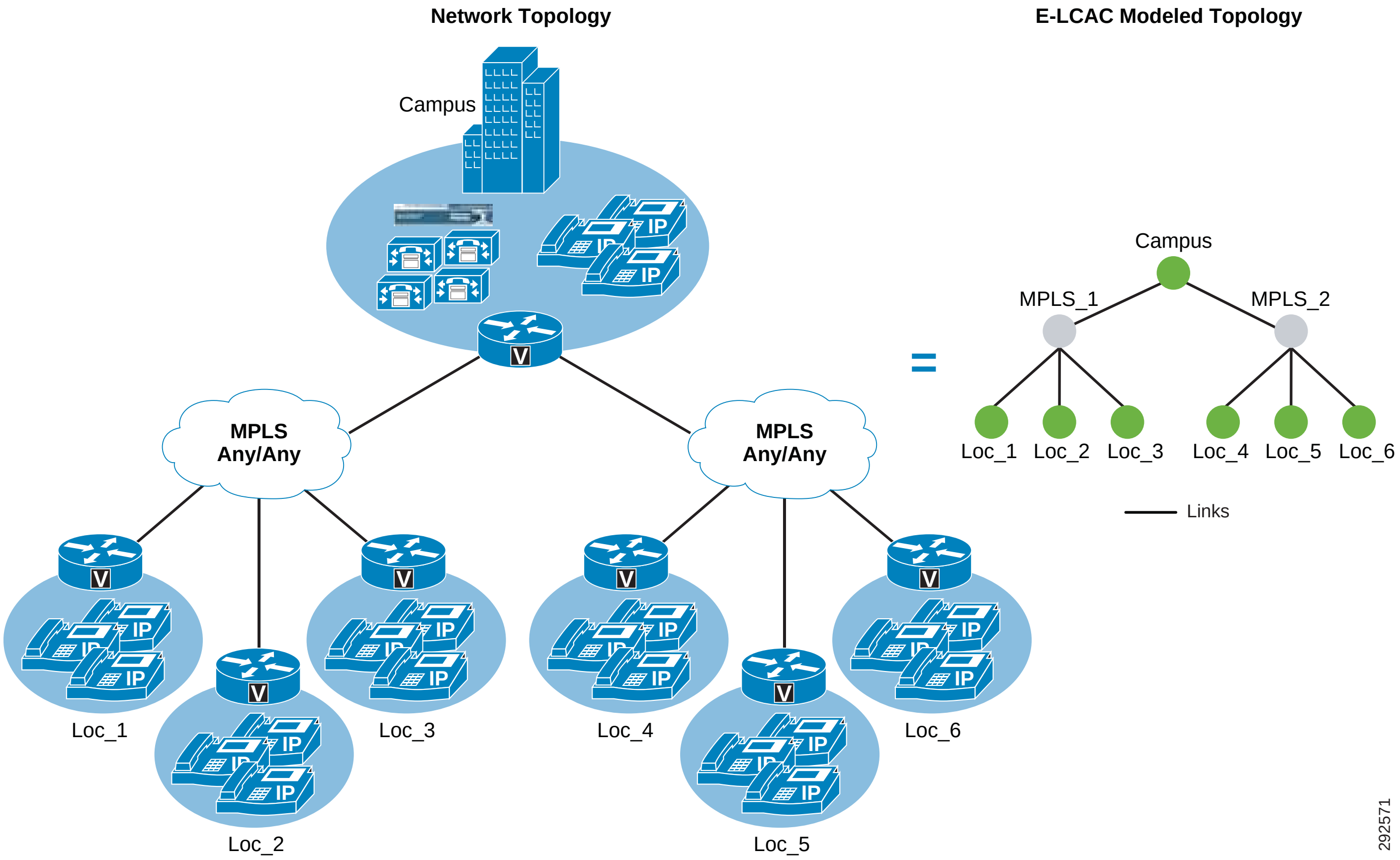
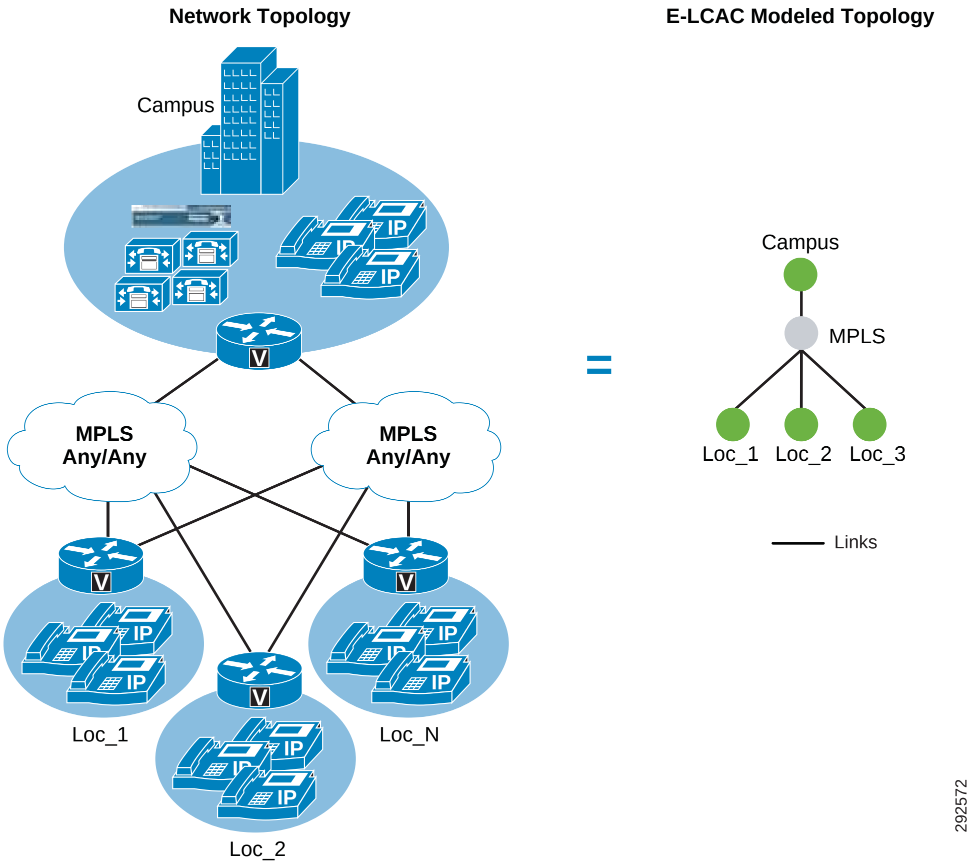


Figure 11-61 shows multiple MPLS clouds from different providers, where each site has one connection to each cloud and uses the MPLS clouds in either an equal-cost load-balanced manner or in a primary/backup scenario. In any case, this design is equivalent to the dual data center design where a single location represents both clouds and a single link represents the lowest capacity link of the two.

Figure 11-61 Remote Sites Connected to Dual MPLS Clouds



Design Recommendations

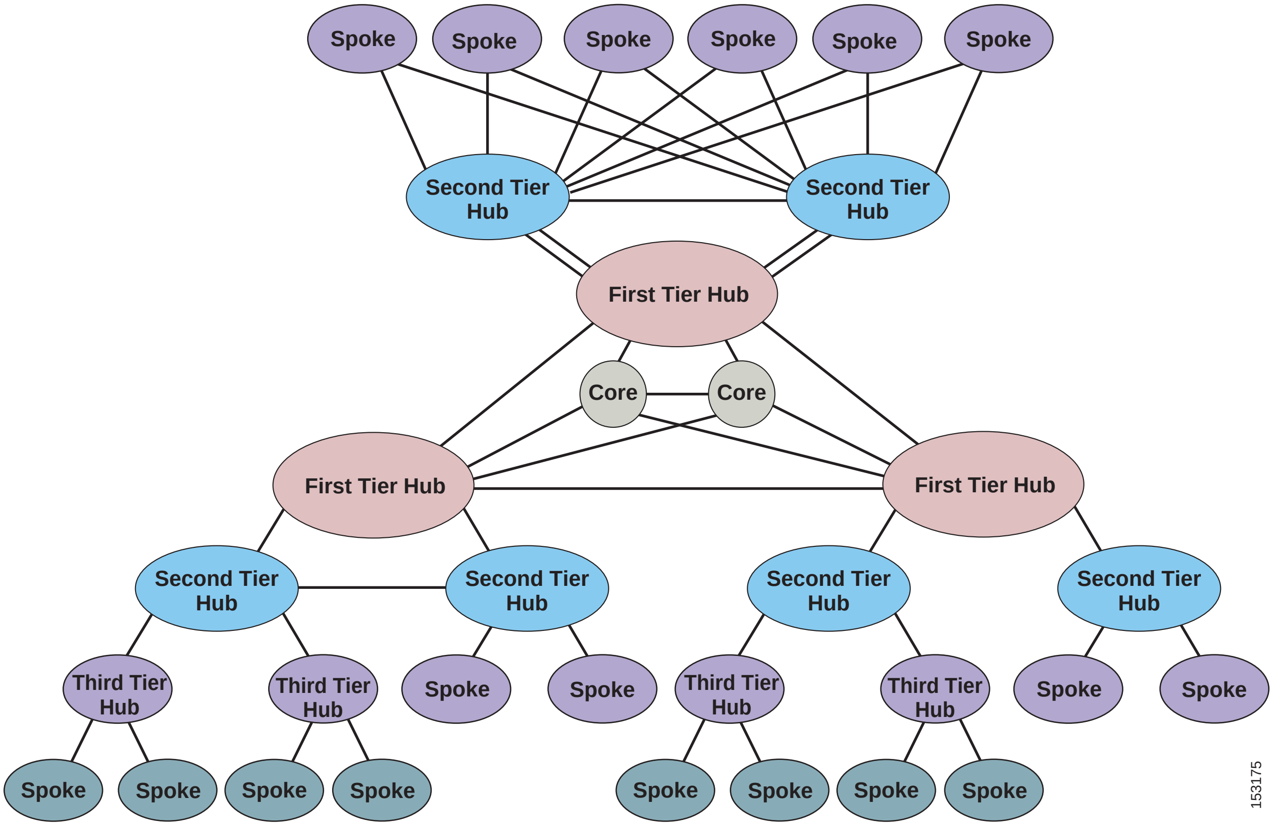
- The MPLS cloud should be configured as a location that does not contain any endpoints but is used as a hub to interconnect locations.
- The MPLS cloud serves as a transit location for interconnecting multiple variable-sized bandwidth WAN uplinks to other remote locations.
- Remote sites with connectivity to dual MPLS clouds should treat those connections as a single link and size to the lowest capacity of the links in order to avoid oversubscription during network failure conditions.

Generic Topologies

In the context of this chapter, a generic topology is a network topology that cannot be reduced to a simple hub-and-spoke, a two-tier hub-and-spoke, or a simple MPLS-based network.

As Figure 11-62 illustrates, a generic topology can present full-mesh features, hub-and-spoke features, partial-mesh features, or possibly all of them combined in a single network. It may also present dual connections between sites, as well as multiple paths from one site to another.

Figure 11-62 A Generic Topology



The complex nature of these networks requires the adoption of topology-aware call admission control mechanisms based on RSVP. In particular, these mechanisms can properly control bandwidth in presence of any of the following topology aspects:

- Remote sites dual-homed to different hub sites
- Multiple IP WAN links between any two sites, either in a primary/backup configuration or in an active/active load-balanced configuration
- Redundant hubs or data centers with a dedicated connection
- Fully-meshed core networks
- Multiple equal-cost IP paths between any two sites
- Multi-tiered architectures

The remainder of this section contains design best practices for generic network topologies according to the Unified CM deployment model adopted:

- Centralized Unified CM Deployments, page 11-99

  One or more Unified CM clusters are located at a given site, but only endpoints and gateways are located at all other sites.

- Distributed Mixed Call Processing Deployments, page 11-104

  Call control applications are distributed in various topologies.

Centralized Unified CM Deployments

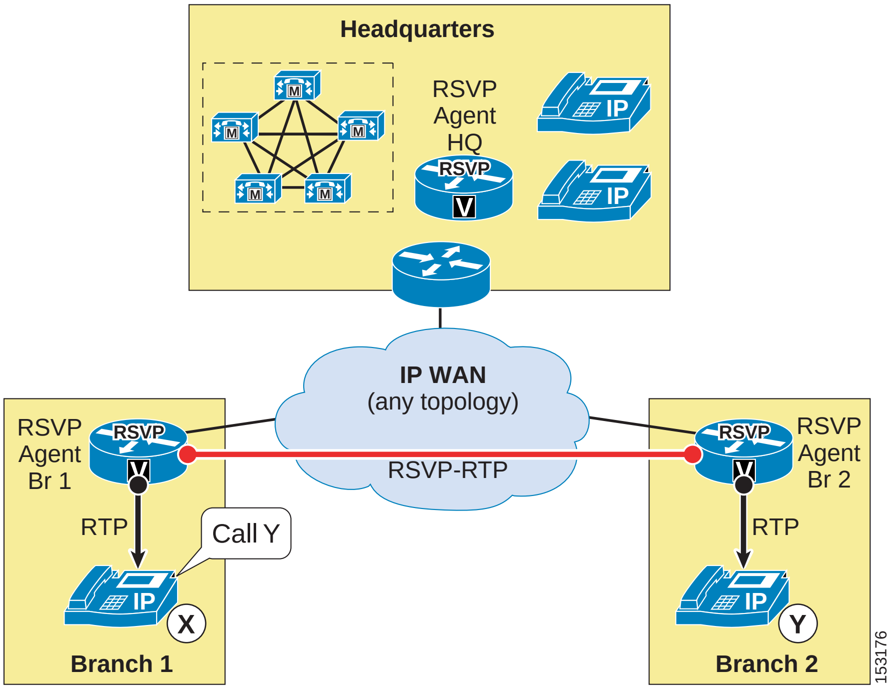
Centralized Unified CM deployments using a generic topology can be categorized into two sub-types:

- Single Unified CM Cluster, page 11-99
- Co-Located Unified CM Clusters, page 11-100

Single Unified CM Cluster

The recommendations in this section apply to a single Unified CM cluster deployed in a generic network topology, as illustrated in Figure 11-63.

Figure 11-63 A Single Unified CM Cluster in a Generic Topology



The following guidelines apply to this type of deployment:

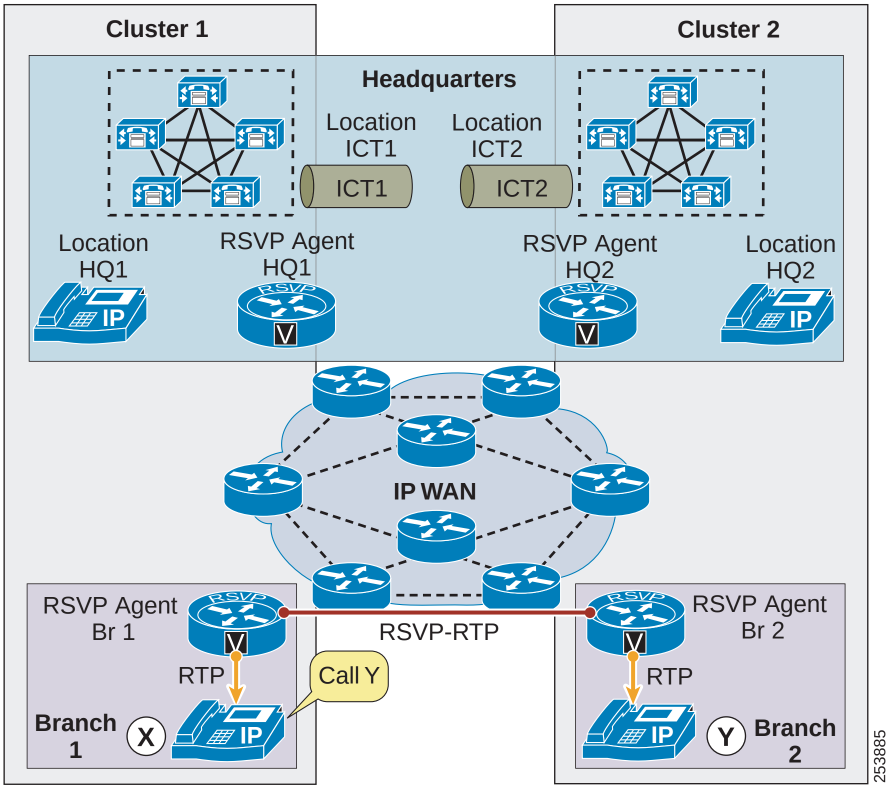
- Enable the Cisco IOS RSVP Agent feature on a Cisco IOS router at each site, including the central site where Unified CM resides. At smaller sites, this router may coincide with the IP WAN router and PSTN gateway, while at larger sites they may be different platforms.
- In Unified CM, define a location for each site, and leave all bandwidth values as **Unlimited**.
- Assign all devices located at each site to the appropriate location (this includes endpoints, gateways, conferencing resources, and the Cisco RSVP Agents themselves).
- Ensure that each Cisco RSVP Agent belongs to a media resource group (MRG) contained in the media resource group list (MRGL) of all devices at that site.
- In the Unified CM service parameters, set the **Default inter-location RSVP Policy** to **Mandatory** or **Mandatory (video desired)** and set the **Mandatory RSVP mid-call error handle option** to **Call fails following retry counter exceeded**.
- Enable RSVP on every WAN router interface in the network where congestion might occur, and configure the RSVP bandwidth based on the provisioning of the priority queue. (See RSVP Design Best Practices, page 11-57.)
- If you need to provision bandwidth separately for voice and video calls, also configure an RSVP application ID on the same WAN router interfaces.
- If the Cisco RSVP Agent is not co-resident with the IP WAN router, enable RSVP on the LAN interfaces connecting the agent to the WAN router.

Co-Located Unified CM Clusters

The recommendations in this section apply to deployments where multiple Unified CM clusters are located on the same LAN or MAN. However, the same considerations may also be valid if the sites where the Unified CM clusters reside are connected via a lower bandwidth link. Due to the design, any call cluster-to-cluster will engage an RSVP Agent for an endpoint in each cluster.

Figure 11-64 illustrates a deployment with two Unified CM clusters located at a given site (HQ) and a number of remote sites with endpoints and gateways, which are controlled either by Cluster 1 (for example, Branch 1) or Cluster 2 (for example, Branch 2).

Figure 11-64 Co-Located Unified CM Clusters in a Generic Topology



The following guidelines apply to the deployment in Figure 11-64:

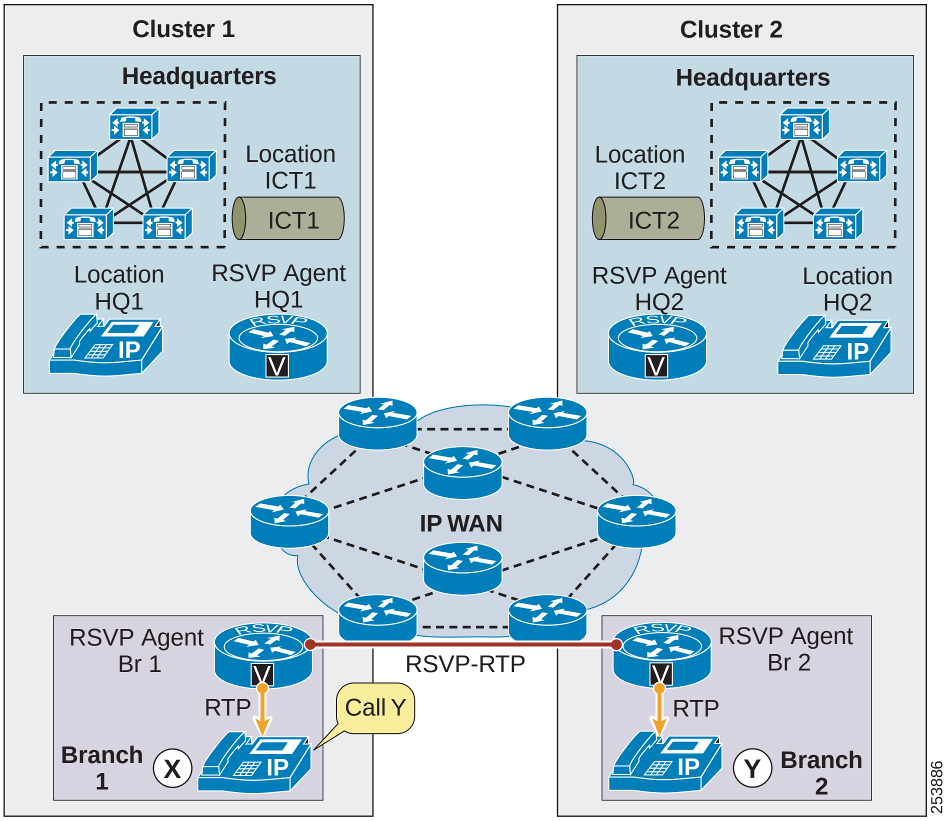
- Enable the Cisco IOS RSVP Agent feature on a Cisco IOS router at each site, including the central site where Unified CM resides. At smaller sites, this router may coincide with the IP WAN router and PSTN gateway, while at larger sites they may be different platforms.
- Depending on the amount of call traffic between the central site and remote sites and between clusters, consider co-locating the RSVP Agents for both clusters at the central site on a single or multiple Cisco Integrated Services Routers (ISR). A single ISR can host multiple RSVP Agents controlled by different clusters.
- In each Unified CM cluster, define a location for each site, and leave all bandwidth values as unlimited.
- Assign all devices located at each site to the appropriate location. This includes endpoints, gateways, conferencing resources, and the Cisco RSVP Agents themselves.
- Ensure that each Cisco RSVP Agent belongs to a media resource group (MRG) contained in the media resource group list (MRGL) of all devices at that site.
- In the Unified CM service parameters, set the **Default inter-location RSVP Policy** to **Mandatory** or **Mandatory (video desired)** and set the **Mandatory RSVP mid-call error handle option** to **Call fails following retry counter exceeded** on both clusters.
- Enable RSVP on every WAN router interface in the network where congestion might occur, and configure the RSVP bandwidth based on the provisioning of the priority queue. (See RSVP Design Best Practices, page 11-57.)

- If you need to provision bandwidth separately for voice and video calls, also configure an RSVP application ID on the same WAN router interfaces.
- If the Cisco RSVP Agent is not co-resident with the IP WAN router, enable RSVP on the LAN interfaces connecting the agent to the WAN router.
- Ensure that RSVP SIP Preconditions are enabled on the SIP intercluster trunks (see Migration from Enhanced Locations Call Admission Control to RSVP SIP Preconditions, page 11-78, for the steps).
- Ensure that an inter-location RSVP policy of **Mandatory** or **Mandatory (video desired)** is set between the SIP intercluster trunk location and all locations, including itself (see intra-location RSVP policy below).
- Ensure that an intra-location RSVP policy of **Mandatory** or **Mandatory (Video Desired)** is set on the SIP intercluster trunk. An intra-location RSVP policy is set by selecting the location and configuring a policy for itself. This effectively ensures that calls within this location will engage RSVP call admission control. This is important for calls hairpinned on the trunk from call transfers or forwards back to the originating cluster.
- For calls from cluster to cluster within the central site, RSVP Agents will be engaged. This is due to the design and the ability for supplementary services to function across clusters, keeping intact end-to-end RSVP across the clusters. There are a few variations that can be supported. Consult with your Cisco account team to determine the best possible call admission control design for co-located clusters.

Distributed Unified CM Deployments

RSVP SIP Preconditions provides call admission control for distributed deployments of Unified Communication Manager clusters in a generic network topology. This section contains an example of a dual-cluster deployment with RSVP SIP Preconditions support between the clusters (see Figure 11-65). Hybrids and variations of this model are expected, and this is only a high-level example of a simple design, with best practices and design considerations called out.

Figure 11-65 Distributed Unified CM Deployment



The following guidelines apply to the deployment in Figure 11-65:

- Enable the Cisco IOS RSVP Agent feature on a Cisco IOS router at each site, including the central site where Unified CM resides. At smaller sites, this router may coincide with the IP WAN router and PSTN gateway, while at larger sites they may be different platforms.
- In each Unified CM cluster, define a location for each site, and leave all bandwidth values as unlimited.
- Assign all devices located at each site to the appropriate location. This includes endpoints, gateways, conferencing resources, and the Cisco RSVP Agents themselves.
- Ensure that each Cisco RSVP Agent belongs to a media resource group (MRG) contained in the media resource group list (MRGL) of all devices at that site.
- In the Unified CM service parameters, set the **Default inter-location RSVP Policy** to **Mandatory** or **Mandatory (video desired)**, and set the **Mandatory RSVP mid-call error handle option** to **Call fails following retry counter exceeded** on both clusters.

- Enable RSVP on every WAN router interface in the network where congestion might occur, and configure the RSVP bandwidth based on the provisioning of the priority queue. (See RSVP Design Best Practices, page 11-57.)
- If you need to provision bandwidth separately for voice and video calls, also configure an RSVP application ID on the same WAN router interfaces.
- If the Cisco RSVP Agent is not co-resident with the IP WAN router, enable RSVP on the LAN interfaces connecting the agent to the WAN router.
- Ensure that RSVP SIP Preconditions are enabled on the SIP intercluster trunks (see Migration from Enhanced Locations Call Admission Control to RSVP SIP Preconditions, page 11-78, for the steps).
- Ensure that an inter-location RSVP policy of **Mandatory** or **Mandatory (video desired)** is set between the SIP intercluster trunk location and all locations, including itself (see intra-location RSVP policy below).
- Ensure that an intra-location RSVP policy of **Mandatory** or **Mandatory (Video Desired)** is set on the SIP intercluster trunk. An intra-location RSVP policy is set by selecting the location and configuring a policy for itself. This effectively ensures that calls within this location will engage RSVP call admission control. This is important for calls hairpinned on the trunk from call transfers or forwards back to the originating cluster.

Distributed Mixed Call Processing Deployments

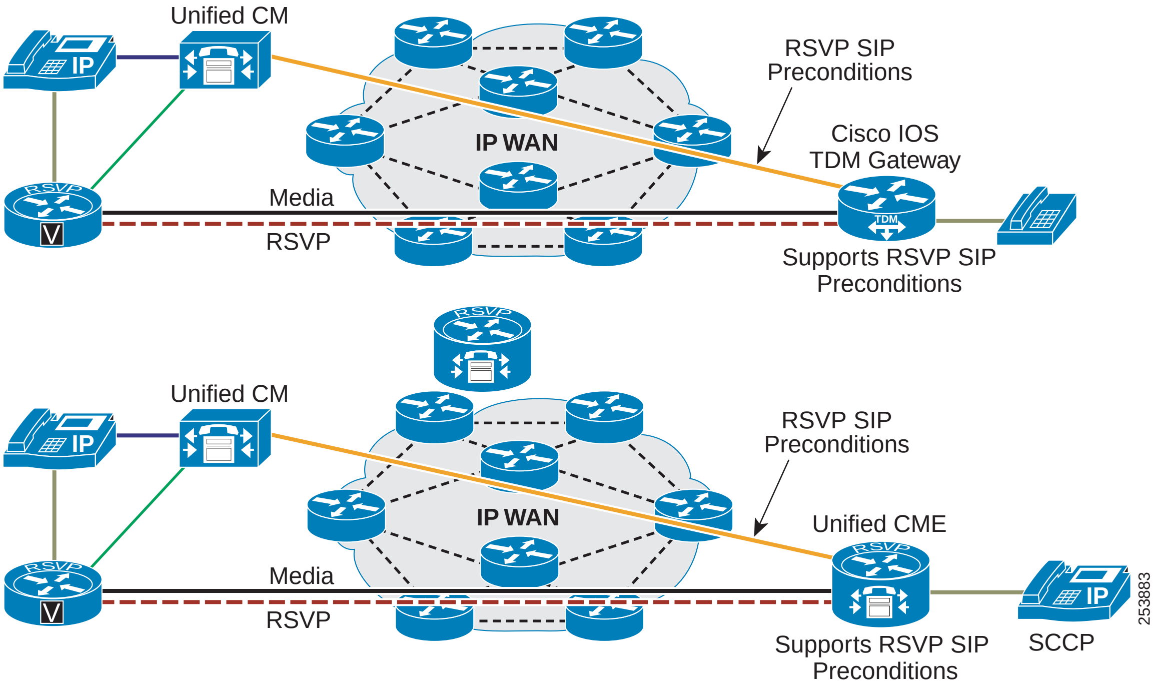
RSVP SIP Preconditions provides call admission control for distributed deployments of Unified Communications call control applications in a generic network topology.

This section contains a list of the supported RSVP SIP Preconditions deployment models. Hybrids and variations of these models are expected, and these are only high-level examples of the design possibilities, with best practices and design considerations called out. (Consult the specific product documentation for more information on configuration of these features.)

The following guidelines apply to the deployments in Figure 11-66:

- The deployment supports Unified CME SCCP integration for audio-only calls.
- Enable RSVP on every WAN router interface in the network where congestion might occur, and configure the RSVP bandwidth based on the provisioning of the priority queue. (See RSVP Design Best Practices, page 11-57.)
- If you need to provision bandwidth separately for voice and video calls, also configure an RSVP application ID on the same WAN router interfaces.
- Follow the recommendations listed in the Design Considerations for Unified CM Interoperability with SIP Cisco IOS TDM Gateway and Unified CME, page 11-88.

Figure 11-66 Unified CM to Cisco IOS Gateway (TDM) and to Unified CME (SCCP)



The following guidelines apply to the deployments in Figure 11-67:

- Follow the guidelines, best practices, limitations, and restrictions for SIP Cisco IOS TDM gateway and Unified CME interoperability listed in the *Cisco IOS SIP Configuration Guide*, available at

  http://www.cisco.com/en/US/docs/ios/voice/sip/configuration/guide/12_4t/sip_12_4t_book.html

- Ensure that the RSVP policy configured on each Unified CME or SIP Cisco IOS TDM gateway is consistent in order to avoid failed or unprotected calls. Use the following options under the **dial-peer** configuration when enabling RSVP reservations for the SIP Cisco IOS TDM gateway or Unified CME:

  ```
  req-qos guaranteed-delay audio
  acc-qos guaranteed-delay audio
  ```

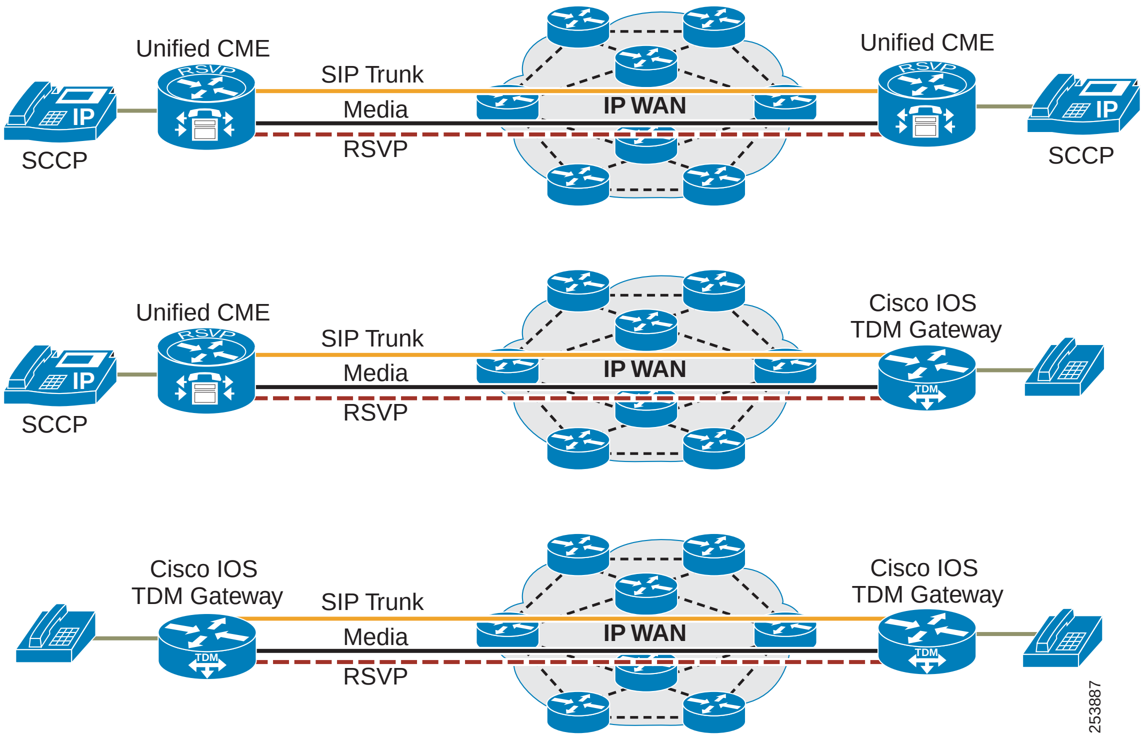
  This configuration ensures that for each voice call, the SIP Cisco IOS TDM gateway will request an RSVP reservation using the guaranteed delay service. The fact that both the requested QoS and the acceptable QoS specify this RSVP service means that the RSVP reservation is mandatory for the call to succeed. (That is, if the reservation cannot be established, the call will fail.)

- If Application ID is used, ensure that it is consistent across all of products in the solution (SIP Cisco IOS TDM gateway and Unified CME).

- Ensure that inbound and outbound dial peers are correctly matched to ensure that the appropriate dial peers configured with SIP Preconditions are utilized. For further information, refer to the *Cisco IOS SIP Configuration Guide*, available at

  http://www.cisco.com/en/US/docs/ios/voice/sip/configuration/guide/12_4t/sip_12_4t_book.html

Figure 11-67 Unified CME to Unified CME, Unified CME to Cisco IOS Gateway, and Cisco IOS Gateway to Cisco IOS Gateway



The following guidelines apply to the deployments in Figure 11-68:

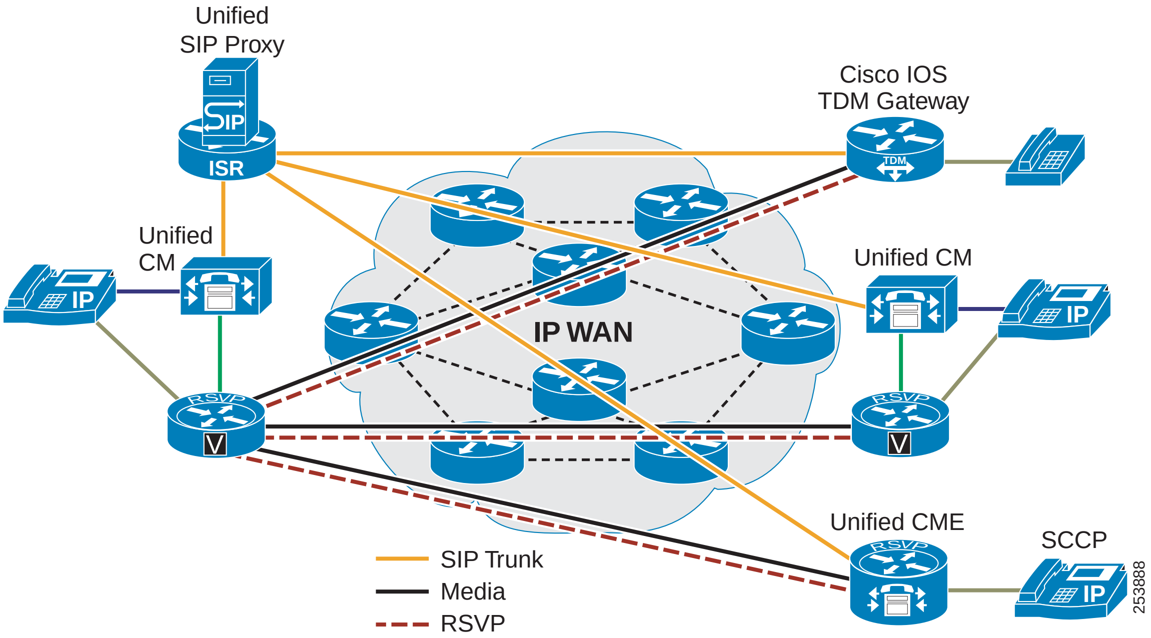
- Follow the guidelines stipulated for Figure 11-66 for Unified CM in distributed call processing deployments, and follow the guidelines stipulated for Figure 11-67 for Unified CME and SIP Cisco IOS TDM Gateways.
- Each Unified CM cluster will typically have a single trunk directed to the Cisco Unified SIP Proxy. Ensure that RSVP SIP Preconditions (end-to-end RSVP) is enabled on that trunk.
- Ensure that RSVP SIP Preconditions are enabled on the SIP trunk to the Cisco Unified SIP Proxy (see Migration from Enhanced Locations Call Admission Control to RSVP SIP Preconditions, page 11-78, for the steps).
- Ensure that an inter-location RSVP policy is configured on each Unified CM cluster between the IP phone locations and the SIP trunk location. This ensures that the SIP preconditions will be enabled for all calls engaged on the SIP trunk to the Cisco Unified SIP Proxy.
- If there are potential SIP destinations that do not support RSVP SIP Preconditions, then ensure that RSVP SIP Preconditions fallback to local RSVP is configured on the SIP trunk to allocate an RSVP Agent for those call flows. And if RSVP SIP Preconditions fallback is enabled, ensure that the RSVP Agent associated to the SIP trunk is in a physical site that will ensure RSVP path protection for those call flows.
- Unified CME and SIP Cisco IOS TDM gateways will also typically have a single SIP dial peer directed to the Cisco Unified SIP Proxy. Ensure that RSVP SIP Preconditions (SIP Preconditions support) is configured on those dial peers, as stipulated in the *Cisco IOS SIP Configuration Guide*, available at

  http://www.cisco.com/en/US/docs/ios/voice/sip/configuration/guide/12_4t/sip_12_4t_book.html

- For information on configuration, guidelines, best practices, limitations, and restrictions for Cisco Unified SIP Proxy, refer to the documentation at

  http://www.cisco.com/en/US/prod/collateral/modules/ps2797/data_sheet_c78-521390_ps2797_Products_Data_Sheet.html

Figure 11-68 All Components Capable of RSVP SIP Preconditions via Cisco Unified SIP Proxy



The following guidelines apply to the deployments in Figure 11-69:

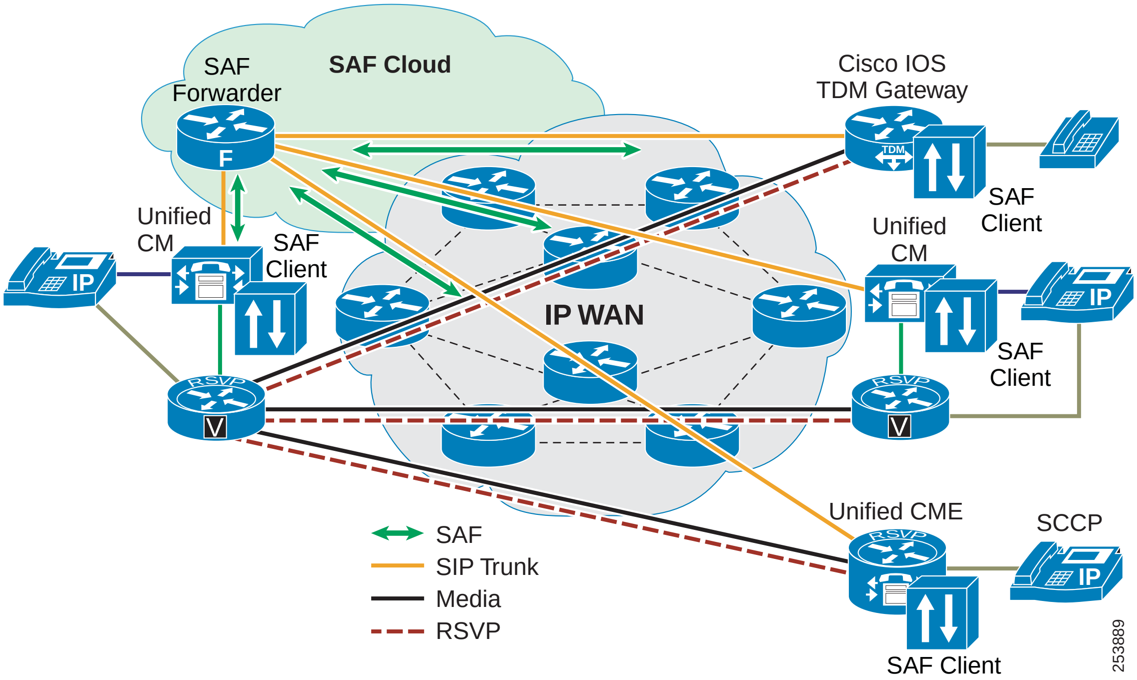
- Follow the guidelines stipulated for Figure 11-66 for Unified CM in distributed call processing deployments, and follow the guidelines stipulated for Figure 11-67 for Unified CME and SIP Cisco IOS TDM Gateways.
- Ensure that Services Advertisement Framework (SAF) and Call Control Discovery (CCD) are enabled in the network and functioning across the deployed products: For details on SAF configuration, refer to the following documents:
  - *Cisco IOS Service Advertisement Framework Configuration Guide*

    http://www.cisco.com/en/US/docs/ios/saf/configuration/guide/15_0/saf_15_0_book.html
  - *Cisco Unified Communications Manager Features and Services Guide*

    http://www.cisco.com/en/US/products/sw/voicesw/ps556/prod_maintenance_guides_list.html
- Enable RSVP SIP Preconditions on the SAF-enabled SIP trunk (see Migration from Enhanced Locations Call Admission Control to RSVP SIP Preconditions, page 11-78, for the steps).
- Enable RSVP SIP Preconditions on the SAF-enabled SIP trunk. Ensure that only SAF-enabled SIP trunks are used with Call Control Discovery in an RSVP SIP Preconditions deployment. SAF-enabled H.323 trunks will not function in an RSVP SIP Preconditions deployment.
- Ensure that an inter-location RSVP policy of **Mandatory** or **Mandatory (video desired)** is set between the SAF-enabled SIP trunk location and all locations, including itself (see the intra-location RSVP policy below). This ensures that the SIP preconditions will be enabled for all calls engaged on the SIP trunk to and from the SAF network.

- Ensure that an intra-location RSVP policy of **Mandatory** or **Mandatory (Video Desired)** is set on the SAF-enabled SIP trunk. An intra-location RSVP policy is set by selecting the location and configuring a policy for itself. This effectively ensures that calls within this location will engage RSVP call admission control. This is important for calls hairpinned on the trunk from call transfers or forwards back to the originating cluster.
- When using SAF-enabled SIP trunks in RSVP SIP Preconditions environments, Cisco recommends ensuring that CCD Automatic Failover is enabled and that calls that fail call admission control are routed to a local route group. For more information see the section on Call Control Discovery Automatic PSTN Failover as well as the DP chapter Local Route Group.

Figure 11-69 All Components Capable of RSVP SIP Preconditions via Service Advertisement Framework (SAF) and Call Control Discovery (CCD)



Call Admission Control Design Recommendations for TelePresence Video Interoperability Architectures

Video interoperability refers to the support for point-to-point video calls between Cisco TelePresence endpoints, Cisco Unified Communications video endpoints, and third-party video endpoints without requiring a Multipoint Control Unit (MCU). This section discusses the features in Enhanced Locations CAC and the design considerations and recommendations applicable to Quality of Service (QoS) for interoperable video calls.

This section explains how to supplement the design to allow for video interoperation between Unified Communications and TelePresence endpoints. For information on Unified Communications endpoints, refer to the chapter on Unified Communications Endpoints, page 18-1; and for information on TelePresence endpoints, refer to the Cisco TelePresence Endpoints product documentation at http://www.cisco.com/en/US/products/ps7060/index.html.

Note Third-party video endpoints should follow the same guidelines and recommendations as Cisco Unified Communications endpoints. Throughout this section, the term *UC endpoints* is used to refer to both third-party endpoints and Cisco Unified Communications Endpoints.

Starting with Cisco Unified CM 9.0, the Cisco TelePresence solution provides the ability to reserve network bandwidth and perform admission control for TelePresence calls. It is important to be familiar with Enhanced Locations call admission control (CAC) and RSVP CAC prior to designing TelePresence video interoperability. This section addresses both Enhanced Locations CAC and RSVP with regard to TelePresence video interoperability, and each CAC mechanism has its own benefits, design considerations, and requirements.

Additionally, TelePresence video interoperability in Unified CM enables Cisco Telepresence System (CTS) endpoints to communicate with non-CTS endpoints, provided that the installed CTS software supports such interoperability. For further information, refer to the document on *Interoperability Between CTS Endpoints and Other Cisco Endpoints or Devices*, available at

http://www.cisco.com/en/US/docs/telepresence/interop/endpoint_interop.html

Supported CAC Deployment Scenarios and Design Considerations

The design considerations for TelePresence video interoperability CAC are based on the following main deployment scenarios:

- Mixed Single Cluster — Mixed UC video endpoints and TelePresence endpoints registered to a single cluster

  This is a single-cluster design where TelePresence and UC video endpoints are registered to the same cluster. The call processing deployment model can be any of the single-cluster designs such as clustering over the WAN, multi-site centralized call processing, or single campus designs.

- Dedicated Multi-Cluster — UC video endpoints and TelePresence endpoints on separate dedicated clusters

  This is a multi-cluster design where the TelePresence endpoints are registered to a different cluster than the UC video endpoints. The call processing deployment model can be multi-site centralized or multi-site distributed cluster designs. All releases of Cisco Unified CM 8.*x* support this model for deploying UC and TelePresence in the same enterprise. However, only Unified CM 8.6 and later releases support the new capability for UC video and TelePresence call interoperability in point-to-point video calls without the use of a video MCU.

- Mixed Multi-Cluster — UC video endpoints and TelePresence endpoints mixed in multi-cluster distributed deployments

  This is a multi-cluster design where TelePresence endpoints are spread across multiple clusters serviced by a single Cisco TelePresence System Manager (CTS-Manager). There are also deployment scenarios where some TelePresence endpoints can be co-located with the same Unified CM cluster as the UC endpoints (mixed single cluster model) while other TelePresence endpoints are registered to a dedicated cluster, and both clusters are serviced by a single CTS-Manager for TelePresence. The call processing deployment model can be multi-site, multi-cluster centralized, or multi-site/multi-cluster distributed designs. This is a hybrid model that combines aspects of the dedicated multi-cluster model with mixed single-cluster model where all endpoints are registered to the same cluster.

Cisco Unified Communications Manager Session Management Edition (SME) is also supported in the multi-cluster models. However, because Session Management Edition is a variation of the multi-site distributed call processing deployment model and is typically employed to interconnect large numbers of Unified Communications systems through a single front-end system, there are no specific guidelines for it with regard to TelePresence video interoperability.

Enhanced Locations CAC Design Considerations and Recommendations

When designing Enhanced Locations (E-L) CAC for TelePresence Video Interoperability, follow the design recommendations and considerations listed in this section.

Design Recommendations

The following design recommendations apply to TelePresence video interoperability solutions that employ Enhanced Locations (E-L) CAC:

- E-L CAC for TelePresence video interoperability is supported in all three deployment models: mixed single cluster, dedicated multi-cluster, and mixed multi-cluster.
- When deploying Unified Communications video and TelePresence video interoperability, ensure that the Unified CM service parameter **Use Video Bandwidth Pool for Immersive Video Calls** is set to **false**. This enables the immersive bandwidth pool for TelePresence calls.
- In E-L CAC TelePresence endpoints can be managed in the same location as Unified Communications video endpoints. If TelePresence calls are not to be tracked through E-L CAC, then set the immersive location and links bandwidth pool to **unlimited**. This will ensure that CAC will not be performed on TelePresence or SIP trunks classified as immersive. If TelePresence calls are to be tracked through E-L CAC, then set immersive location and links bandwidth pool to a value according to the bit rate and number of calls to be allowed over the locations and link paths.
- E-L CAC performs call admission control end-to-end on location pairs; therefore, cross-cluster call transfers and forwards do not require QSIG tunneling with path replacement in order to perform end-to-end E-L CAC. However, Cisco recommends using QSIG path replacement when possible to optimize the call signaling path because this diminishes the number of call signaling legs in complex call forwarding or transfer scenarios.
- Intercluster SIP trunks should be associated with the shadow location.

- Only point-to-point video calls are supported between UC and TelePresence endpoints. No ad-hoc conferencing is supported unless a video MCU is available.
- Cisco Unified CM uses two different cluster-wide QoS service parameter to differentiate between the Differentiated Services Code Point (DSCP) settings of UC video endpoints and TelePresence endpoints. TelePresence endpoints use the **DSCP for Telepresence calls** QoS parameter while the Cisco UC video endpoints use the **DSCP for video calls** QoS service parameter.
- For sites that deploy only UC endpoints and no TelePresence endpoints, ensure that the CS4 DSCP class is added to the AF41 QoS traffic class on inbound WAN QoS configurations to account for the inbound CS4 marked traffic, thus ensuring QoS treatment of CS4 marked media.
- For sites that deploy only UC TelePresence endpoints and no UC endpoints, ensure that the AF41 DSCP class is added to the CS4 QoS traffic class on inbound WAN QoS configurations to account for the inbound AF41 marked traffic, thus ensuring QoS treatment of AF41 marked media.

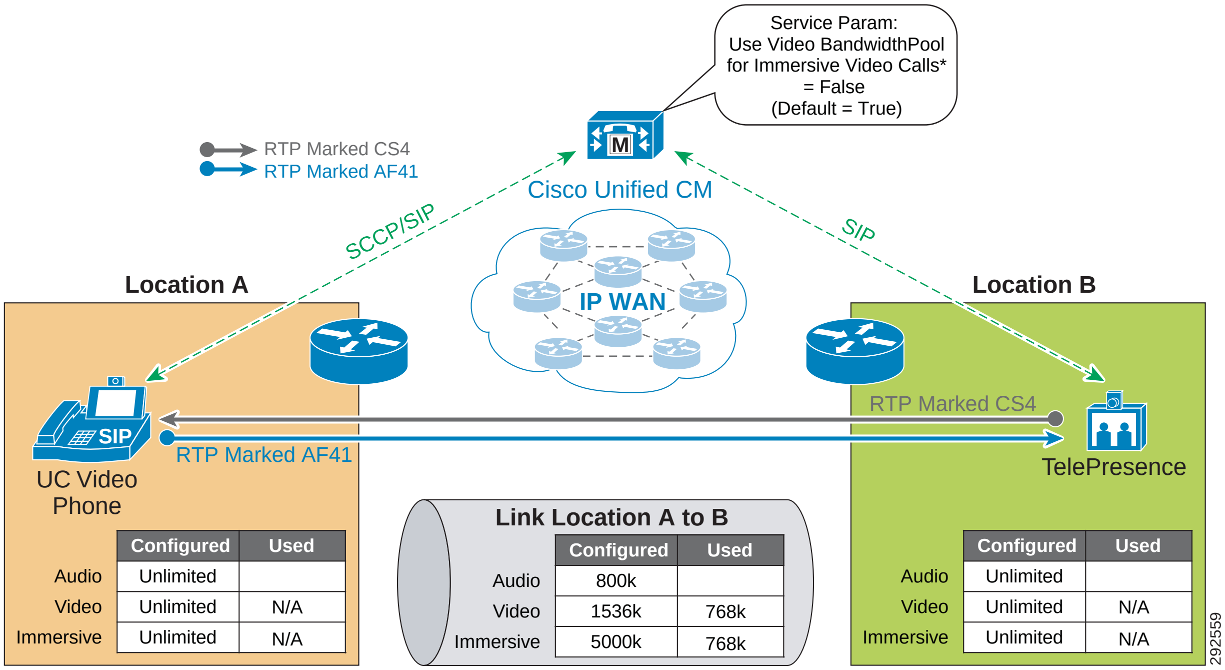
Design Considerations

When deploying Enhanced Locations CAC for TelePresence video interoperable calls, consider the affects of DSCP marking for both QoS classes.

DSCP QoS Marking

The Differentiated Services Code Point (DSCP) QoS markings for TelePresence video interoperable calls are asymmetric, with AF41 used for the UC endpoints and CS4 for the TelePresence endpoints. AF41 and CS4 are default configurations in Unified CM, and changes to these defaults should align with the QoS configuration in the network infrastructure, as applicable. TelePresence endpoints mark video calls with a DSCP value of CS4, which is consistent with the default **DSCP for Telepresence calls** setting. UC endpoints mark calls with a DSCP value of AF41, which is consistent with the default **DSCP for Video calls** setting. Figure 11-70 illustrates the media marking and bandwidth accounting.

Figure 11-70 Bandwidth Deductions and Media Marking in a Multi-Site Deployment with Enhanced Locations CAC



Location A

	Configured	Used
Audio	Unlimited	
Video	Unlimited	N/A
Immersive	Unlimited	N/A

Link Location A to B

	Configured	Used
Audio	800k	
Video	1536k	768k
Immersive	5000k	768k

Location B

	Configured	Used
Audio	Unlimited	
Video	Unlimited	N/A
Immersive	Unlimited	N/A

Bandwidth Accounting for TelePresence Video Interoperability Calls

Enhanced Locations CAC for TelePresence-to-UC video interoperable calls deducts bandwidth from both the video and immersive locations and links bandwidth pools, as illustrated in Figure 11-70. This is by design to ensure that both types of QoS classified streams have the bandwidth required for media in both directions of the path between endpoints.

Enhanced Locations CAC accounts for the bidirectional media of both AF41 and CS4 class traffic. In asymmetrically marked flows, however, the full allocated bit rate of the AF41 class is used in one direction but not the other. In the other direction, the full allocated bit rate is marked CS4. This does not represent additional bandwidth consumption but simply a difference in marking and queuing in the network for each QoS class. This manner of bandwidth accounting is required to protect each flow in each direction.

For more information on the call flows for Enhanced Locations CAC and TelePresence interoperable calls, see the section on Enhanced Locations CAC for TelePresence Immersive Video, page 11-32.

RSVP CAC Design Considerations and Recommendations

In an RSVP solution for TelePresence video interoperability, the goal is to ensure RSVP CAC for both end-to-end UC endpoint calls and TelePresence video interoperability calls, while also ensuring that end-to-end TelePresence calls never invoke RSVP (RSVP Agent) because it is not currently supported and can cause specific TelePresence features to fail.

TelePresence video interoperability in RSVP deployments is very easy to achieve and provides better CAC support than Enhanced Locations CAC solutions, provided that a few design rules and deployment models are adhered to. The supported designs are the mixed single cluster and dedicated multi-cluster designs. The mixed multi-cluster designs are not recommended due to the configuration complexity of ensuring that end-to-end TelePresence calls across clusters do not invoke RSVP and that all other interoperable point-to-point call scenarios do invoke RSVP.

For details about RSVP architecture and general design and deployment considerations, see the section on Unified Communications Architectures Using Resource Reservation Protocol (RSVP), page 11-41. It is important to understand the RSVP principles and solution prior to reading this section independently as this section will uniquely cover recommendations and considerations specific to Telepresence video interoperability with regards to RSVP.

Design Recommendations

The following design recommendations apply to TelePresence video interoperability solutions that employ RSVP for call admission control:

- Locations-based RSVP is supported in mixed single cluster and dedicated multi-cluster designs. As mentioned previously, a mixed multi-cluster design can be used but at the cost of complex configuration of trunks, dial plan, and locations-based RSVP policy. Therefore, mixed multi-cluster designs are not recommended and will not be treated in these design recommendations.
- Both Local RSVP and RSVP SIP Preconditions are supported for intercluster calls, provided they are designed and configured according to the guidelines in the relevant sections of this chapter.

- When designing solutions for TelePresence video interoperability with RSVP, ensure that TelePresence-to-TelePresence calls never invoke RSVP because that functionality is not currently supported and can cause specific TelePresence features to fail. However, for TelePresence calls to/from UC video endpoints, RSVP should be invoked. To achieve this, ensure the following:
  - TelePresence endpoints are in CAC locations separate from UC endpoints.
  - TelePresence locations set their RSVP policy to **no reservation** for the location pairing with other TelePresence locations as well as their own location. For more information on RSVP Policy for location pairs, see the section on RSVP Policy, page 11-69.
  - TelePresence locations set their RSVP policy to **mandatory** or **mandatory video desired** for the location pairings with UC video endpoint locations.
  - For dedicated multi-cluster deployments, intercluster trunks require an RSVP policy pairing between UC endpoint locations as well as TelePresence endpoint locations.

    For dedicated TelePresence clusters, intercluster trunks pointing to dedicated UC endpoint clusters should have an RSVP policy location pairing with TelePresence endpoints set to **mandatory** or **mandatory video desired**. Intercluster trunks pointing to TelePresence clusters should have an RSVP policy location pairing with TelePresence endpoints set to **no reservation**.

    For dedicated UC endpoint clusters, intercluster trunks pointing to both dedicated UC clusters and TelePresence clusters should have an RSVP policy location pairing with UC endpoints set to **mandatory** or **mandatory video desired**.

Design Considerations

When deploying RSVP CAC for TelePresence video interoperable solutions, consider the following major factors:

RSVP Performs CAC for Both UC and TelePresence Endpoints

Unlike Enhanced Locations CAC, RSVP performs CAC for both the TelePresence and UC endpoints for TelePresence video interoperating calls. It also overwrites the endpoint QoS marking, thus ensuring the correct QoS marking from agent to agent. Figure 11-71 illustrates this point

Figure 11-71 Bandwidth Deductions and Media Marking in a Multi-Site Deployment with RSVP CAC

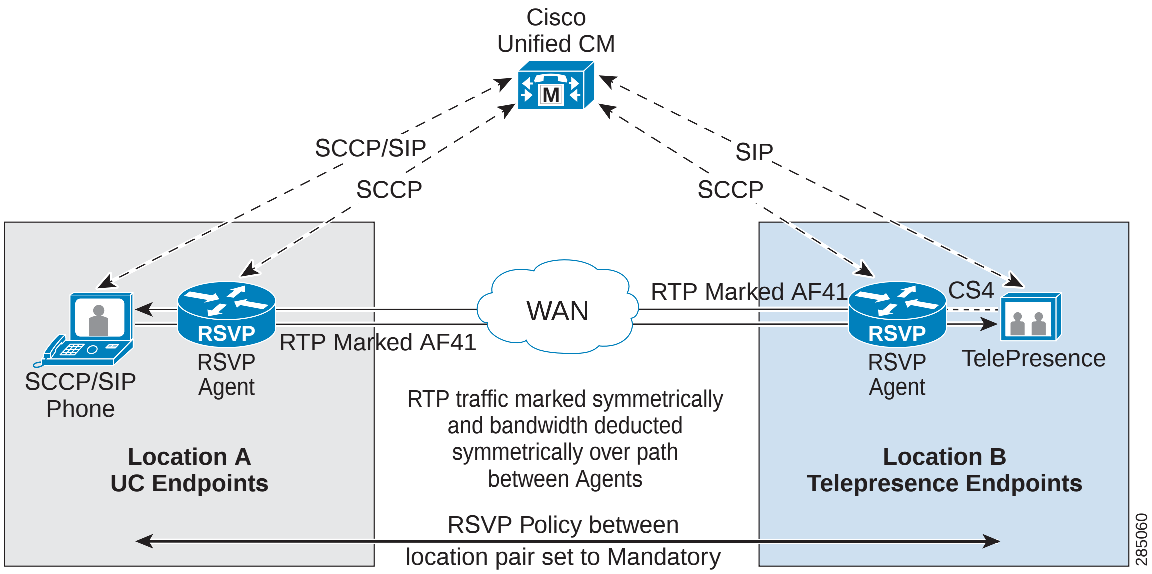


Figure 11-71 illustrates a call between a TelePresence endpoint and a Cisco Unified IP Phone 9900 Series video phone, where RSVP Agents are invoked for RSVP CAC. Two salient points are illustrated here:

- RSVP Agents re-mark the RTP media traffic so that the media is marked symmetrically between RSVP Agents with a DSCP value of AF41, which is consistent with the **DSCP for Video calls** setting. This provides symmetrically marked RTP traffic between endpoints, which is something that the locations CAC solution cannot achieve.
- RSVP deducts bandwidth over the media path between the RSVP Agents for the UC endpoint and the TelePresence endpoint. This provides CAC for audio and video streams in both directions for TelePresence video interoperability calls. This is also something that locations CAC inherently cannot achieve.

RSVP Should Not Be Used on Calls Between Endpoints Located in the Same Physical Site in a Mixed Single-Cluster Deployment

When deploying mixed single-cluster designs, where TelePresence endpoints and UC video endpoints are registered to the same Unified CM cluster and located in the same physical site or campus, it is important not to engage RSVP for calls between these devices. The RSVP policy location pair between the UC endpoint's location and the TelePresence endpoint's location should be set in these cases to **no reservation**. This is illustrated in Figure 11-72.

Figure 11-72 RSVP Policy Setting for Calls Between UC and TelePresence Endpoints in the Same Site

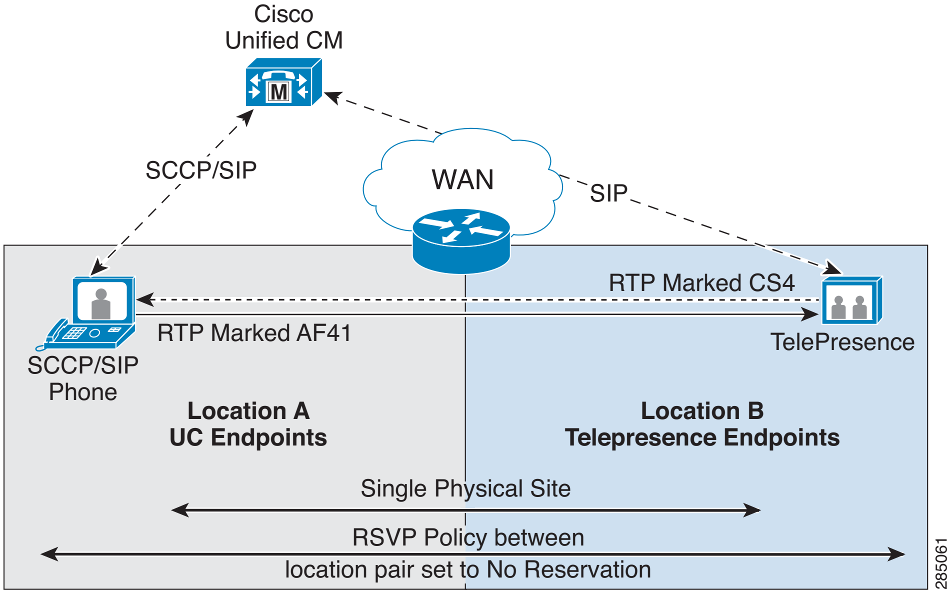
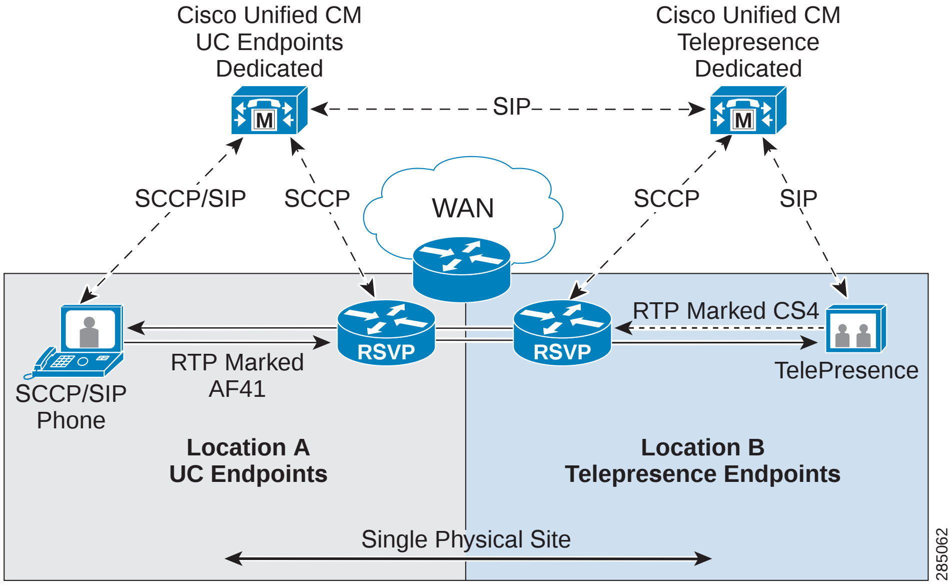


Figure 11-72 illustrates a TelePresence endpoint and a Cisco Unified IP Phone 9900 Series video phone in the same physical site but in separate CAC locations. The location pair policy is set to **no reservation** so that calls between TelePresence and UC endpoints in the same physical site or campus do not invoke RSVP. Note that the RTP streaming is asymmetrical. This will be the case but is usually inconsequential over the LAN.

RSVP Should Be Used on Calls Between Endpoints in Different Clusters

For dedicated multi-cluster deployments it is necessary to invoke RSVP even when the UC endpoint and the TelePresence endpoint are located in the same physical site. This will utilize RSVP Agent resources but the media will not be routed over the WAN uplink. Figure 11-73 illustrates this point.

Figure 11-73 Bandwidth Deductions and Media Marking in a Multi-Cluster, Single-Site Deployment with RSVP CAC



Note that the RSVP Agents as well as the WAN edge router in Figure 11-73 could all be either co-located or separated as depicted. For information on co-locating multiple RSVP Agents registered to separate clusters, see the section on Multiple Clusters Sharing a Single Platform for RSVP Agent, page 11-124.

Guidelines for Session Management Edition Deployments

For Session Management Edition deployments, follow the same guidelines that are stipulated in the section on Unified CM Session Management Edition with RSVP Deployments, page 11-120. Keep in mind, however, that RSVP is not recommended in mixed multi-cluster deployments.

Call Admission Control Design Recommendations for Unified CM Session Management Edition Deployments

Cisco Unified Communications Manager Session Management Edition (SME) has two main forms of call admission control available to it for admitting trunk-to-trunk audio and video calls: one is Enhanced Locations call admission control (CAC) and the other is Resource Reservation Protocol (RSVP) using RSVP-enabled Locations. This section covers design guidelines and best practices specific to Session Management Edition deployments. It does not cover the basic functions of Enhanced Locations CAC, Locations-based RSVP, or RSVP SIP Preconditions. Cisco highly recommends that you become familiar with the applicable sections of this chapter as a prerequisite to understanding the following design guidelines.

Session Management Edition with Enhanced Locations CAC

Unified CM Session Management Edition (SME) is typically used for interconnecting multiple Unified CM clusters, third-party UC systems (IP- and TDM-based PBXs), PSTN connections, and centralized UC applications as well as for dial-plan and trunk aggregation. The following is a list of recommendations and design considerations to follow when deploying Unified CM SME with Enhanced Locations (E-L) CAC. For more information on Unified CM SME, refer to the *Cisco Unified Communications Manager Session Management Edition Deployment Guide*, available at

http://www.cisco.com/en/US/products/ps10661/products_implementation_design_guides_list.html

Recommendations and Design Considerations

- All leaf clusters that support E-L CAC should be enabled for intercluster E-L CAC with SME.
- SME can be used as a centralized bootstrap hub for the E-L CAC intercluster hub replication network. See LBM Hub Replication Network, page 11-23, for more information.
- All trunks to leaf clusters supporting E-L CAC should be SIP trunks placed in the shadow location to enable E-L CAC on the trunk between SME and the leaf clusters supporting E-L CAC.
- For TelePresence video interoperability, see the section on Call Admission Control Design Recommendations for TelePresence Video Interoperability Architectures, page 11-109.
- Connectivity from SME to any trunk or device other than a Unified CM that supports E-L CAC (some examples are third-party PBXs, gateways, Unified CM clusters prior to release 9.0 that do not support E-L CAC, voice messaging ports or trunks to conference bridges, Cisco Video Communications Server, and so forth) should be configured in a location other than a phantom or shadow location. The reason for this is that both phantom and shadow locations are non-terminating locations; that is, they relay information about locations and are effectively placeholders for user-defined locations on other clusters. Phantom locations are legacy locations that allow for the transmission of location information in versions of Unified CM prior to 9.0, but they are not supported with Unified CM 9.*x* Enhanced Locations CAC. Shadow locations are special locations that enable trunks between Unified CM clusters that support E-L CAC to accomplish it end-to-end.
- SME can be used as a locations and link management cluster. See Figure 11-74 as an example of this.
- SME can support a maximum of 2,000 locations configured locally.

Figure 11-74 Unified CM SME as a Location and Link Management Cluster

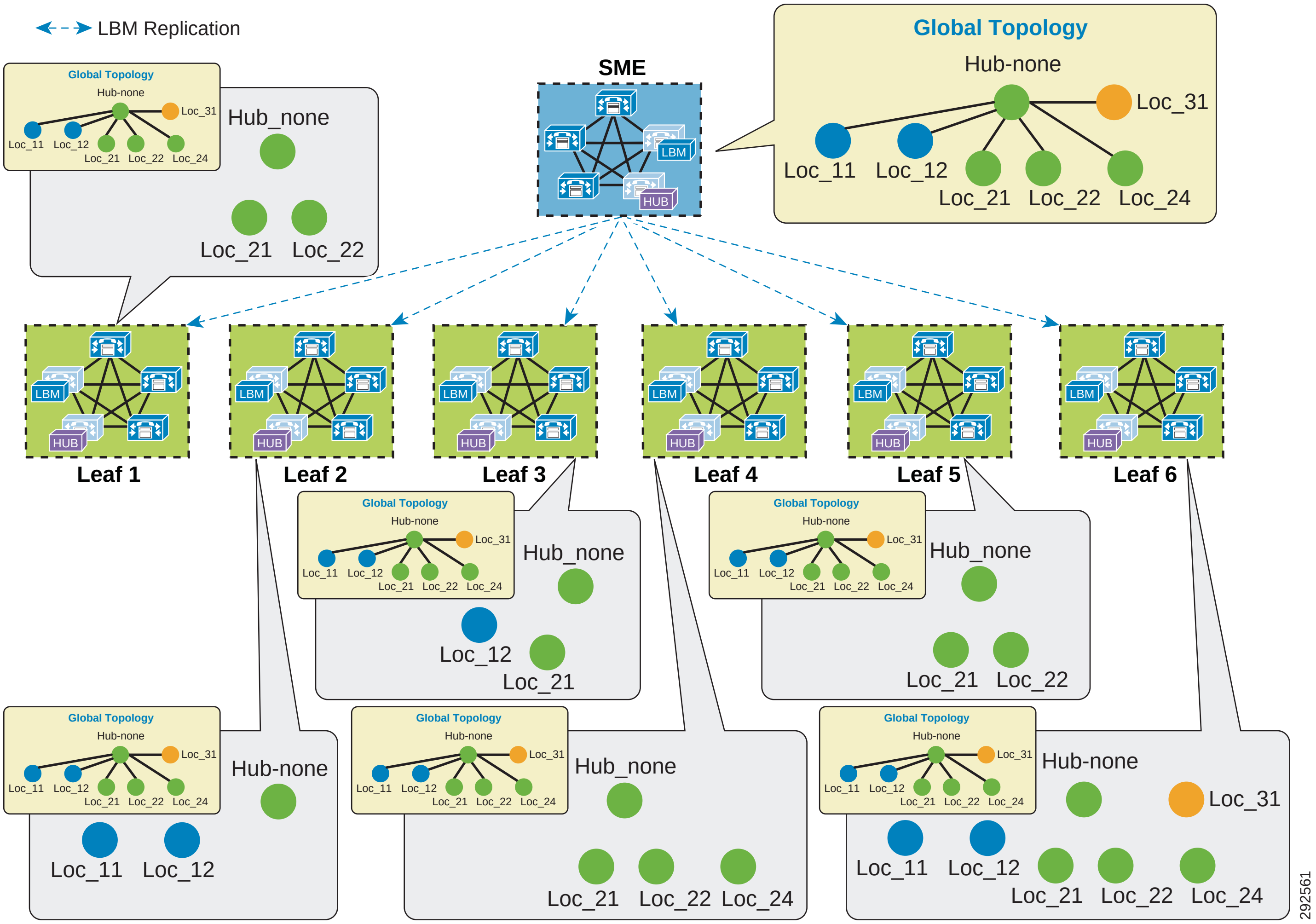
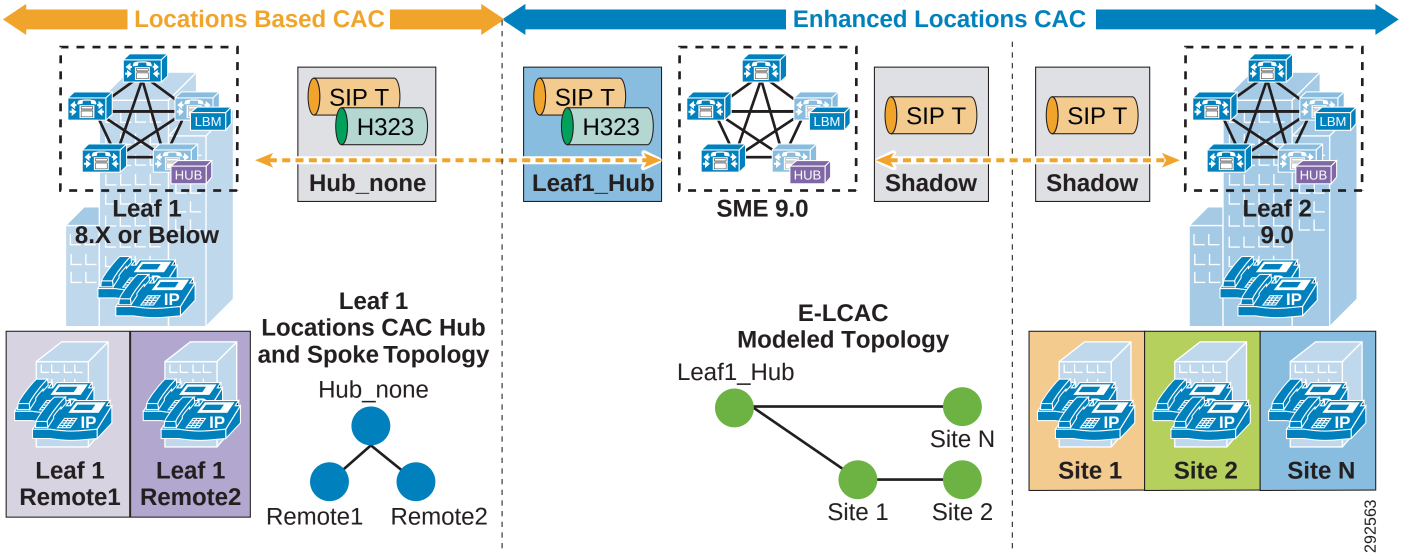


Figure 11-74 illustrates SME as a location and link management cluster. The entire location and link global topology is configured and managed in SME, and the leaf clusters configure locally only the locations that they require to associate with the end devices. When intercluster E-L CAC is enabled and locations and links are replicated, each leaf cluster will receive the global topology from SME and overlay this on their configured topology and use the global topology for call admission control. This simplifies configuration and location and link management across multiple clusters, and it diminishes the potential for misconfiguration across clusters. For more information and details on the design and deployment see the section on Location and Link Management Cluster, page 11-28.

Figure 11-75 illustrates an SME design where intercluster E-L CAC has been enabled on one or more leaf clusters (right) and where one or more leaf clusters are running a version of Unified CM prior to 9.0 and are running traditional locations CAC (left). In this type of a deployment the locations managed by traditional locations CAC cannot be common or shared locations between E-L CAC-enabled clusters. Leaf 1 has been configured in a traditional hub and spoke, where devices are managed at various remote sites. SME and the other leaf clusters that are enabled for intercluster E-L CAC share a global topology, as illustrated in the E-L CAC Modeled Topology. Leaf1_Hub is a user-defined location in SME assigned to the SIP or H.323 intercluster trunk that represents the hub of the Leaf 1 topology. This allows SME to deduct bandwidth for calls to and from Leaf 1 up to the Leaf1_Hub. In this way SME and Leaf 2 manage the E-L CAC locations and links while Leaf 1 manages its remote locations with traditional locations CAC.

Figure 11-75 SME Design with Enhanced Locations CAC and Traditional Locations CAC in Leaf Clusters



QSIG Path Replacement Over Intercluster Trunks

In addition to providing features such as Call Back (on Busy/No Reply) between phones in Unified CM leaf clusters, QSIG also provides path replacement, which optimizes call signaling between clusters when, for example, a call is transferred or forwarded from one cluster to another. With E-L CAC, end-to-end QSIG path replacement is not required because E-L CAC deducts bandwidth over the correct locations and links path end-to-end between clusters. Nonetheless, QSIG path replacement is beneficial in optimizing the signaling path between clusters that do not support E-L CAC, as well as for third-party PBXs and gateways where supported, in order to avoid hair-pinning the trunk signaling with every forward or transfer off the cluster, PBX, or gateway. For more information on QSIG over intercluster trunks, refer to the *Cisco Unified Communications Manager Session Management Edition Deployment Guide*, available at

http://www.cisco.com/en/US/products/ps10661/products_implementation_design_guides_list.html

Unified CM Session Management Edition with RSVP Deployments

Unified CM Session Management Edition can be deployed in a number of RSVP deployments either with SIP Preconditions support between clusters or without it. For an overview of RSVP in Unified Communications designs and as a prerequisite to understanding the following content, see the section on Unified Communications Architectures Using Resource Reservation Protocol (RSVP), page 11-41, which includes detailed sections on Unified CM RSVP-enabled Locations and RSVP SIP Preconditions.

Session Management Edition Design with Leaf Clusters without RSVP SIP Precondition Support

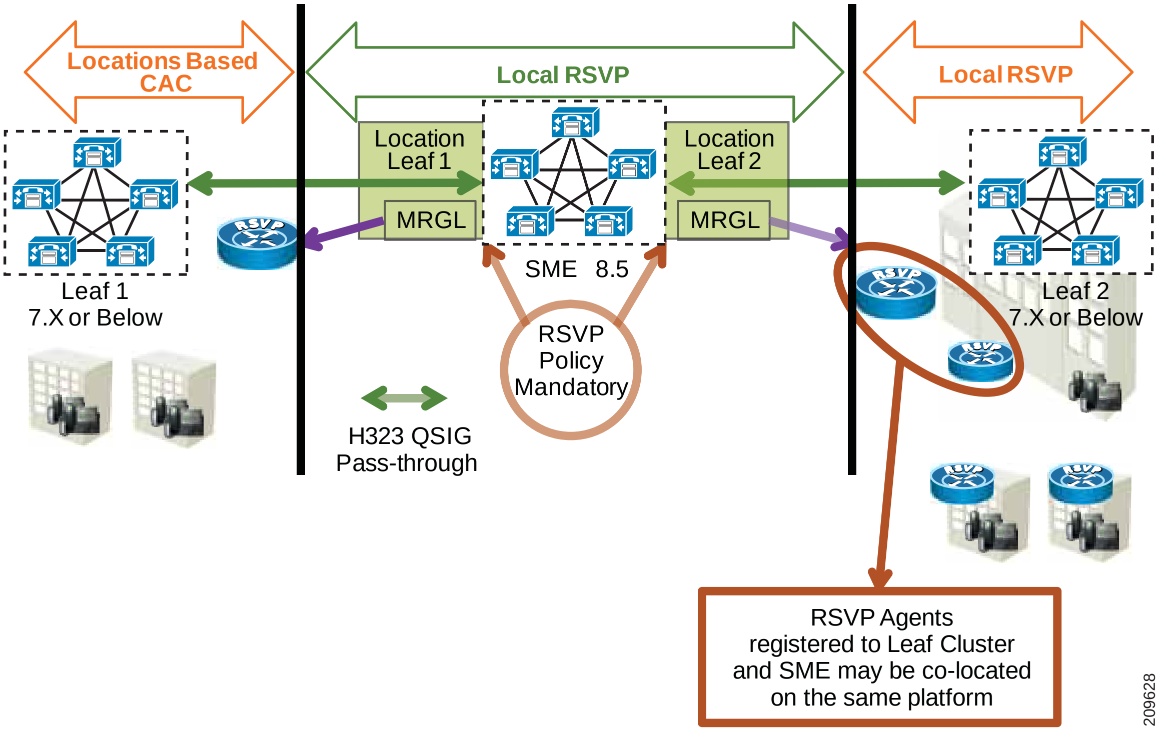
Unified CM Session Management Edition can be deployed with leaf clusters that maintain their own CAC support locally within the leaf cluster. In such deployments where the leaf clusters manage their own CAC mechanism for their own devices in remote sites, or where those leaf clusters are in a campus deployment and do not require CAC for intra-cluster calls (calls that remain within the leaf cluster), Session Management Edition can be leveraged to manage CAC for audio and video calls between those leaf clusters (intercluster calls) with RSVP. In these cases Session Management Edition is leveraging the Cisco RSVP Agents to handle CAC across the links between the leaf clusters but is not managing CAC for the links that the leaf clusters manage for their intra-cluster calls.

Requirements

- Intra-cluster calls are managed by the leaf cluster CAC.
- Intercluster calls are managed by Session Management Edition RSVP-enabled Locations CAC.
- Cisco highly recommends that the WAN bandwidth managed by Session Management Edition for intercluster calls should not be shared with the WAN bandwidth for intra-cluster calls managed by the leaf clusters. If the bandwidth from the same WAN links is managed by two separate CAC mechanisms, then there is the potential for double bandwidth counting because the two separate CAC mechanisms are not aware of each other.

Figure 11-76 depicts two leaf clusters deployed in different regional sites connecting to one another through Session Management Edition for intercluster connectivity. Leaf cluster 1 is using Locations CAC to manage any remote sites in its CAC domain, while Leaf cluster 2 is using RSVP to manage any remote sites under its CAC domain. Session Management Edition in this case leverages RSVP-enabled Locations to manage the CAC domain between leaf clusters using remote RSVP Agents. The remote RSVP Agents are simply standard RSVP Agents registered with Session Management Edition and associated to the leaf cluster trunk(s) through media resource group (MRG) and media resource group list (MRGL) functions, but they are physically located at a campus site and/or co-located with the leaf cluster and thus could be "remote" from the Session Management Edition cluster. These RSVP Agents should be at a head-end network WAN that interconnects the leaf clusters and is enabled to support RSVP.

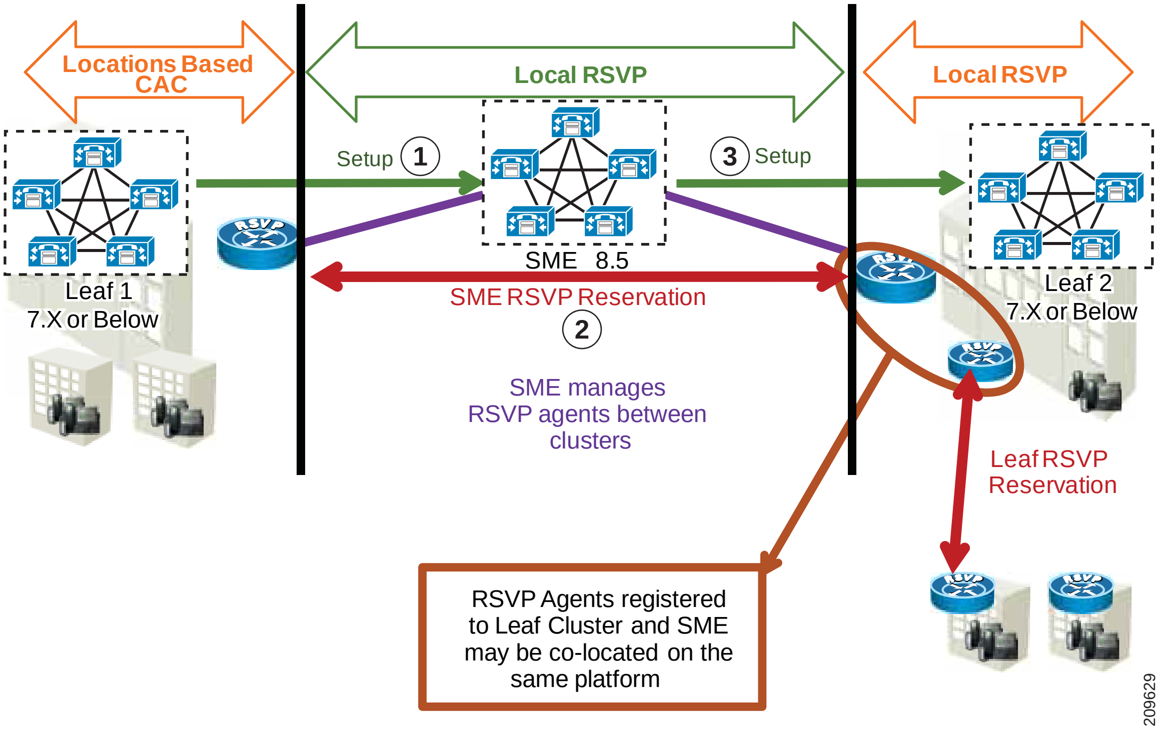
Figure 11-76 Session Management Edition Deployment with Leaf Clusters without SIP Precondition Support



In a call setup between Leaf 1 and Leaf 2, the call flow works as follows (see Figure 11-77):

1. Leaf 1 sets up a call to Session Management Edition. This is done over an H.323 or SIP intercluster trunk. QSIG Path Replacement support is preferred for H.323 prior to Unified CM 8.5 and SIP with Unified CM 8.5 and later releases.
2. Session Management Edition receives the call from Leaf 1 and allocates two RSVP Agents. The RSVP Agents are associated to the trunks through the media resource group and list functions (see Figure 11-76). Once allocated, the RSVP Agents reserve the bandwidth over the path between them.
3. If the bandwidth request is successful, the call is extended from Session Management Edition to Leaf 2 over the SIP or H.323 intercluster trunk. If the reservation fails, then the call is not extended to Leaf 2 and, depending on the configuration of call processing in Session Management Edition, the call could be extended elsewhere or torn down from Leaf 1.

Figure 11-77 Call Flow for Session Management Edition Deployment with Leaf Clusters without SIP Precondition Support



The following notes apply to Figure 11-77:

- If the call is extended to Leaf 2 in step 3, then Leaf 2 can use Locations CAC or RSVP as the admission control for that call leg. If using RSVP as in Figure 11-77, then Leaf 2 will associate an RSVP Agent to the SIP trunk pointing to Session Management Edition and will associate another RSVP Agent with the called party endpoint.
- In cases where the leaf cluster is doing RSVP locally and not with SIP Preconditions, the Session Management Edition Remote RSVP Agent and the leaf cluster RSVP Agent associated with the trunk to Session Management Edition can be co-located on the same routing platform. See Multiple Clusters Sharing a Single Platform for RSVP Agent, page 11-124, for further information.

Session Management Edition Design with Leaf Clusters with SIP Precondition Support

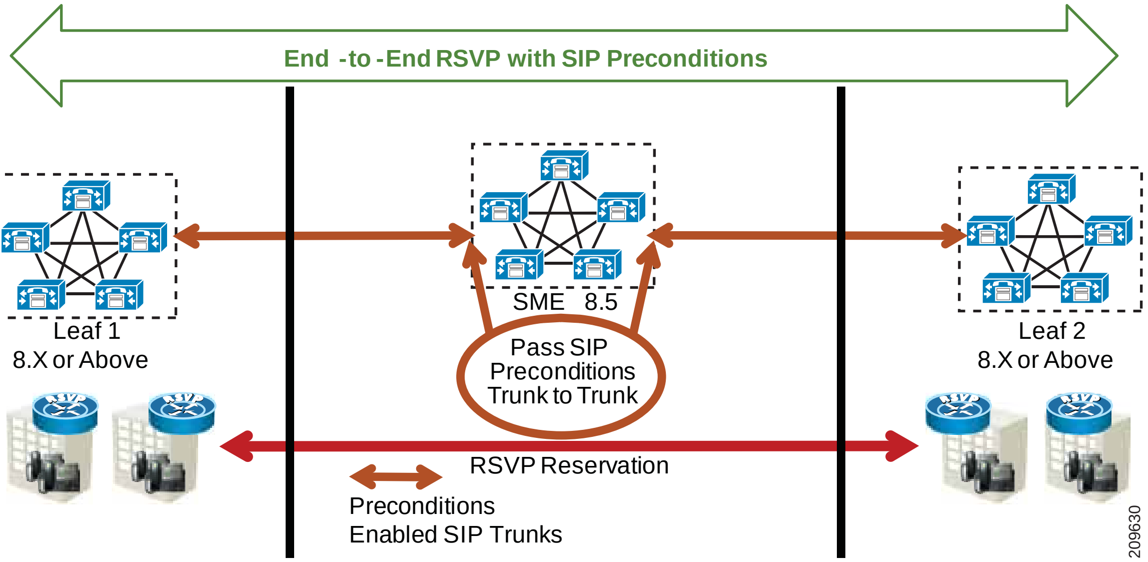
Unified CM Session Management Edition can also be deployed with leaf clusters that manage their CAC domain with RSVP and support RSVP SIP Preconditions. Session Management Edition in this case enables its intercluster trunks with RSVP SIP Preconditions and passes the SIP Preconditions from leaf cluster to leaf cluster, thus providing an end-to-end RSVP Agent deployment.

Requirements

- Leaf clusters with Unified CM 8.0 and later releases (preferably Unified CM 8.6.1 or later)
- Intra-cluster calls managed by the leaf cluster using Unified CM RSVP-enabled Locations
- Intercluster calls managed by Session Management Edition with SIP intercluster trunks (ICTs) configured to pass RSVP SIP Preconditions from leaf cluster to leaf cluster
- QSIG with path-replacement is not required, but Cisco recommends QSIG with path-replacement enabled on all SIP intercluster trunks (ICTs) to optimize call signaling for transferred and forwarded calls (requires Unified CM 8.5 or later release).

Figure 11-78 depicts the described solution where leaf clusters have Unified CM RSVP-enabled Locations for intra-cluster calls, and Session Management Edition as well as the SIP ICTs configured on both of the leaf clusters are enabled with RSVP SIP Preconditions. Session Management Edition in turn passes the SIP Preconditions along from leaf Unified CM to leaf Unified CM, which in turn pass those preconditions down to the branch RSVP Agents, thus providing an end-to-end RSVP reservation directly from branch to branch across multiple cluster boundaries.

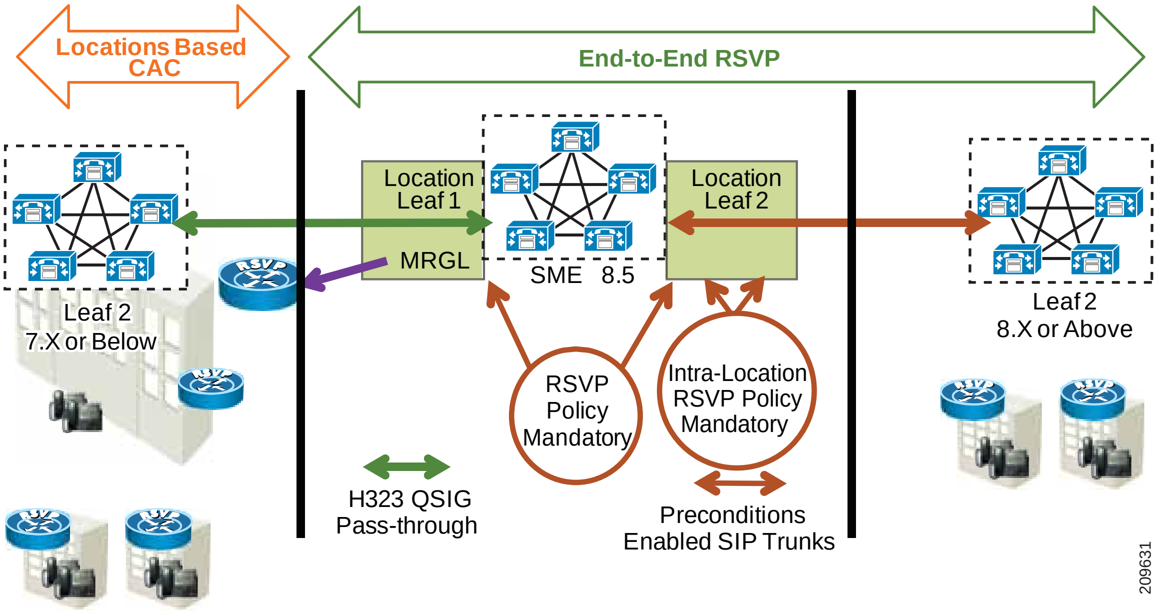
Figure 11-78 Session Management Edition Deployment with Leaf Clusters with SIP Precondition Support



Session Management Edition Design with Mixed Leaf Clusters (with and without SIP Precondition Support)

Unified CM Session Management Edition can also support a mixed environment where on one trunk it is supporting RSVP SIP Preconditions on the trunk to a leaf cluster that supports RSVP locally and RSVP SIP Preconditions (see example Leaf 2 in Figure 11-79) and on another trunk Session Management Edition associates an RSVP Agent and is doing RSVP locally to the trunk to a cluster that does not support RSVP SIP Preconditions (see Leaf 1 in Figure 11-79).

Figure 11-79 Session Management Edition Design with Mixed Leaf Clusters (with and without SIP Precondition Support)



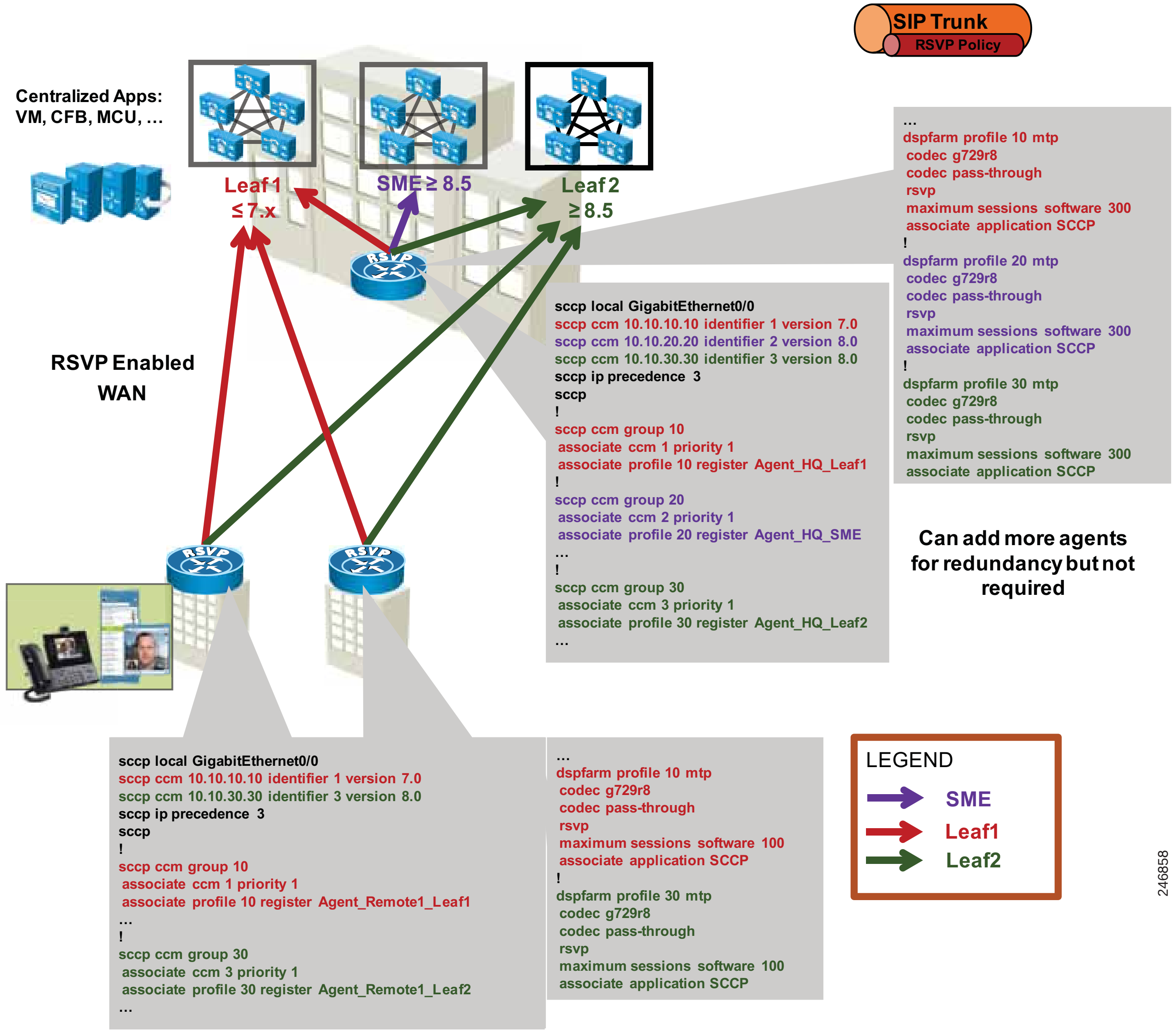
Multiple Clusters Sharing a Single Platform for RSVP Agent

In some cases there might be the need to share a single router platform supporting RSVP Agent across multiple Unified CM clusters. In this case a single router platform supporting RSVP Agent, such as a Cisco Integrated Services Router (ISR), can be configured with multiple RSVP Agents, with each agent registered to a separate Unified CM cluster, each with dedicated software sessions. For information on the number of supported RSVP Agent sessions per platform, refer to the *Cisco RSVP Agent Data Sheet*, available at

http://www.cisco.com/en/US/products/ps6832/products_data_sheets_list.html

Figure 11-80 illustrates this with a Session Management Edition deployment co-located with two leaf clusters at the headquarters site. In this example one leaf cluster is Unified CM 7.*x* while the other leaf cluster and the Session Management Edition cluster is version 8.5. Each of the three clusters is configured with an RSVP Agent that supports 300 software sessions at the headquarters, and both leaf clusters share a Cisco ISR platform for RSVP Agent supporting 100 sessions each at two remote locations.

Figure 11-80 Session Management Edition Co-Located with Leaf Clusters (Multiple Clusters Sharing a Single Platform)



```
…
dspfarm profile 10 mtp
 codec g729r8
 codec pass-through
 rsvp
 maximum sessions software 300
 associate application SCCP
!
dspfarm profile 20 mtp
 codec g729r8
 codec pass-through
 rsvp
 maximum sessions software 300
 associate application SCCP
!
dspfarm profile 30 mtp
 codec g729r8
 codec pass-through
 rsvp
 maximum sessions software 300
 associate application SCCP
```

```
sccp local GigabitEthernet0/0
sccp ccm 10.10.10.10 identifier 1 version 7.0
sccp ccm 10.10.20.20 identifier 2 version 8.0
sccp ccm 10.10.30.30 identifier 3 version 8.0
sccp ip precedence 3
sccp
!
sccp ccm group 10
 associate ccm 1 priority 1
 associate profile 10 register Agent_HQ_Leaf1
!
sccp ccm group 20
 associate ccm 2 priority 1
 associate profile 20 register Agent_HQ_SME
…
!
sccp ccm group 30
 associate ccm 3 priority 1
 associate profile 30 register Agent_HQ_Leaf2
…
```

```
sccp local GigabitEthernet0/0
sccp ccm 10.10.10.10 identifier 1 version 7.0
sccp ccm 10.10.30.30 identifier 3 version 8.0
sccp ip precedence 3
sccp
!
sccp ccm group 10
 associate ccm 1 priority 1
 associate profile 10 register Agent_Remote1_Leaf1
…
!
sccp ccm group 30
 associate ccm 3 priority 1
 associate profile 30 register Agent_Remote1_Leaf2
…
```

```
…
dspfarm profile 10 mtp
 codec g729r8
 codec pass-through
 rsvp
 maximum sessions software 100
 associate application SCCP
!
dspfarm profile 30 mtp
 codec g729r8
 codec pass-through
 rsvp
 maximum sessions software 100
 associate application SCCP
```

CHAPTER 12

IP Video Telephony

Revised: April 30, 2013; OL-27282-05

Enterprises with Unified Communication systems may have widely deployed voice services such as point-to-point calls and conferencing within the enterprise as well as calls to the PSTN for external connectivity. When adding video to such networks, the following approaches can be used:

- Enable Voice Devices to Support Video Calls, page 12-1
- Integrating Voice Network with Existing Video Network, page 12-1

Enable Voice Devices to Support Video Calls

An enterprise may enable existing IP phones to support cameras or use software on PC-based systems to provide video capability in conjunction with IP phones. Where possible, newer devices that support video can be deployed, thus keeping the existing call control infrastructure intact.

This approach has the following advantages:

- Existing dial plan — Enterprises can use the existing dial plan and existing call agents to support video calls in the network.
- Call admission control — A single call admission control entity in the enterprise provides a way to optimize use of the network bandwidth.
- Existing network — The enterprise can use the existing network topology by adding the needed bandwidth to accommodate video on its network.
- Users — Enterprises can enable all existing users to make video calls.
- Video codec — A standardizing video codec enables the enterprise to optimize bandwidth usage for video calls, thus reducing the need for transcoding or transrating resources.
- Existing IP phones — The enterprise can leverage existing IP phones to add video by using soft clients or by connecting cameras IP phones.

Integrating Voice Network with Existing Video Network

An enterprise that has an existing video network used for conference room video calls or for video devices for executives, can integrate the video network with its voice network, thus enabling enterprise users to call video devices.

This approach has the following advantages:

- Separate dial plans — A separate dial plan for the video network enables enterprises to plan for video resources such as videoconferencing bridges and video PSTN gateways. Users can dial a prefix to access resources from the other network.

- Existing network — The enterprise can use the existing overlapped network topology as separate video and voice networks.
- Management and monitoring — Separate management and monitoring for voice and video make troubleshooting and problem isolation easier.
- Trunk protocol — Enterprises can choose the desired protocol to interwork between the voice and video call agents. This can enable enterprises to use similar methods for call transfers, DTMF, MWI, or other functionality.
- Independence from call agent features — Enterprises can leverage capabilities that video call agents provide for video endpoints, such as optimizing the bit rate of video traffic for calls, and those capabilities may be independent of the features of voice call agents.

An enterprise can choose either of the above approaches or can combine them to achieve its objective of enabling its users to make video calls.

Video is fully integrated into Cisco Unified Communications Manager (Unified CM), and there are also many video endpoints available from Cisco and its strategic partners. For example, Cisco Jabber is just as easy to deploy, manage, and use as a Cisco Unified IP Phone.

What's New in This Chapter

Table 12-1 lists the topics that are new in this chapter or that have changed significantly from previous releases of this document.

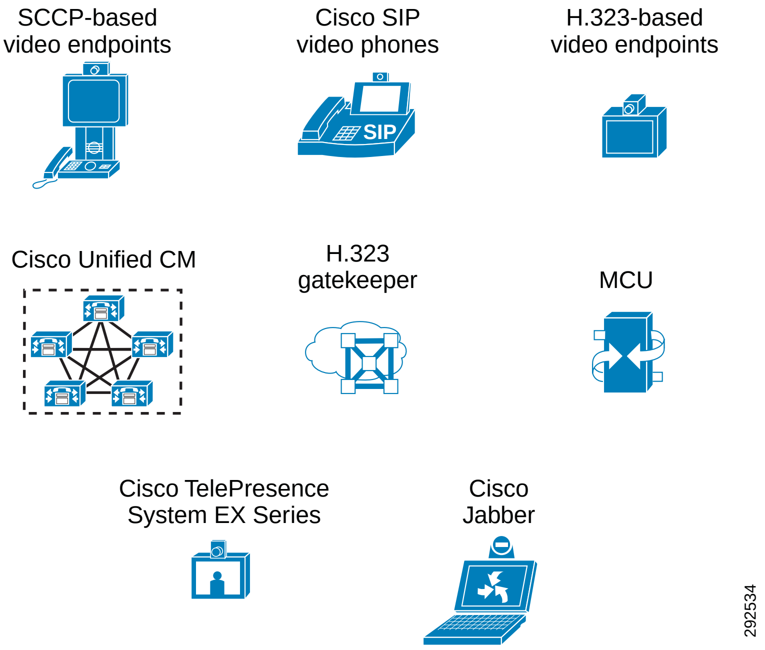
Table 12-1 New or Changed Information Since the Previous Release of This Document

New or Revised Topic	Described in:	Revision Date
Removal of call admission control information from this chapter	See the chapter on Call Admission Control, page 11-1, for this information.	April 30, 2013
Numerous updates for Cisco Unified Communications System Release 9.0	Various sections throughout this chapter	June 28, 2012

IP Video Telephony Solution Components

The Cisco IP Video Telephony solution consists of Cisco Unified Communications Manager (Unified CM); Cisco Multipoint Control Units (MCUs), Session Initiation Protocol (SIP), and Skinny Client Control Protocol (SCCP) conference calls; Cisco H.320 Gateways; Cisco IOS H.323 Gatekeeper; Cisco TelePresence System EX60 and EX90; Cisco Cius; Cisco Unified IP Phones 9900 Series; Cisco Unified IP Phones with video capabilities (for example, Cisco Unified IP Phone 9900 Series); Cisco Jabber; third-party SCCP video endpoint solutions; and the existing range of H.323 or SIP-compliant products from partners such as Polycom, Lifesize, Sony, and others. (See Figure 12-1.)

Figure 12-1 IP Video Telephony Components


SCCP-based video endpoints
Cisco SIP video phones
SIP
H.323-based video endpoints
Cisco Unified CM
H.323 gatekeeper
MCU
Cisco TelePresence System EX Series
Cisco Jabber
292534

Administration Considerations

This section discusses the following configuration elements in Unified CM Administration that pertain to Video Telephony:

- Protocols, page 12-4
- Endpoints, page 12-5
- Regions, page 12-6
- Call Admission Control, page 12-8
- Quality of Service, page 12-9
- Retry Video Call as Audio, page 12-10
- Dial Plan, page 12-11
- Trunks, page 12-11
- Security, page 12-11

Protocols

Although Cisco Unified CM supports a large number of protocols, SIP is the preferred call control protocol when using video with Unified CM. Any device can call any other device, but video is supported only on SCCP, H.323, and SIP devices. Specifically, video is not supported in the following protocols in Cisco Unified CM Release 9.*x*:

- Computer Telephony Integration (CTI) applications (TAPI and JTAPI)
- Media Gateway Control Protocol (MGCP)

Therefore, Unified CM currently supports the types of calls listed in Table 12-2.

Table 12-2 Types of Calls Supported in Unified CM Release 9.x

Calling Device Type	Called Device Type				
	SCCP	**H.323**	**MGCP**	**TAPI/JTAPI**	**SIP**
SCCP	Audio and video	Audio and video	Audio only	Audio only	Audio and video
H.323	Audio and video	Audio and video	Audio only	Audio only	Audio and video
MGCP	Audio only	Audio only	Audio only	Audio only	Audio only
TAPI/JTAPI	Audio only	Audio only	Audio only	Audio only	Audio only
SIP	Audio and video	Audio and video	Audio only	Audio only	Audio and video

Table 12-3 lists the audio and video algorithms and protocols currently supported in Unified CM.

Table 12-3 Capabilities Supported in Unified CM Release 9.x

H.323	SCCP	SIP
H.261	H.261	H.261
H.263, H.263+	H.263, H.263+	H.263, H.263+
H.264	H.264	H.264
G.711 A-law and mu-law	G.711 A-law and mu-law	G.711 A-law and mu-law
G.723.1	G.723.1	G.723.1
G.728	G.728	G.728
G.729, G.729a, G.729b, and G.729ab	G.729, G.729a, G.729b, and G.729ab	G.729, G.729a, G.729b, and G.729ab
G.722	G.722	G.722
G.722.1		
		iLBC
		iSAC
		AAC-LD

Table 12-3 Capabilities Supported in Unified CM Release 9.x (continued)

H.323	SCCP	SIP
H.224 far-end camera control (supported by Unified CM but not by all endpoints); No protocol interworking	H.224 far-end camera control (supported by Unified CM but not by all endpoints); No protocol interworking	H.224 far-end camera control (supported by Unified CM but not by all endpoints); No protocol interworking
Out-of-band DTMF (H.245 alphanumeric) RFC2833 AVT Tones (only for H.323 intercluster trunk to SIP calls)	Out-of-band DTMF RFC2833 AVT Tones	RFC2833 AVT Tones Unsolicited SIP Notify KPML

Note Cisco recommends, whenever possible, registering new and existing H.323 devices to the Cisco TelePresence Video Communication Server (VCS) as a gatekeeper and using H.323-SIP interworking to connect to Unified CM, peering VCS and Unified CM through a SIP trunk.

Endpoints

IP phones, TelePresence personal units, and rich media software clients are the most common video endpoints within the Cisco Unified Communications System. To add video to the IP phones for users, enterprises can use a camera for the Cisco Unified IP Phone 9971, use endpoints such as the Cisco TelePresence System EX60 or EX90 personal TelePresence units, or connect the IP phone to a PC running a software client that supports the Cisco Audio Session Tunnel (CAST). Enterprises can also deploy soft clients such as Cisco Jabber. Third-party endpoints that support protocols such as H.323 and SIP can also be deployed with Cisco Unified CM.

Whenever possible, SIP should be used instead of H.323 as a call control protocol for the endpoints. However, the decision to use H.323 is primarily determined by the use of H.239 for data sharing (for example, sharing PC screens during video calls) or if H.235 is used to pass secure tokens between endpoints for a video call with secure media between the endpoints. The type of user features, such as presence, also govern the type of protocol used for the endpoints, and the protocol choice is primarily dependent on the support for the features needed on the endpoints.

SIP video devices supported by Cisco Unified CM include the Cisco 9900 Series IP Phones, Cisco E20 Video Phone, Cisco Cius, TelePresence personal units (for example, Cisco EX60 and EX90), third-party SIP devices (advanced), or the generic desktop and room system video device. Video can be enabled on the Cisco 9900 Series IP Phones by means of the video capabilities configuration for the devices. Configuration for the Cisco E20 Video Phone is on the phone itself. Cius supports video for calls with its front facing camera. The third-party SIP device (advanced) phone type or the generic video devices are additional options for endpoints from Polycom, Lifesize, Sony, and other manufacturers. While the configuration on Unified CM for these endpoints has not changed from earlier versions, the operation of Unified CM has been optimized to support the Cisco E20 Video Phone, Cisco Cius, Tandberg endpoints, and third-party endpoints more efficiently. Features such as the ability to process Early Offer from the endpoints and process them across SIP trunks without the use of MTP resources provide call signaling optimization and reduce the time to establish media for the call. Unified CM can also now support HD calls from these endpoints so that additional signaling (such as RTCP or parameters passed between the endpoints) is processed and sent across to achieve an optimal video call between the two devices. While Cius needs additional consideration due to its flexible use as a video phone when docked and as a Wi-Fi tablet when undocked, a Wi-Fi network is recommended for audio, and appropriate bandwidth should be used for calls. For additional information on wireless deployments, see the chapter on Mobile Unified Communications, page 25-1.

For further information on the capabilities of the various IP phones and Cisco software clients, see the chapter on Unified Communications Endpoints, page 18-1. For additional information on soft clients such as Cisco Unified Personal Communicator and the Client Services Framework, see the chapter on Cisco Collaboration Clients and Applications, page 24-1.

Selecting the appropriate IP phones and endpoints for users to make video calls depends on the features desired, visual experience required, and capabilities needed for video calls. The available options provide flexibility for designing and deploying clients for different types of users.

Regions

When configuring a region, you set two fields in Unified CM Administration: the Audio Codec and the Video Bandwidth. The audio setting specifies a codec type, while the video setting specifies the amount of bandwidth per call. However, even though the notation is different, the Audio Codec and Video Bandwidth fields actually perform similar functions. The Audio Codec field defines the maximum bit-rate allowed for audio-only calls as well as for the audio channel in video calls. For instance, if you set the Audio Codec for a region to G.711, Unified CM allocates 64 kbps as the maximum bandwidth allowed for the audio channel for that region. In this case, Unified CM will permit calls using either G.711, G.722, G.728, iLBC, or G.729. However, if you set the Audio Codec to G.729, Unified CM allocates only 8 kbps as the maximum amount of bandwidth allowed for the audio channel, and it will permit calls using only G.729 because iLBC, G.728, G.711, and G.722 all take more than 8 kbps.

Note If both endpoints support G.711 and G.722, then G.722 will be negotiated because it is a wideband codec.

The Video Bandwidth field defines the maximum bit-rate allowed for the video channel of the call. However, for historical continuity with the practices used in traditional videoconferencing products, the value used in this field also includes the bandwidth of the audio channel. For instance, if you want to allow calls at 384 kbps using G.711 audio, you would set the Video Bandwidth field to 384 kbps and not 320 kbps.

Note The Audio Codec setting also applies to the audio channel of video calls.

In summary, the Audio Codec field defines the maximum bit-rate used for audio-only calls and for the audio channel of video calls, while the Video Bandwidth field defines the maximum bit-rate allowed for video calls and should include the audio portion of the call.

Choosing the correct audio codec bandwidth limit is very important because each device supports only certain audio codecs. If you set the region to G.729, not all videoconferencing devices are able to support this type of codec. For example, calls between a Cisco Unified IP Phone 9971 and a Cisco TelePresence System EX90 endpoint set to use G.729 would fail, or Unified CM would allocate an audio transcoding resource for the call. (For the most recent list of codecs supported by a particular endpoint, refer to the product documentation for that endpoint.)

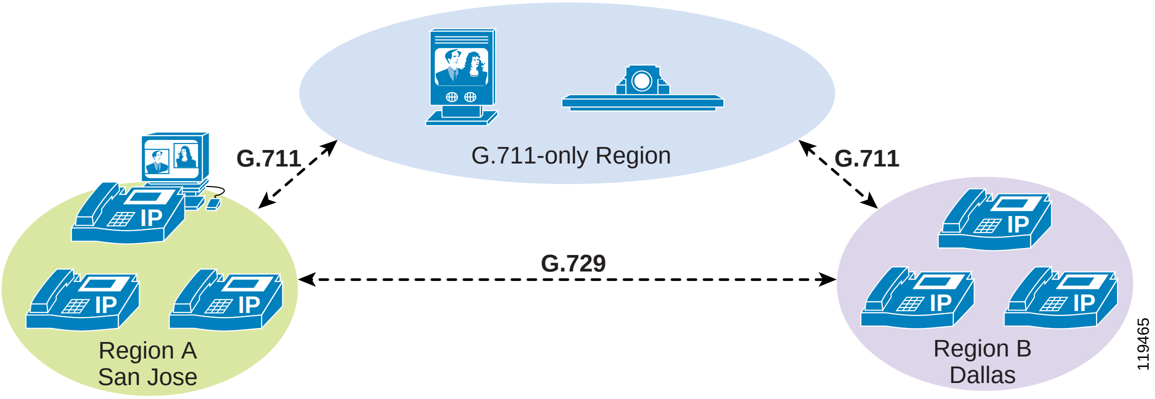
Cisco Unified CM allows transcoding of the audio stream of a video call while still supporting the video stream via a pass-through codec. The pass-through codec is used only for the video stream because the pass-through codec cannot be used for a stream that requires transcoding. The following three conditions must all be true for the pass-through codec to be used:

- The two endpoint devices have a matching CODEC capability.
- **MTP Required** is *not* checked for either endpoint.
- All intermediate resource devices (MTPs and transcoders) support the pass-through codec.

Traditional transcoders do not currently support the pass-through capability, so the call would connect as audio-only and would be transcoded between G.729 and G.711. To avoid this situation without using Cisco IOS Enhanced Transcoders, you would have to set the region to use G.711 instead. However, a region set for G.711 would also use G.711 for audio calls between two IP phones, which you might not want due to the increased consumption of bandwidth over the WAN.

If you want to use G.729 for audio-only calls to conserve bandwidth and to use G.711 for video calls, then you should configure one region to use G.711 for video endpoints that do not support G.729 and a separate region (or regions) to use G.729 for IP phones. (See Figure 12-2.) This method increases the number of regions needed but provides the desired codec and bandwidth allocations.

Figure 12-2 Using G.711 for Video Calls and G.729 for Audio-Only Calls



Note It is possible to configure a pair of regions to prohibit video. If two video-capable devices in that region pair try to call each other, they will connect as audio-only unless Retry Video Call as Audio is not checked, in which case AAR rerouting logic will take over.

Table 12-4 lists some example configurations and their outcomes.

Table 12-4 Scenarios for Various Region Settings

Region Setting	Setting of Retry Video as Audio	Result
Region allows video	Enabled	Video calls allowed
Region allows video	Disabled	Video calls allowed

Table 12-4 Scenarios for Various Region Settings (continued)

Region Setting	Setting of Retry Video as Audio	Result
Region does not allow video	Enabled	Video calls will proceed as audio
Region does not allow video	Disabled	If AAR is not configured, video calls fail (with busy tone and "Bandwidth Unavailable" message displayed)

The Video Call Bandwidth field accepts values in the range of 1 to 32,256 kbps. However, to allow for compatibility with H.323 and H.320 videoconferencing devices, Cisco recommends that you always enter values for this field in increments of either 56 or 64 kbps. Therefore, valid values for this field include 112 kbps, 128 kbps, 224 kbps, 256 kbps, 336 kbps, 384 kbps, and so forth.

When the call speed requested by the endpoint exceeds the bandwidth value configured for the region, Unified CM automatically negotiates the call down to match the value allowed in the region setting. For instance, assume that an H.323 endpoint calls another H.323 endpoint at 768 kbps, but the region is set to allow a maximum of 384 kbps. The incoming H.225 setup request from the calling party would indicate that the call speed is 768 kbps, but Unified CM would change that value to 384 kbps in the outgoing H.225 setup message to the called party. Thus, the called endpoint would think that it was a 384-kbps call to begin with, and the call would be negotiated at that rate. The calling endpoint would show the requested bandwidth as 768 kbps, but the negotiated bandwidth would be 384 kbps.

However, if you set the Video Bandwidth to "None" in the region, Unified CM will either terminate the call (and send an H.225 Release Complete message back to the calling party) or will allow the call to pass as an audio-only call instead, depending on whether or not the called device has the Retry Video Call as Audio option enabled. (See Retry Video Call as Audio, page 12-10.)

As the video resolution for the calls increases, so does the need for bandwidth. For video bandwidth in the region settings, the suggested values are 384 kbps for calls where CIF video resolution is desired, 768 kbps where VGA resolution is desired, and 1.5 Mbps for 720p resolution video calls. While most video endpoints have variable bit-rate encoders, video phones such as the Cisco Unified IP Phone 9900 Series have a constant bit-rate encoder for video. The constant bit-rate encoder provides better motion video and error resiliency.

Some endpoints might support a limited number of resolutions for calls. To restrict the resolution for video calls from such devices to VGA, for example, the region configuration for video call bandwidth can be configured to 768 kbps and the devices can be associated to this region. In this case, calls to endpoints with a higher resolution or conferences to MCUs would negotiate VGA resolution for video.

Some Unified Communications endpoints with wireless capabilities (for example, Cisco Unified IP Phones 9900 Series and Cisco Cius) allow configuration of different bandwidth settings for the wireless media. For further details and design considerations, see Wireless LAN Infrastructure, page 3-54.

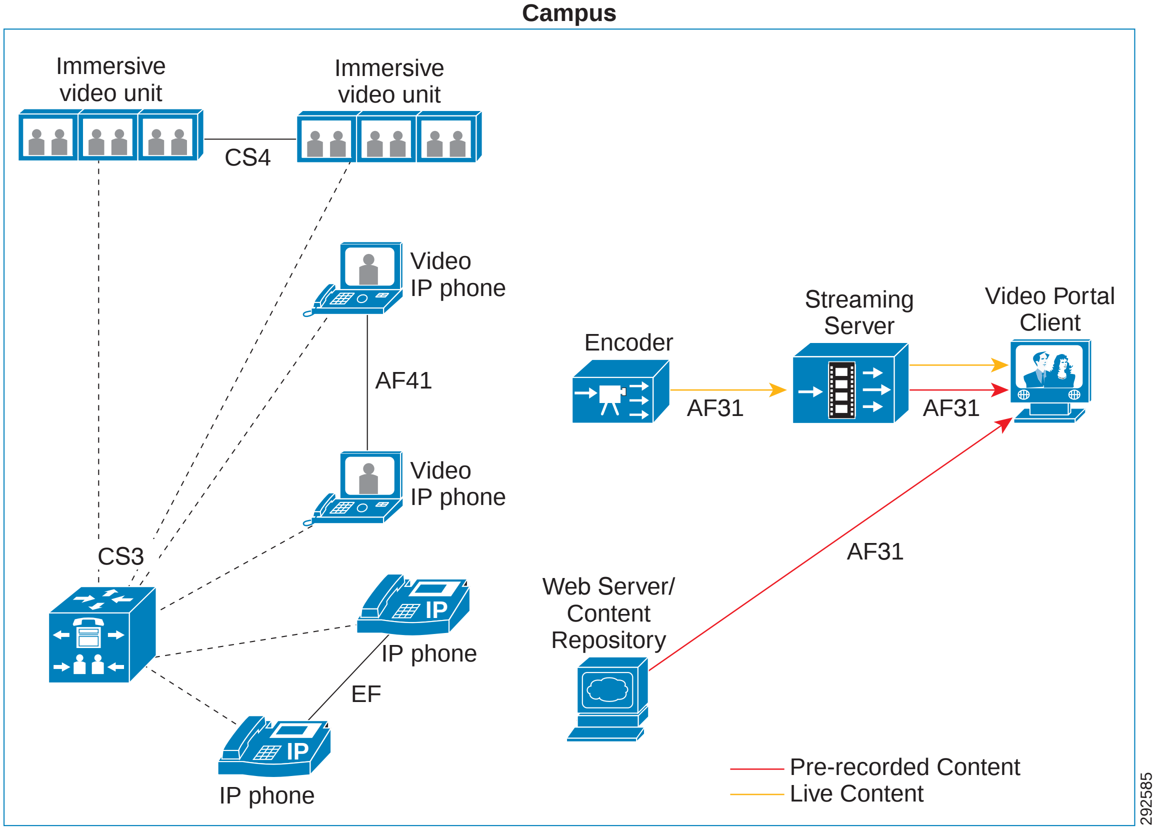
Call Admission Control

Cisco Unified CM can provide call admission control for video calls. For further information, see the chapter on Call Admission Control, page 11-1.

Quality of Service

Cisco recommends using different DSCP markings for different video applications. Unified CM 9.*x* provides support for different DSCP marking for immersive video traffic and videoconferencing (IP video telephony) traffic. By default, Unified CM 9.*x* has preconfigured the recommended DSCP values for TelePresence (immersive video) calls at CS4 and video (IP video telephony) calls at AF41. Figure 12-3 depicts the different video applications in a converged environment using the recommended DSCP values.

Figure 12-3 Recommended QoS Traffic Markings in a Converged Network



Calculating Overhead for QoS

Unlike voice, real-time IP video traffic in general is a somewhat bursty, variable bit rate stream. Therefore video, unlike voice, does not have clear formulas for calculating network overhead because video packet sizes and rates vary proportionally to the degree of motion within the video image itself. From a network administrator's point of view, bandwidth is always provisioned at Layer 2, but the variability in the packet sizes and the variety of Layer 2 media that the packets may traverse from end-to-end make it difficult to calculate the real bandwidth that should be provisioned at Layer 2. However, the conservative rule that has been thoroughly tested and widely used is to over-provision video bandwidth by 20%. This accommodates the 10% burst and the Layer 2 to Layer 4 network overhead.

For more details about Quality of Service, see the QoS information in the chapter on Network Infrastructure, page 3-1.

Retry Video Call as Audio

This check-box is available on SCCP endpoint types that support video and H.323 and SIP devices (clients, gateways and all types of H.323 trunks). When this option is activated (checked), if there is not enough bandwidth to reach the device (for example, if the Unified CM regions or locations do not allow video for that call), then Unified CM will retry the call as an audio-only call. When this option is deactivated (unchecked), Unified CM will not retry the call as audio-only but instead will either fail the call or reroute the call by whatever automated alternate routing (AAR) path is configured. By default, this retry option is enabled (checked).

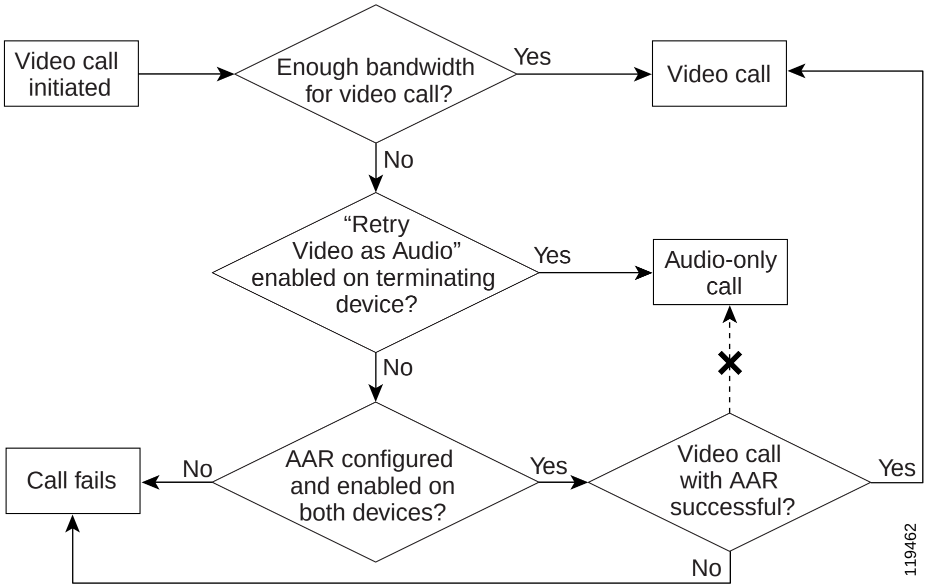
This feature applies to the following scenarios only:

- The region is configured not to allow video.
- The location is configured not to allow video, or the requested video speed exceeds the available video bandwidth for that location when locations are not using an RSVP Policy.
- For calls between Unified CM clusters, the requested video speed exceeds the gatekeeper's zone bandwidth limits.

The Retry Video Call as Audio option takes effect only on the terminating (called) device, thus allowing the flexibility for the calling device to have different options (retry or AAR) for different destinations.

If the video call fails due to bandwidth limitations but automated alternate routing (AAR) is enabled, Unified CM will attempt to reroute the failed call as a video call to the AAR destination. If AAR is not enabled, the failed call will result in a busy tone and an error message being sent to the caller. (See Figure 12-4.)

Figure 12-4 Possible Scenarios for a Video Call


Video call initiated
Enough bandwidth for video call?
Yes
Video call
No
"Retry Video as Audio" enabled on terminating device?
Yes
Audio-only call
No
Call fails
No
AAR configured and enabled on both devices?
Yes
Video call with AAR successful?
Yes
No
119462

See the chapter on Call Admission Control, page 11-1, for further details on the use of AAR.

Dial Plan

From the IP video telephony perspective, it is important to consider the dial plan implications that certain user experiences carry. For instance, Cisco Unified CM 9.0 has added support for URI dialing; however, for URI dialing and DN users to coexist across pre-9.0 Unified CM clusters, the trunk connecting the clusters must be configured to provide only the DN information by setting the **calling and connected party info format** to **deliver DN only in connected party**. Similarly, while a Cisco Video Communication Server (VCS) integrated with Unified CM 9.0 or later releases can take advantage of a URI dial plan scheme without requiring the use of transforms, a VCS integrated with multiple Unified CM clusters of version 9.0 and earlier releases would still require as many transforms because URIs in the VCS would need to be reached by the pre-9.0 Unified CMs. Therefore, Cisco recommends using a numeric dial plan in the VCS for situations where the VCS is integrated with a pre-9.0 Unified CM.

For additional information about URI dialing, see the chapter on Dial Plan, page 9-1.

Trunks

Cisco Unified CM supports various types of trunks. However, the SIP trunk features available in the current release of Unified CM make SIP the preferred choice for new and existing trunk connections. Cisco recommends, whenever possible, registering new and existing H.323 devices to Cisco VCS as a gatekeeper and using H.323-SIP interworking to Unified CM, peering the VCS and Unified CM through a SIP trunk. It is important to consider the bandwidth implications when a VCS is used for interworking because the media of the call being interworked will traverse the VCS (media flow-through).

For situations when use of a VCS is not possible, H.323 trunks can be used to interwork with H.323 gatekeepers that route calls to video endpoints and gateways. H.323 trunks also provide a pass-through functionality for a number of video features used by the H.323 video endpoints, such as H.239 and H.235. The RASAggregator trunk enables Unified CM to provide advanced features such as call restrictions and bandwidth enforcement for calls for the endpoints registered to a Cisco IOS gatekeeper.

SIP trunks can provide interconnection to SIP networks. These trunks support video and SRTP across the trunks. Unified CM can provide a much tighter integration for video communication servers such as the Cisco TelePresence Video Communication Server (VCS). This capability expands the support for high-definition calls so that advanced signaling needed by Cisco VCS, Cisco Video devices, and third-party video endpoints can also work through Unified CM. In addition, Cisco Unified CM SIP trunks support Early Offer for video calls without the need of a media termination point (MTP). This can be important for deployments where calls go through multiple call control servers or where call cut-through time for establishing a bidirectional media path is important. For additional information, see the chapter on Cisco Unified CM Trunks, page 14-1.

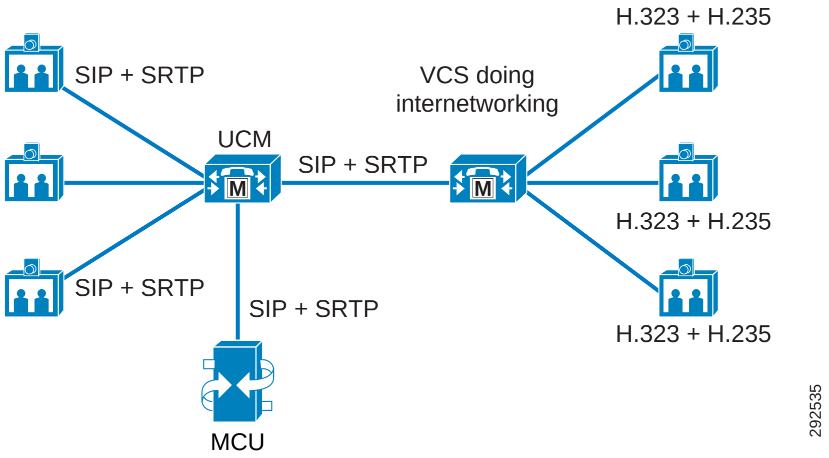
Some deployments may use DNS SRV in their networks. For Cisco TelePresence VCS deployments that use DNS SRV, Unified CM SIP trunks can also use DNS SRV. In such deployments, you need to consider the DNS server scalability and redundancy and also note that the load balancing and redundancy ability is dependent on the DNS server servicing the requests. Thus, the Unified CM trunk load balancing and redundancy will be in addition to the DNS server load balancing and redundancy.

Security

Unified CM 9.*x* supports H.235 pass-through as a security mechanism when interacting with H.323 video devices and has added support for Secure Real-Time Transport Protocol (SRTP) encryption of the video and audio media streams of video calls of Cisco SIP video endpoints. However, interworking of

H.235 to SRTP is not currently supported in Unified CM. Whenever H.235 and SRTP are needed in a video deployment, Cisco recommends registering the H.323 endpoints to a Cisco VCS as a gatekeeper and using SIP-H.323 interworking, while providing SRTP for the SIP video endpoints in the Unified CM side and a secure SIP trunk to the VCS. If the H.323 video endpoints are configured to use H.235 with the VCS, the call can be encrypted end-to-end. Figure 12-5 depicts Unified CM 9.*x* and a VCS working together to provide security end-to-end in a mixed SIP-H.323 network.

Figure 12-5 Unified CM and VCS Providing End-to-End Security in a Mixed SIP-H.323 Network



For further details about security in Unified CM, see the chapter on Unified Communications Security, page 4-1.

Multipoint Conferencing

Whenever three or more parties want to engage in the same video call together, a Multipoint Control Unit (MCU) is required. Cisco Unified CM supports the Cisco MCUs in SCCP, H.323, and SIP modes. Each protocol offers different features, and the protocol and MCU integration should be done based on the conference service types to be deployed. There are three video conference service types:

- Ad-Hoc Video Conferencing, page 12-13
- Meet-Me Video Conferencing, page 12-13
- Scheduled Video Conferencing, page 12-13

Regardless of signaling protocol, the MCU provides the same basic function of receiving the audio and video streams from each participant and sending those streams back out to all other participants in some sort of combined view. There are two types of views in a multipoint video conference:

- Voice Activation, page 12-14
- Continuous Presence, page 12-14

Ad-Hoc Video Conferencing

An ad-hoc video conference refers to an impromptu conference. This conference can be created by a user invoking the Confr function of the IP phone. Cisco Unified CM 9.*x* supports SCCP and SIP MCU integrations for this kind of video conference. The MCU needs to be defined as a media resource in Unified CM for it to be available during the bridge selection process. Cisco Unified CM 9.*x* supports SIP-based TelePresence MCUs through conference bridges, thus providing an additional method of MCU integration for ad-hoc conferences that can support higher resolution.

Only the following events invoke ad-hoc MCU resources:

- The user of an SCCP or SIP endpoint (such as an IP phone or a third-party SCCP video endpoint) presses the Conf, Join, or cBarge softkey to invoke an ad-hoc conference.
- The user of an SCCP or SIP endpoint (such as an IP phone or a third-party SCCP video endpoint) presses the MeetMe softkey to invoke a reservationless meet-me conference.
- The user of Cisco Cius or Cisco Video Software Client (i.e. Cisco Jabber) in softphone mode uses the Join or Conference feature to join multiple calls into a conference.

Participants in either of these types of conferences can include any type of endpoint (that is, video and non-video devices using any signaling protocol that Unified CM supports via any supported gateway type); however, only SCCP endpoints, SIP-based Cisco Unified IP Phones, or Cisco Video Software Clients can invoke the ad-hoc MCU resources. In other words, an H.323 video endpoint cannot invoke an ad-hoc MCU resource, but an SCCP video endpoint can invoke the resource and then join an H.323 video participant to the call. For example, the user at the SCCP endpoint could press the Conf softkey, dial the directory number of an H.323 client, and then press the Conf softkey again to complete the transaction. The H.323 client will be joined as a participant on the SCCP MCU conference.

Note Earlier versions of this document described alternative configurations for H.323 devices that do not support supplementary services (such as placing a call on hold). For details, refer to the section on SCCP MCU Resources in the *Cisco Unified Communications SRND Based on Cisco Unified Communications Manager 7.x*, available at http://www.cisco.com/en/US/docs/voice_ip_comm/cucm/srnd/7x/uc7_0.html.

Meet-Me Video Conferencing

For meet-me video conferencing, the conference initiator creates the conference prior to it by invoking the MeetMe function of the IP phone. The conference initiator then distributes the MeetMe number to the attendees so they can dial in. Unified CM 9.*x* supports SCCP and SIP MCU integrations for meet-me video conferencing. The MCU needs to be defined as a media resource in Unified CM for it to be available during the bridge selection process.

Scheduled Video Conferencing

A scheduled video conference is started by its initiator dialing into an IVR to create the conference or through a middle-ware time management system to schedule it. Unified CM relies on the services available in the MCU and/or middle-ware in this scenario for the creation and the logic control of the conference. Unified CM 9.*x* supports H.323 and SIP MCUs for scheduled video conferences. Cisco strongly advises using H.323 MCUs registered to a Cisco TelePresence System Video Conference Server (VCS as a gatekeeper, and configuring H.323-SIP interworking from the VCS to Unified CM to provide support to this kind of conference.

Voice Activation

Voice-activated (switched) conferences take in the audio and video streams of all the participants, decide which participant is the dominant speaker, and send only the dominant speaker's video stream back out to all other participants. The participants then see a full-screen image of the dominant speaker (and the current speaker sees the previous dominant speaker). The audio streams from all participants are mixed together, so everyone hears everyone else, but only the dominant speaker's video is displayed.

You can use any of the following methods to select the dominant speaker:

- Voice activation mode

  Using this mode, the MCU automatically selects the dominant speaker by determining which conference participant is speaking the loudest and the longest. To determine loudness, the MCU calculates the strength of the voice signal for each participant. As conditions change throughout the conversation, the MCU automatically selects a new dominant speaker and switches the video to display that participant. A hold timer prevents the video from switching too hastily. To become the dominant speaker, a participant has to speak for a specified number of seconds and be more dominant than all other participants.

- Manual selection of the dominant speaker through the MCU’s web-based conference control user interface

  The conference controller (or chairperson) can log onto the MCU's web page, highlight a participant, and select that person as the dominant speaker. This action disables voice activity detection, and the dominant speaker remains constant until the chairperson either selects a new dominant speaker or re-enables voice activation mode.

- Configuring the MCU to cycle through the participant list automatically, one participant at a time

  With this method, the MCU stays on each participant for a configured period of time and then switches to the next participant in the list. The conference controller (or chairperson) can turn this feature on and off (re-enable voice activation mode) via the web interface.

Continuous Presence

Continuous-presence conferences display some or all of the participants together in a composite view. The view can display the participants in a variety of different layouts. Each layout offers the ability to make one of the squares voice-activated, which is useful if there are more participants in the conference than there are squares to display them all in the composite view. For instance, if you are using a four-way view but there are five participants in the call, only four of them will be displayed at any given time. You can make one of the squares in this case voice-activated so that participants 4 and 5 will switch in and out of that square, depending on who is the dominant speaker. The participants displayed in the other three squares would be fixed, and all of the squares can be manipulated via the conference control web-based user interface.

Note Cisco strongly advises against the use of Asynchronous Continuous Presence.

Note For H.323 and SIP clients with built-in MCUs, Unified CM does not allow an H.323 client to generate a second call, thereby negating the functionality of the built-in MCU.

Secure Conferencing

Unified CM 9.*x* supports secure conferencing with SIP MCU integration types. With secure conferencing, Unified CM uses HTTPS to communicate to the MCU for conference scheduling, and it uses TLS and SRTP for call signaling and media payload encryption, respectively. However, the conference is secure only if all the participants' endpoints support video encryption.

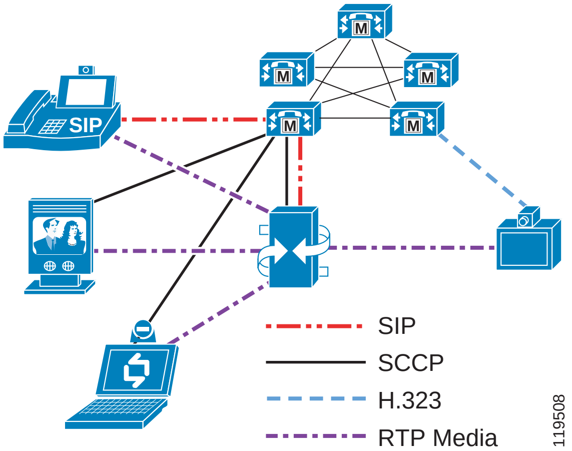
For more information about secure conferencing, see the chapter on Unified Communications Security, page 4-1.

MCU Resources for Ad-Hoc Conferences

For reservationless conferences via the MeetMe softkey, the signaling protocol used by the other endpoints does not have to support being placed on hold and transferred. For these types of conferences, each endpoint dials the MeetMe dial-in number arranged by the endpoint that initiated the conference.

Figure 12-6 illustrates how H.323 endpoints and Cisco IP Phones can participate in the same ad-hoc conference. In this example, the conference was initiated by an SCCP endpoint using the Conf softkey to invite the three members.

Figure 12-6 Ad-Hoc Conference Between SCCP, SIP, and H.323 Endpoints



Ad-hoc conferences support voice-activated mode as well as continuous presence, depending on the conferencing bridge used.

Media Resource Groups and Lists

When a user of an SCCP or SIP phone activates the Conf, Join, or MeetMe softkey, Unified CM uses the Media Resource Manager to select conference bridges. Conference bridges or MCUs resources are configured in the media resource groups (MRGs). The media resource group lists (MRGLs) specify a prioritized list of MRGs and can be associated with the endpoints. The Media Resource Manager uses MRGLs of the endpoints for selecting the conference bridge. How you group the resources is completely at your discretion, but it is typically done either by geographical placement (so that all endpoints at a given site use the conference bridges closest to them) or by endpoint type (so that video-capable endpoints use a video-capable MCU while audio-only endpoints use a different conference bridge resource).

Cisco Unified CM has the Intelligent Bridge Selection feature, which provides a method for selecting conference resources based on the capabilities of the endpoints in the conference. If there are two or more video endpoints when the conference is invoked and a videoconferencing resource is available, Intelligent Bridge Selection chooses that resource for the conference. On the other hand, if no videoconferencing resource is available or if there are no video-capable endpoints in the conference, Intelligent Bridge Selection chooses an available audio resource for the conference. Intelligent Bridge Selection provides an added functionality to select secure conference bridges for secure conferences. However, secure conference bridge selection is dependent on device capabilities. Unified CM may decide to allocate secure conference bridges in lieu of video or audio conference bridges. Flexibility to change the behavior of the Intelligent Bridge Selection functionality is provided through service parameter configurations in Unified CM.

Intelligent Bridge Selection has the following advantages over other methods of conference bridge selection:

- Conference bridge selection by conference type – either secure, video, or audio conferences
- Simplified media resource configuration
- Optimized use of MCU video ports that potentially would have been used for audio-only conferences with other methods of bridge selection

All the conference bridge resources and MCUs can be in one MRGL, and Intelligent Bridge Selection will then select the conference bridge based on the need to do just an audio conference or a video conference.

Unified CM also supports an alternate way of selecting conference bridges, which can be specified by service parameter configurations. In this mode, Unified CM applies the following criteria to select the conference bridge resource to use, in the order listed here:

1. The priority order in which the media resource groups (MRGs) are listed in the media resource group list (MRGL)
2. Within the selected MRG, the resource that has been used the least

If the MCU is placed at the top of the MRGL for the phone, the MCU will always be chosen even for audio-only conferences that do not involve any video-capable participants. In this scenario, the MCU resources might be wasted on audio-only conferences and not be available to satisfy the request for a video conference when it occurs.

Note Meet-me conferences do not use the Intelligent Bridge Selection feature.

Intelligent Bridge Selection

Cisco Unified CM includes the Intelligent Bridge Selection feature, which provides a method for selecting conference resources based on the capabilities of the endpoints in the conference. If there are two or more video endpoints when the conference is invoked and a videoconferencing resource is available, Intelligent Bridge Selection chooses that resource for the conference. On the other hand, if no videoconferencing resource is available or if there are no video-capable endpoints in the conference, Intelligent Bridge Selection chooses an available audio resource for the conference.

Intelligent Bridge Selection provides an added functionality to select secure conference bridges for secure conferences. However, secure conference bridge selection is dependent on device capabilities. Unified CM may decide to allocate secure conference bridges in lieu of video or audio conference bridges. Flexibility to change the behavior of the Intelligent Bridge Selection functionality is provided through service parameter configurations.

Cisco Business Edition

When using video with Cisco Business Edition 3000, keep in mind that Business Edition 3000 does not support video conference bridge registrations; therefore, no multipoint calls are supported in Business Edition 3000.

Cisco Business Edition 6000 does support video conference bridges and provides the ability for multipoint calls to the endpoints it services.

H.323 and SIP MCU Resources

When configured in H.323 or SIP mode, the MCU provides the MC function and behaves like an H.323 or SIP peer to Unified CM. Because of SIP features available in the current release of Unified CM, SIP is the preferred choice for integration of MCUs. Cisco recommends registering H.323 MCU resources to a Cisco VCS and using its H.323-SIP interworking capabilities to peer it with Unified CM through a SIP trunk.

H.323 and SIP MCU conferences can be invoked in a number of different ways, but they all fall into two major categories:

- Scheduled
- Reservationless

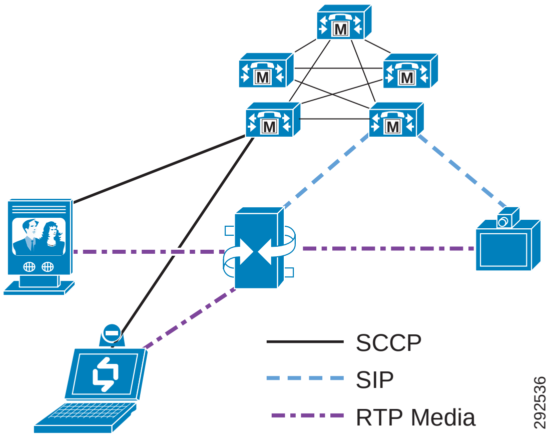
A scheduled conference uses a scheduling application to reserve the MCU resources in advance of the call. The scheduling function typically is provided through a web-based user interface such as Cisco Unified MeetingPlace, or Cisco Unified Video Conferencing Manager. The scheduling application usually generates an invitation that provides the user with the date and time of the conference, the number of ports reserved for the conference, and the dial-in information. Alternatively, the scheduling system can be configured to dial out to some or all of the participants at the beginning of the conference.

For a reservationless conference, the MCU has a certain number of resources that are available on demand. To create a conference, a user simply dials into the MCU at any time. If that user is the first participant to dial in, the MCU dynamically creates a new conference using the settings defined in the service template. Subsequent users who dial into the same conference number are joined to that conference.

Any type of endpoint can create and participate in scheduled and reservationless H.323 or SIP conferences. For instance, an SCCP endpoint can dial into the SIP MCU to create a reservationless conference just as well as a SIP endpoint can.

Figure 12-7 illustrates how SIP and SCCP endpoints can participate in the same SIP conference. In this example, the conference was initiated by an SCCP endpoint that dialed into the SIP MCU to create a new reservationless conference, and the other two parties subsequently dialed into that conference.

Figure 12-7 SCCP and SIP Endpoints in a Reservationless Conference



H.323 and SIP conferences support both voice-activated and continuous-presence modes.

Sizing the MCU

There are several factors involved in determining the types and number of conferences that an MCU can support. These sizing factors are different for different models of MCUs. MCUs can also make available more ports when using standard definition (SD) mode as compared to the high definition (HD) mode.

Calculating the size of MCUs depends on the following factors:

- The type of resolution for the video conference
- The total number of ports that the MCU can support
- The number of ports that the MCU can dedicate to each protocol
- Whether cascading conferences are needed between MCUs

For specific information about the number of ports supported, refer to the product documentation for your MCU hardware, available on Cisco.com. Due to the almost infinite number of possible variations, it is very difficult to provide any concrete design guidance in this document. Many customers end up with a mixture of SIP or SCCP ad-hoc conferences, H.323 and SIP reservationless conferences, and H.323 and SIP scheduled conferences. The MCUs must be sized to accommodate all of those types of conferences at the correct speeds and video layouts. Needless to say, this can become quite complex to determine. Please consult with your Cisco sales representative for assistance on sizing the MCUs for your particular environment.

IVR for Dial-In Conference

Dial-in conferences typically use an interactive voice response (IVR) system to prompt users to enter the conference ID and the password (if one is configured) of the conference they want to join. You can use either of the following types of IVRs with the Cisco MCUs:

- The IVR built into the MCU
- Cisco Unified IP IVR

The built-in IVR of the MCU has the following characteristics:

- Can prompt to create a conference or join by conference ID
- Can prompt for the password of the conference
- Supports both in-band and out-of-band (H.245 alpha-numeric) DTMF
- Cannot be customized to provide more flexible menus or functionality

  The only thing that can be customized is the recorded audio file that is played to the user.

If you want to have a single dial-in number and then prompt the user for the conference ID, you can use Cisco Unified IP IVR in conjunction with the MCU.

Cisco Unified IP IVR has the following characteristics:

- Can prompt for the conference ID and the password (among other things)
- Supports only out-of-band DTMF

  That is, the calling device must support an out-of-band DTMF method, such as H.245 alpha-numeric on H.323 devices. These out-of-band DTMF messages are then relayed by Unified CM to the Cisco IP IVR server. If the calling device supports only in-band DTMF tones, the Cisco IP IVR server will not recognize them and the calling device will be unable to enter the conference.

- Can be highly customized to provide more flexible menus and other advanced functionality

  Customizations can include such things as verifying the user's account against a back-end database before permitting that user to enter into the conference, or queuing the participants until the chairperson joins.

Note Because Cisco Unified IP IVR supports only out-of-band signaling, it will not work with H.323 endpoints that use in-band DTMF tones.

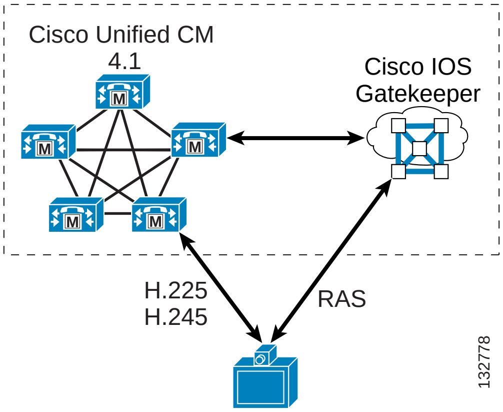
With Cisco Unified IP IVR, users dial a CTI route point that routes the call to the Cisco Unified IP IVR server instead of dialing a route pattern that routes directly to the MCU. After collecting the DTMF digits of the conference ID, the Cisco Unified IP IVR then transfers the call to the route pattern that routes the call to the MCU. This transfer operation requires that the calling device supports having its media channels closed and reopened to a new destination. For example, an H.323 video device that calls the Cisco Unified IP IVR will initially negotiate an audio channel to the Cisco Unified IP IVR server and then, after entering the appropriate DTMF digits, it will be transferred to the MCU, at which point Unified CM will invoke the Empty Capabilities Set (ECS) procedure described earlier in this chapter to close the audio channel between the endpoint and the Cisco Unified IP IVR server and open new logical channels between the endpoint and the MCU. If the H.323 video endpoint does not support receiving an ECS from Unified CM, it will react by disconnecting the call or, worse, crashing and/or rebooting.

Gatekeepers

Because of SIP features available in the current release of Unified CM and the robust H.323 video support in Cisco VCS, H.323 video devices should be registered to Cisco VCS as a gatekeeper whenever possible. The Cisco VCS can then provide local call resolution to the devices and SIP-H.323 interworking with a neighboring Unified CM through a SIP trunk. In case registration is not possible, however, the following sections offer guidance about Unified CM and H.323 gatekeeper integration.

Unified CM and the gatekeeper work as a team to manage H.323 video endpoints. The gatekeeper handles all Registration, Admission, and Status (RAS) signaling, while Unified CM handles all of the H.225 call signaling and H.245 media negotiations. Therefore, you have to deploy gatekeepers along with the Unified CM servers if RAS signaling procedures are required for the H.323 endpoints in your network, as illustrated in Figure 12-8.

Figure 12-8 Unified CM and Cisco IOS Gatekeeper Provide RAS Signaling for H.323 Endpoints



RAS signaling is required any time either of the following conditions exists:

- The endpoint does not use a fixed IP address.

  If the endpoint uses a static IP address, Unified CM does not require RAS procedures to locate the endpoint. Instead, the endpoint is provisioned in Unified CM Administration with its static IP address, and calls to that H.323 client's directory number are routed directly to that static IP address. If the endpoint does not use a static IP address, then Unified CM must query the gatekeeper to obtain the endpoint's current IP address each time Unified CM extends a call to the endpoint.

- The endpoint requires RAS procedures to place calls to E.164 addresses.

  Most H.323 videoconferencing endpoints are capable of dialing another endpoint directly only when dialing by IP address (that is, the user enters the IP address of the destination endpoint in dotted-decimal format and then pushes the call button). However, if the user dials an E.164-formatted number (a numeric value not in the dotted-decimal format of an IP address) or an H.323-ID (in the format of *username* or *username@domain*), most endpoints today provide only one way to resolve these types of destinations – by a RAS query to their gatekeeper. A growing number of endpoints, however, can be configured so that, for any call to an E.164 address, they skip any RAS procedures and instead send an H.225 SETUP message directly to a specified IP address. This method of operation is known as peer-to-peer mode. Tandberg H.323 endpoints are one example that use this mode, in which you can either configure a gatekeeper address for them to register with, or configure the IP address of the Unified CM server they should use. In the latter case, the endpoint sends all calls directly to the specified IP address, bypassing the need for RAS procedures with any gatekeeper.

In addition to managing RAS procedures for H.323 video endpoints, gatekeepers also continue to play an important role in managing dial plan resolution and bandwidth restrictions between Unified CM clusters in large multisite distributed call processing environments. A gatekeeper can also integrate with large numbers of H.323 VoIP gateways within the organization, or it can act as a session border controller between an enterprise IP Telephony network and a service provider VoIP transport network.

Therefore, as it pertains to Cisco IP Video Telephony deployments, the Cisco IOS Gatekeeper can perform one or both of the following roles:

- Endpoint gatekeeper

  An endpoint gatekeeper is configured to manage all RAS procedures for calls to, from, and between H.323 clients, MCUs, and H.320 video gateways. The endpoint gatekeeper directs all such calls to the appropriate Unified CM cluster so that Unified CM can perform all of the H.225 call routing and H.245 media negotiations.

- Infrastructure gatekeeper

  An infrastructure gatekeeper is configured to manage all dial plan resolution and bandwidth restrictions (call admission control) between Unified CM clusters, between a Unified CM cluster and a network of H.323 VoIP gateways, or between a Unified CM cluster and a service provider's H.323 VoIP transport network.

In previous Cisco Unified CM releases, the endpoint gatekeeper and the infrastructure gatekeeper had to run on separate routers, and each endpoint gatekeeper could service only a single Unified CM cluster. If multiple Unified CM clusters existed within the enterprise, a separate endpoint gatekeeper had to be deployed for each Unified CM cluster. With the current Cisco Unified CM release, it is possible to combine these roles on a single gatekeeper, using it as an endpoint gatekeeper for one or more Unified CM clusters and as the infrastructure gatekeeper for managing calls between clusters or between a cluster and other H.323 VoIP networks. However, for the following reasons (among others), Cisco recommends that you still separate these roles onto two or more gatekeepers:

- Scalability

  Depending on the Cisco IOS router platform you choose to deploy and your estimated busy hour call volume, you might need several gatekeepers to handle the load.

- Geographical resiliency

  Putting all of your eggs into one basket may not be wise in a large, multi-national VoIP network. Having gatekeepers placed throughout your network (typically by geography) can provide better fault isolation in the event of a gatekeeper failure.

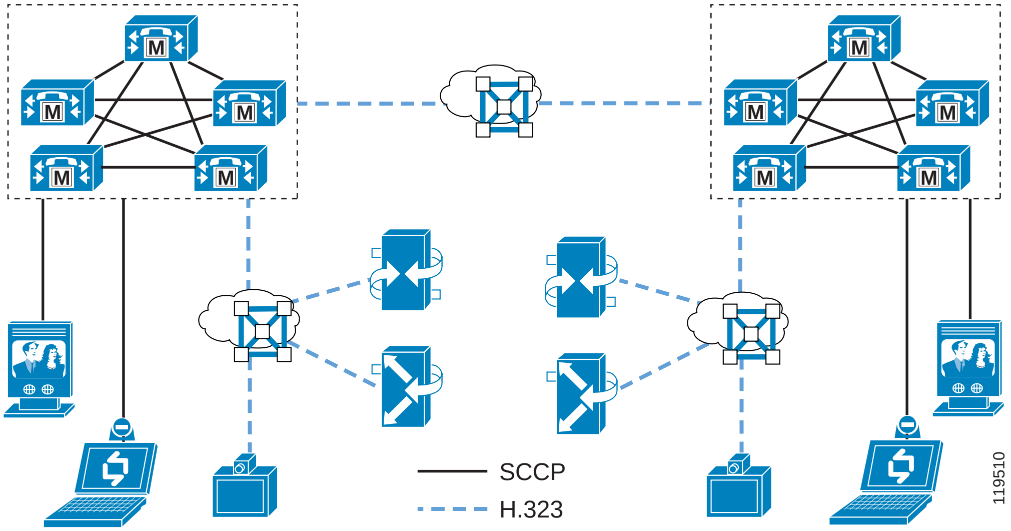
- Incompatibilities

  Some configuration aspects of the gatekeeper are global in nature (they pertain to all endpoints registered with that gatekeeper). For example, the command **arq reject-unknown-prefix**, which may be useful in some H.323 VoIP transport environments, conflicts with the use of the **gw-type-prefix** <*prefix*> **default-technology** command, which is used in endpoint gatekeepers to route calls to Unified CM. While Cisco IOS does not stop you from configuring both commands on the same gatekeeper, the **arq reject-unknown-prefix** command takes precedence and, therefore, calls to unknown numbers will be rejected instead of being routed to Unified CM. In this case, you would have to use one gatekeeper for the H.323 VoIP transport network and another gatekeeper for the Unified CM cluster(s).

  Another example of incompatibility can occur in the way you configure the gatekeeper for redundancy. Most Cisco H.323 voice devices, including Cisco Voice Gateways and Unified CM, support the H.323v3 Alternate Gatekeeper feature, which would allow you to configure the gatekeepers as a gatekeeper cluster using the Gatekeeper Update Protocol (GUP) to keep in sync with each other. However, many H.323 video endpoints do not support Alternate Gatekeeper, so the gatekeepers must be configured to use Hot Standby Routing Protocol (HSRP) for redundancy. You cannot mix and match these two redundancy methods on the same gatekeeper. In this case, you might decide to use a gatekeeper cluster for those endpoints that support Alternate Gatekeeper and an HSRP pair of gatekeepers for those that do not.

Figure 12-9 illustrates a network scenario with two Unified CM clusters. Each cluster consists of SCCP and H.323 clients, H.323 MCUs, and H.320 gateways. To manage the RAS aspects of the H.323 clients, MCUs, and H.320 gateways, an endpoint gatekeeper is deployed with each cluster. A separate infrastructure gatekeeper manages dial plan resolution and bandwidth between the clusters. Gatekeeper redundancy is not shown in the figure, although each of these gatekeepers may actually be multiple gatekeepers configured for either Alternate Gatekeeper or HSRP-based redundancy.

Figure 12-9 Two Unified CM Clusters with Required Gatekeepers



Endpoint Gatekeepers

An endpoint gatekeeper is required any time both of the following conditions are met:

- The cluster contains H.323 clients, H.323 MCUs, or H.320 gateways (collectively referred to as H.323 endpoints). If none of these types of endpoints exists (for example, if all clients are SCCP endpoints and there are no MCUs or H.320 gateways), then an endpoint gatekeeper is not needed.
- And either of the following conditions is true:
  - The H.323 endpoints require RAS procedures to initiate calls to E.164 addresses. As mentioned earlier, a growing number of devices are capable of peer-to-peer call signaling, in which case there is no need for those devices to register with a gatekeeper.
  - The H.323 endpoints do not use static IP addresses.

The role of the endpoint gatekeeper is simply to handle the RAS aspects of communications with the endpoints, providing a place for these H.323 endpoints to register. The endpoint gatekeeper responds to all call requests made to, from, or between these endpoints by directing the call to the appropriate Unified CM server(s) so that Unified CM can perform all of the call routing and bandwidth control functions. To accomplish this call routing and bandwidth control, you configure Unified CM to register H.323 trunk(s) with the gatekeeper and configure the gatekeeper to route calls to those trunks for all calls to, from, or within that zone.

Cisco Unified CM should register to endpoint gatekeepers using a type of H.323 trunk called the RASAggregator trunk. This type of trunk is used for all H.323 client, H.323 MCU, or H.320 gateway zones, while the gatekeeper-controlled intercluster trunk and gatekeeper-controlled H.225 trunk are used to integrate with infrastructure gatekeepers.

Provisioning H.323 Clients

H.323 clients are provisioned much the same way as other phones are, in that you create a new phone (model type = H.323 Client), assign a directory number to it, and assign it a calling search space, device pool, and so forth. You configure the H.323 clients in Unified CM in one of the following ways. The method you use depends on whether or not the client uses a static IP address and whether or not the client requires RAS procedures to dial E.164 addresses.

- Gatekeeper controlled

  This type of configuration is used for clients that do not have a static IP address assigned to them (they use a DHCP-assigned address) and that require RAS procedures to dial E.164 addresses. A RASAggregator trunk is used to communicate to and from these clients. (See Figure 12-10 and Figure 12-11.)

- Non-gatekeeper controlled, asynchronous

  This type of configuration is used for clients that have a static IP address assigned to them but that require RAS procedures to dial E.164 addresses. While Unified CM can signal directly to them without the need of a gatekeeper to resolve their IP addresses, they are not able to signal directly to Unified CM but instead must query the gatekeeper to resolve the E.164 address they are trying to dial (thus, asynchronous communications). To support these types of clients, you must have at least one gatekeeper-controlled client defined in Unified CM for each zone on the gatekeeper, even if all the clients actually use static IP addresses. In this case, the non-gatekeeper controlled client may be a "dummy" client that does not actually exist. Its purpose is merely to create the RASAggregator trunk so that the gatekeeper will be able to route calls from the clients to Unified CM. (See Figure 12-12 and Figure 12-13.)

- Non-gatekeeper controlled, synchronous

  This type of configuration is used for clients that have a static IP address and are also capable of peer-to-peer signaling (that is, they do not require RAS procedures to dial E.164 numbers). Unified CM signals directly to them, and they signal directly to Unified CM (thus, synchronous communications). No gatekeeper or RASAggregator trunk is needed for this type of client. (See Figure 12-14 and Figure 12-15.)

Figure 12-10 through Figure 12-15 illustrate the call signaling flows used in these three scenarios.

Figure 12-10 Call to Gatekeeper-Controlled Client from Unified CM

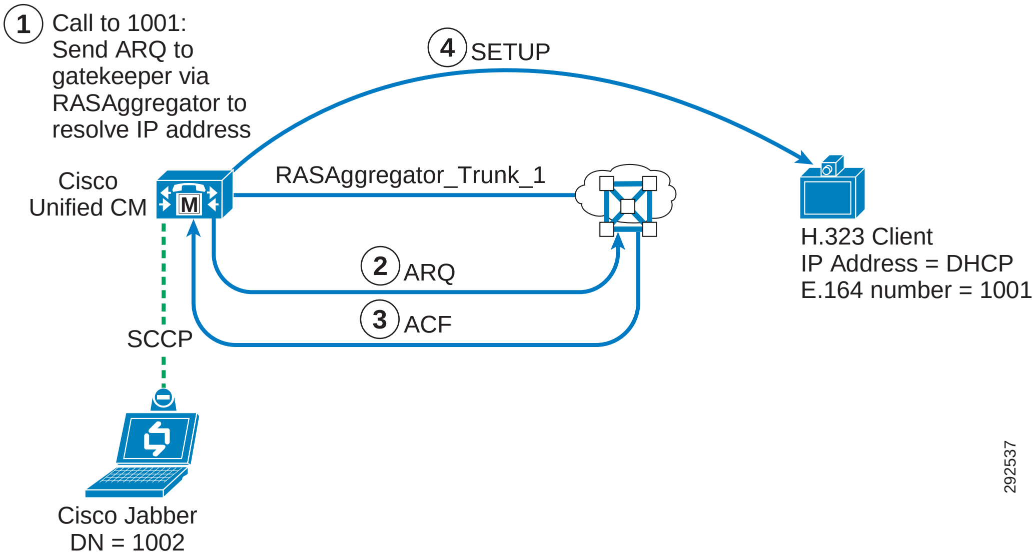


Figure 12-11 Call from Gatekeeper-Controlled Client to Unified CM

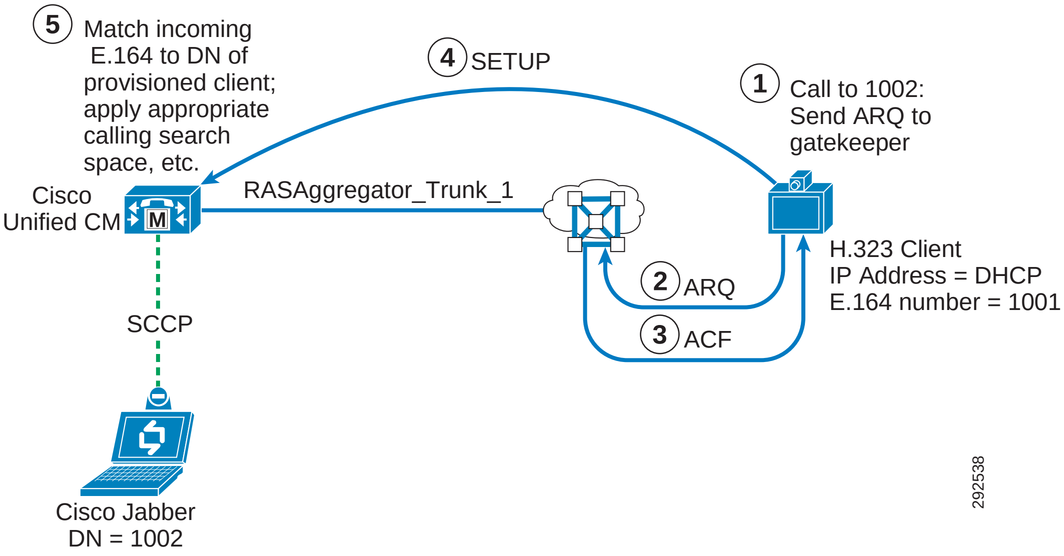


Figure 12-12 Call to Non-Gatekeeper Controlled Client from Unified CM (Asynchronous)

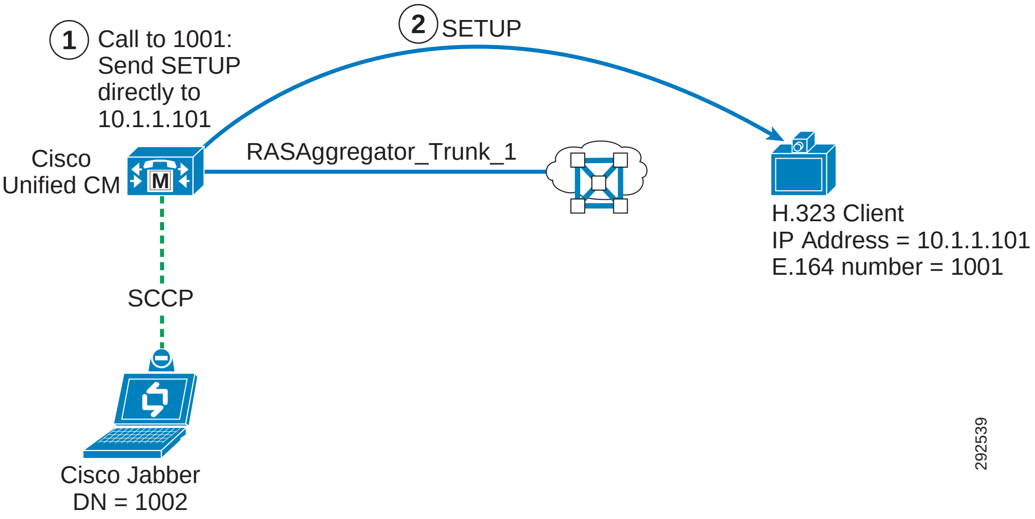


Figure 12-13 Call from Non-Gatekeeper Controlled Client to Unified CM (Asynchronous)

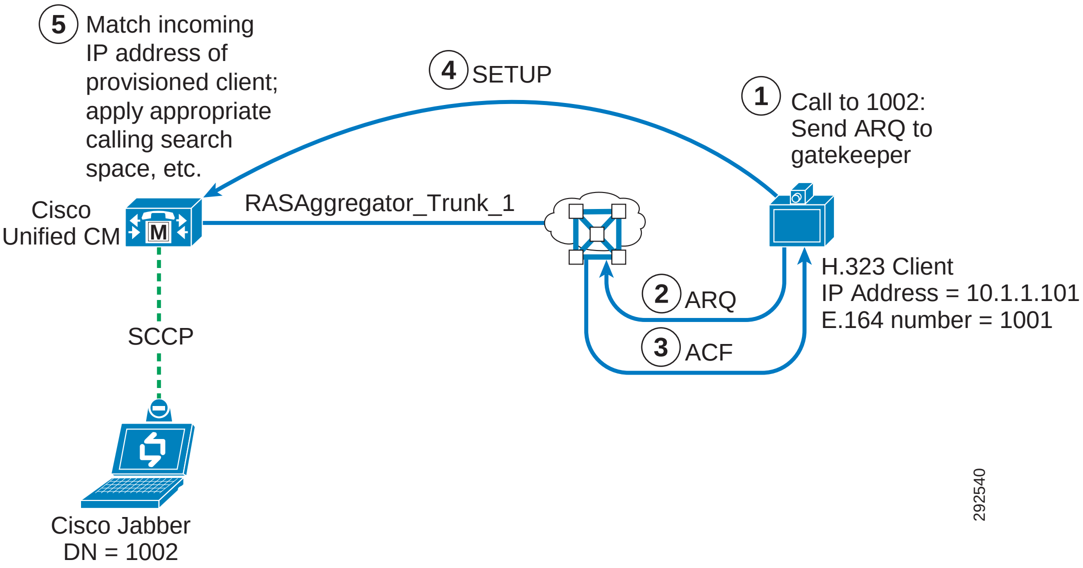


Figure 12-14 Call to Non-Gatekeeper Controlled Client from Unified CM (Synchronous)

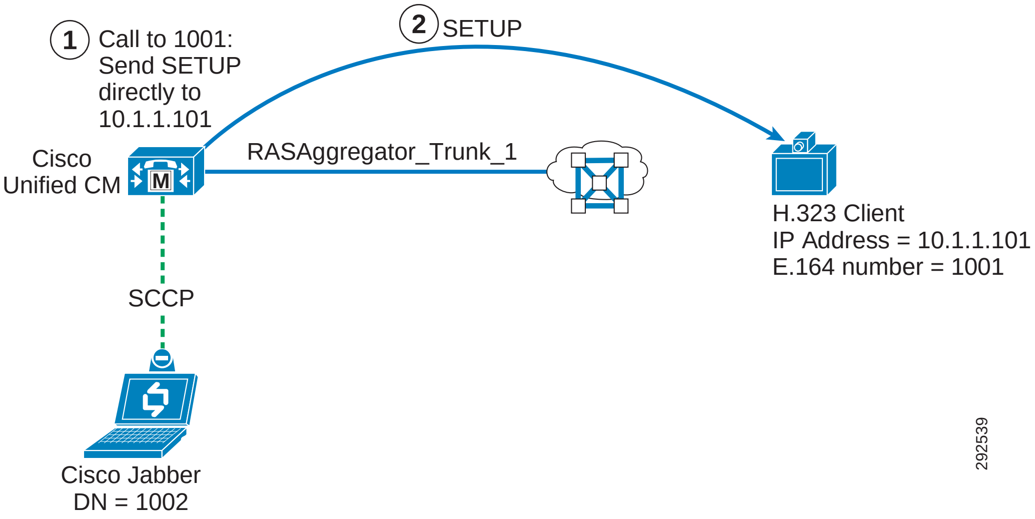
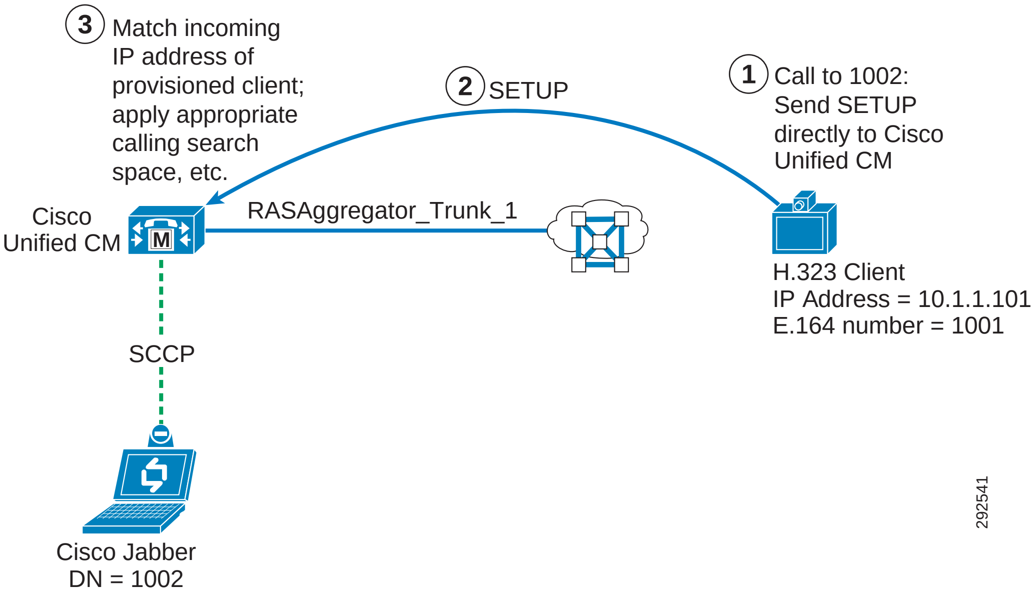


Figure 12-15 Call from Non-Gatekeeper Controlled Client to Unified CM (Synchronous)



Gatekeeper-Controlled Clients

When you configure an H.323 client as gatekeeper-controlled, you may enter any alpha-numeric string (such as a descriptive name) in the Device Name field, check the **Gatekeeper-controlled** box, and fill in the following fields:

- Device Pool

  The device pool you want the client to use. All H.323 clients (whether gatekeeper-controlled or non-gatekeeper controlled) that are registered in the same zone must use the same device pool. If you accidentally assign different device pools across the endpoints, Unified CM will register multiple RASAggregator trunks within the zone, and an inbound call might be rejected by Unified CM if the call is directed to the wrong RASAggregator trunk.

- Gatekeeper

  A drop-down list of gatekeeper IP addresses. You must define the gatekeeper in Unified CM before configuring any gatekeeper-controlled H.323 clients.

- Technology Prefix

  The technology prefix used by the RASAggregator trunk to register in the client zone on the gatekeeper. This technology prefix must match what is configured as the default technology prefix on the gatekeeper. All gatekeeper-controlled H.323 clients that are registered in the same zone must use the same technology prefix. If you accidentally assign different technology prefixes across the endpoints, Unified CM will register multiple RASAggregator trunks within the zone, and an inbound call might be rejected by Unified CM if the call is directed to the wrong RASAggregator trunk. Cisco recommends that you use **1#** for this prefix.

- Zone Name

  The (case-sensitive) name of the client zone as configured in the gatekeeper. All gatekeeper-controlled H.323 clients that are registered in the same zone must use the same zone name. If you accidentally assign different zone names (remember, the field is case sensitive) across the endpoints, Unified CM will attempt to register multiple RASAggregator trunks with the gatekeeper (but the one with the incorrect zone name will fail to register), and an inbound call might be rejected by Unified CM if the call is directed to the wrong RASAggregator trunk.

Also, you must set the Unified CM service parameter **Send Product ID and Version ID** to **True**. This parameter allows the RASAggregator trunk to register with the gatekeeper as an H323-GW, so that the gatekeeper can direct all H.323 calls to, from, or within the client zone to the RASAggregator trunk.

Non-Gatekeeper Controlled Clients

When provisioning an H.323 client as non-gatekeeper controlled, you must enter the static IP address of the client into the Device Name field and leave all of the settings under the Gatekeeper-controlled section blank (unchecked). Unified CM then uses the static IP address to reach the client any time a call is extended to its directory number.

If the client is configured to use peer-to-peer mode, then no further configuration is required. If the client requires RAS procedures to place calls to E.164 addresses, then you must also configure a dummy gatekeeper-controlled H.323 client in order to create the RASAggregator trunk, by filling in the following fields:

- Device Name

  A descriptive name that identifies this client as a dummy client used for the purpose of creating the RASAggregator trunk for the client zone.

- Device Pool

  The device pool you chose when configuring the non-gatekeeper controlled H.323 client(s). If the device pool assigned to the dummy client is different than that assigned to the real clients, inbound calls from the real clients might be rejected by Unified CM.

- Gatekeeper

  A drop-down list of gatekeeper IP addresses. You must define the gatekeeper in Unified CM before configuring the dummy gatekeeper-controlled H.323 client.

- Technology Prefix

  The technology prefix used by the RASAggregator trunk to register in the client zone on the gatekeeper. This technology prefix must match what is configured as the default technology prefix on the gatekeeper. Cisco recommends that you use **1#** for this prefix.

- Zone Name

  The (case-sensitive) name of the client zone as configured in the gatekeeper.

Also, you must set the Unified CM service parameter **Send Product ID and Version ID** to **True**. This parameter allows the RASAggregator trunk to register with the gatekeeper as an H323-GW, so that the gatekeeper can direct all H.323 calls to, from, or within the client zone to the RASAggregator trunk.

Provisioning H.323 MCUs

H.323 MCUs are provisioned in Unified CM as H.323 gateways, and then route patterns are configured to extend calls to these devices. When provisioning an H.323 gateway, you must enter the static IP address and TCP signaling port of the MCU into the Device Name field. Unified CM then uses the static IP address and TCP port to reach the MCU any time a call matches the route pattern(s) associated with it.

Note The Cisco Unified Videoconferencing 3500 and 5000 Series MCUs do not listen on TCP port 1720 by default. (The Cisco Unified Videoconferencing 3500 and 5000 Series MCUs listen on port 2720 by default.) You must verify which TCP port they are listening on, and either change it to 1720 or provision the correct port in Unified CM.

If the MCU is configured to use peer-to-peer mode, then no further configuration is required. (Cisco Unified Videoconferencing MCUs do not currently support peer-to-peer mode, but some third-party MCUs do.) If the MCU requires RAS procedures to place calls to E.164 addresses, then you must also configure a dummy gatekeeper-controlled H.323 client in order to create the RASAggregator trunk, by filling in the fields for Device Name, Device Pool, Gatekeeper, Technology Prefix, and Zone Name as discussed in the section on Non-Gatekeeper Controlled Clients, page 12-27.

MCU Service Prefixes

H.323 MCUs can use either E.164 addresses or technology prefixes (also referred to as service prefixes in the MCU) as the dial-in number(s) to reach reservationless or scheduled H.323 conferences running on them. Cisco recommends that you configure the MCUs to use E.164 addresses by setting the MCU Mode to **MCU** instead of **Gateway** in the MCU administration screens. If the **MCU** setting is not available on the model of MCU you are using, then you must use the following special configuration to properly route calls placed from other H.323 endpoints to the MCU:

If the MCU is configured in **Gateway** mode or is another vendor's MCU that (for whatever reason) requires its conference IDs to register as technology prefixes instead of as E.164 addresses, then the service prefix(s) of the MCU must begin with a **#** character. For example, if the MCUs service prefix is 8005551212, then you must provision the service prefix on the MCU as #8005551212. Thus, when other H.323 endpoints dial 8005551212, the gatekeeper will not find a matching technology prefix registered and will instead route the call to the RASAggregator trunk that is registered with the default technology prefix in the zone of the endpoint that is placing the call. Unified CM must then prepend the # character to the beginning of the called number before extending the call to the MCU. This character is prepended on the route pattern(s) associated with the H.323 gateway representing the MCU. Calls to the MCU from SCCP clients will therefore also have this # character prepended to the calling number.

If the MCU is configured in **MCU** mode or is another vendor’s MCU that uses E.164 addresses for its conference IDs, then you do not have to prepend the # character. Also note that, if the MCU uses peer-to-peer mode and hence does not need to register its technology prefixes with any gatekeeper, then this situation does not apply and you do not have to prepend a # character.

Provisioning H.320 Gateways

As with H.323 MCUs, H.320 gateways are provisioned in Unified CM as H.323 gateways, and then route patterns are configured to extend calls to these devices. When provisioning an H.323 gateway, you must enter the static IP address and TCP signaling port of the H.320 gateway into the Device Name field. Unified CM then uses the static IP address and TCP port to reach the gateway any time a call matches the route pattern(s) associated with it.

Note The Cisco Unified Videoconferencing 3500 and 5000 Series Gateways do not listen on TCP port 1720 by default. (The Cisco Unified Videoconferencing 3500 and 5000 Series Gateways listen on port 1820 by default.) You must verify which TCP port they are listening on, and either change it to 1720 or provision the correct port in Unified CM.

If the gateway is configured to use peer-to-peer mode, then no further configuration is required. If the gateway requires RAS procedures to place calls to E.164 addresses, then you must also configure a dummy gatekeeper-controlled H.323 client in order to create the RASAggregator trunk, by filling in the fields for Device Name, Device Pool, Gatekeeper, Technology Prefix, and Zone Name as discussed in the section on Non-Gatekeeper Controlled Clients, page 12-27.

Gateway Service Prefixes

H.320 gateways use technology prefixes (also referred to as service prefixes in the gateway) as the prefix that users should dial to reach an ISDN destination. For calls to route correctly, you must configure the service prefix(s) of the gateway to begin with a # character. For example, if the gateway's service prefix that clients dial to reach an ISDN number is 9, then you must provision the service prefix on the gateway as #9. In this way, when H.323 clients dial 9 plus the PSTN number (such as 918005551212), the gatekeeper will not find a matching technology prefix registered and will instead route the call to the Unified CM trunk that is registered with the default technology prefix. Unified CM must then prepend the # character to the beginning of the called number before extending the call to the gateway. Note that, if the gateway uses peer-to-peer mode and hence does not need to register its technology prefixes with any gatekeeper, then this situation does not apply and you do not have to prepend a # character.

Gatekeeper Zone Configuration

The preceding sections discuss how to provision the endpoints in Unified CM Administration. You must also configure the endpoint gatekeeper(s) with the appropriate zone definitions. You must configure a zone for each type of endpoint (client, MCU, or gateway) and, optionally, for each device pool associated with these endpoints in Unified CM.

Each zone is configured to route all calls placed to, from, or within the zone to the RASAggregator trunk registered in that zone. You configure the zones on the endpoint gatekeeper by using the following command syntax:

zone local *<zone_name> <domain_name> <ip_address>* **invia** *<zone_name>*
outvia *<zone_name>* **enable-intrazone**

The command argument **invia** applies to calls placed to the zone from any other zone, **outvia** applies to calls placed from the zone to any other zone, and **enable-intrazone** applies to calls placed within the zone. The following sections illustrate the use of these commands.

Client Zones

The number of client zones you have to configure within each endpoint gatekeeper depends on the following factors:

- The device pools to which the H.323 clients are associated

  The device pool determines which Unified CM servers are primary, secondary, and tertiary servers for each H.323 client. If you assign all H.323 clients to the same device pool, then you need to define only a single client zone in the endpoint gatekeeper. In other words, for each device pool used by H.323 clients, you must configure a separate client zone in the gatekeeper.

- Whether the endpoint gatekeeper provides services for a single Unified CM cluster or multiple Unified CM clusters

  Each client zone is configured to route calls to a particular RASAggregator trunk. Therefore, if one endpoint gatekeeper is used to service multiple Unified CM clusters, then you must define a separate client zone for each cluster that the gatekeeper services.

To illustrate, the following examples show how client zones may be configured. Example 12-1 shows a single client zone defined for a single Unified CM cluster in which all H.323 clients are associated with the same device pool. Example 12-2 shows a single Unified CM cluster in which the H.323 clients are divided between two different device pools.

Note Some of the commands shown in the following examples are the default values applied in the Cisco IOS Gatekeeper and, therefore, would not have to be configured explicitly, nor would they appear in the running configuration. They are included here for thoroughness but are marked by a **!** at the beginning of the command line.

Example 12-1 Client Zone for a Single Unified CM Cluster and Single Device Pool

```
gatekeeper
zone local clients domain.com invia clients outvia clients enable-intrazone
gw-type-prefix 1# default-technology
no use-proxy clients default inbound-to terminal
no use-proxy clients default outbound-from terminal
! no arq reject-unknown-prefix
endpoint ttl 60
no shutdown
```

Example 12-2 Client Zones for a Single Unified CM Cluster and Two Device Pools

```
gatekeeper
zone local dp1-clients domain.com invia dp1-clients outvia dp1-clients enable-intrazone
zone local dp2-clients domain.com invia dp2-clients outvia dp2-clients enable-intrazone
gw-type-prefix 1# default-technology
```

```
no use-proxy dp1-clients default inbound-to terminal
no use-proxy dp1-clients default outbound-from terminal
no use-proxy dp2-clients default inbound-to terminal
no use-proxy dp2-clients default outbound-from terminal
! no arq reject-unknown-prefix
endpoint ttl 60
no shutdown
```

Disabling The Use of Proxy

The Cisco IOS Gatekeeper, formerly known as the Cisco Multimedia Conference Manager (MCM), previously offered an H.323 proxy function that has been at End of Life (EOL) for some time and is not compatible with Unified CM, but the commands in the gatekeeper to use a proxy for all calls to and from terminals (clients) are still enabled by default. You must disable this function for each client zone by using the following command syntax:

```
gatekeeper
no use-proxy <zone_name> default [inbound-to | outbound-from] terminals
```

The Cisco MCM proxy was replaced by a solution called the Cisco IOS Multiservice IP-to-IP Gateway and the associated via-zone-enabled Cisco IOS Gatekeeper. This document does not discuss the IP-to-IP Gateway, but Cisco Unified CM leverages the via-zone and IP-to-IP gateway constructs by registering its RASAggregator trunks with the gatekeeper, effectively mimicking an IP-to-IP gateway so that the gatekeeper will route all invia, outvia, and enable-intrazone calls to the RASAggregator trunk as if it were an IP-to-IP gateway.

Client Zone Prefixes

For H.323 client zones, there is no need to configure zone prefixes or technology prefixes of any kind, except for the default technology prefix. Instead, the **invia**, **outvia**, **enable-intrazone**, and **gw-type-prefix** *<1#>* **default-technology** commands ensure that all calls placed are routed to the RASAggregator trunk associated with the zone in which the call originated.

MCU Zones

The number of MCU zones you have to configure within each endpoint gatekeeper depends on the following factors:

- The device pools to which the MCUs are associated

  The device pool determines which Unified CM servers are primary, secondary, and tertiary servers for each MCU. If you assign all MCUs to the same device pool, then you need to define only a single MCU zone in the endpoint gatekeeper. In other words, for each device pool used by MCUs, you must configure a separate MCU zone in the gatekeeper.

- Whether the endpoint gatekeeper provides services for a single Unified CM cluster or multiple Unified CM clusters

  Each MCU zone is configured to route calls to a particular RASAggregator trunk. Therefore, if one endpoint gatekeeper is used to service multiple Unified CM clusters, then you must define a separate MCU zone for each cluster that the gatekeeper services.

Gatekeeper configuration for MCU zones is similar to the configurations shown in Example 12-1 and Example 12-2 for MCUs.

Disabling The Use of Proxy

By default, the Cisco IOS Gatekeeper is set to not use a proxy for calls to and from MCUs or gateways. However, if you have enabled the use of proxy for those types of endpoints, you must disable it for each MCU zone by using the following command syntax:

```
gatekeeper
no use-proxy <zone_name> default [inbound-to | outbound-from] [MCU | gateway]
```

If your MCU is registering as an MCU, then use the **MCU** argument at the end of the **no use-proxy** command; if your MCU is registering as a gateway, then use the **gateway** argument instead.

MCU Zone Prefixes

For H.323 MCU zones, there is no need to configure zone prefixes or technology prefixes of any kind, except for the default technology prefix. Instead, the **invia**, **outvia**, **enable-intrazone**, and **gw-type-prefix** *<1#>* **default-technology** commands ensure that all calls placed are routed to the RASAggregator trunk associated with the zone in which the call originated.

If your MCUs are registering their service prefixes as technology prefixes instead of E.164 addresses, use the special configuration described previously for prepending a # character to the MCU's service prefixes (see MCU Service Prefixes, page 12-28). Due to the way the Cisco IOS Gatekeeper selects a via-zone for calls to a technology prefix, when the endpoint dials the service prefix of the MCU, the call will fail if the gatekeeper finds a matching technology prefix registered. You must ensure that the client does not dial the # character, so that the gatekeeper will not find a matching technology prefix and will instead route the call to the RASAggregator trunk associated with the zone in which the call originated.

H.320 Gateway Zones

The number of H.320 gateway zones you have to configure within each endpoint gatekeeper depends on the following factors:

- The device pools to which the H.320 gateways are associated

  The device pool determines which Unified CM servers are primary, secondary, and tertiary servers for each H.320 gateway. If you assign all gateways to the same device pool, then you need to define only a single gateway zone in the endpoint gatekeeper. In other words, for each device pool used by H.320 gateways, you must configure a separate gateway zone in the gatekeeper.

- Whether the endpoint gatekeeper provides services for a single Unified CM cluster or multiple Unified CM clusters

  Each gateway zone is configured to route calls to a particular RASAggregator trunk. Therefore, if one endpoint gatekeeper is used to service multiple Unified CM clusters, then you must define a separate gateway zone for each cluster that the gatekeeper services.

Gatekeeper configuration for gateway zones is similar to the configurations shown in Example 12-1 and Example 12-2 for gateways.

Disabling The Use of Proxy

By default, the Cisco IOS Gatekeeper is set to not use a proxy for calls to and from gateways. However, if you have enabled the use of proxy for those types of endpoints, you must disable it for each H.320 gateway zone by using the following command syntax:

```
gatekeeper
no use-proxy <zone_name> default [inbound-to | outbound-from] gateway
```

Gateway Zone Prefixes

There is no need to configure zone prefixes of any kind for H.320 gateway zones. Instead, the **invia**, **outvia**, **enable-intrazone**, and **gw-type-prefix** *<1#>* **default-technology** commands ensure that all calls placed are routed to the RASAggregator trunk associated with the zone in which the call originated.

You must also use the special configuration described previously for prepending a # character to the gateway's service prefixes (see Gateway Service Prefixes, page 12-29). Due to the way the Cisco IOS Gatekeeper selects a via-zone for calls to a technology prefix, when the endpoint dials the service prefix of the gateway, the call will fail if the gatekeeper finds a matching technology prefix registered. You must ensure that the client does not dial the # character, so that the gatekeeper will not find a matching technology prefix and will instead route the call to the RASAggregator trunk associated with the zone in which the call originated.

Zone Subnets

As mentioned previously, the H.323 specification permits a single gatekeeper to manage multiple zones. However, the gatekeeper needs a way to decide which zone an endpoint should be placed in when it receives a Registration Request (RRQ) from that device. The RRQ message contains a Gatekeeper Identifier field that enables the endpoint to indicate the zone in which it would like to register. However, many H.323 video endpoints do not populate this field, and if the gatekeeper has multiple zones defined, it will not know which zone to place the endpoint into. Therefore, you must use of the **zone subnet** command to tell the gatekeeper which zone to associate with the endpoint. This command defines which IP addresses or IP address ranges are permitted to register in each zone. The command syntax requires that you enter a network mask. Therefore, you can specify either a particular host address by entering a 32-bit (/32) network mask or a range of addresses by specifying a smaller network mask.

Because MCUs, H.320 gateways, and Unified CM servers typically use fixed IP addresses but H.323 clients can use DHCP addresses, Cisco recommends that you define **zone subnet** commands only for the MCU and gateway zones but leave the client zones open so that any IP address is permitted in them. Note that you must also permit the Unified CM servers to register in the MCU and gateway zones, as illustrated in Example 12-3.

Note Some of the commands shown in the following example are the default values applied in the Cisco IOS Gatekeeper and, therefore, would not have to be configured explicitly, nor would they appear in the running configuration. They are included here for thoroughness but are marked by a **!** at the beginning of the command line.

Example 12-3 Defining Zone Subnets

```
gatekeeper
no zone subnet MCUs default enable
zone subnet MCUs [MCUs_IP_addr]/32 enable
zone subnet MCUs [RASAggregators_IP_addr]/32 enable
no zone subnet gateways default enable
zone subnet gateways [gateways_IP_addr]/32 enable
zone subnet gateways [RASAggregators_IP_addr]/32 enable
! zone subnet clients default enable
no zone subnet clients [MCUs_IP_addr]/32 enable
no zone subnet clients [gateways_IP_addr]/32 enable
```

The configuration in Example 12-3 explicitly permits the MCU and the RASAggregator for the MCU zone to register in the MCU zone, and it explicitly permits the gateway and RASAggregator for the gateway zone to register in the gateway zone. It also explicitly denies the MCU and gateway from registering in the client zone, while implicitly permitting all other IP addresses (including the RASAggregator for the client zone) to register in the client zone.

Endpoint Time to Live

Endpoints send lightweight Registration Requests (RRQs) to their gatekeeper periodically to maintain their registration status. The frequency with which they send these RRQs is referred to as the Time to Live (TTL) value. The endpoint may specify the TTL it wishes to use in the body of its RRQs. The gatekeeper may then honor the endpoint’s requested TTL value by echoing it in the Registration Confirm (RCF) response or, alternatively, may override the endpoint’s request by specifying a different TTL value in the RCF.

If the TTL value is not specified in the RRQ, the gatekeeper should specify one in its RCF response. The endpoint should then honor the TTL specified by the gatekeeper. The Cisco IOS Gatekeeper honors all TTL values specified by the endpoints. However, many H.323 video endpoints do not specify a TTL value in their RRQs. In such cases, the Cisco IOS Gatekeeper defaults to specifying a TTL value of 1800 seconds (30 minutes). The Cisco IOS Gatekeeper will flush the endpoint's registration after three TTL intervals have passed without receiving any messages from the endpoint (3 * 30 minutes = 90 minutes).

A large TTL value can cause problems with H.323 clients that do not use static IP addresses. For example, with the default TTL value of 1800 seconds, if you disconnect the client from the network and move it to another location in which it receives a different DHCP address, it will fail to register with the gatekeeper (Registration Reject (RRJ) cause value "duplicate alias") until three TTL intervals have passed, and the gatekeeper will flush that endpoint's original registration.

Therefore, Cisco recommends that you consider reducing the TTL value to as low a number as possible without causing any negative effect on your network. The Cisco IOS Gatekeeper permits you to set the TTL value anywhere in the range of 60 seconds to 3600 seconds. In most cases, 60 seconds should work well. However, if your gatekeeper is already heavily utilized, adjusting the TTL from the default of 1800 seconds to 60 seconds might cause it to become overwhelmed.

Use the following command syntax to set the TTL value:

```
gatekeeper
endpoint ttl <seconds>
```

Supported Gatekeeper Platforms

To act as an endpoint gatekeeper with Cisco Unified CM, the Cisco IOS Gatekeeper must run Cisco IOS Release 12.3(11)T or greater. For minimum Cisco IOS release requirements on the infrastructure gatekeeper, refer to the latest *Cisco Unified Communications System Release Notes for IP Telephony* available at:

http://www.cisco.com/go/unified-techinfo

To determine which release and feature set you should use for your router platform, use the Cisco Feature Navigator (requires a Cisco.com login account), available at:

http://tools.cisco.com/ITDIT/CFN/jsp/index.jsp

For more information, also refer to the *Cisco IOS H323 Gatekeeper Data Sheet*, available at:

http://www.cisco.com/en/US/prod/collateral/voicesw/ps6788/vcallcon/ps4139/data_sheet_c78_561921.html

Summary of Endpoint Gatekeepers

This section summarizes some key points to remember about endpoint gatekeepers and provides some example configurations that combine techniques used in the previous examples.

- Configure a separate zone in the endpoint gatekeeper for each type of endpoint (clients, MCUs, and H.320 gateways). If the endpoints are associated with multiple device pools, configure multiple zones for each type of endpoint.
- Configure a RASAggregator trunk to register in each zone. This trunk is automatically created when you configure gatekeeper-controlled H.323 clients in Unified CM Administration. However, for non-gatekeeper controlled H.323 clients, H.323 MCUs, and H.320 gateways, you must configure a dummy gatekeeper-controlled H.323 client in order to create the RASAggregator trunk for that zone.
- Set the service parameter **Send Product ID and Version ID** to **True** in order for the RASAggregator trunk to register with the gatekeeper as an IP-to-IP gateway. This setting enables the RASAggregator to be selected by the gatekeeper for all calls placed to, from, or within each zone due to the use of the **invia**, **outvia**, **enable-intrazone**, and **gw-type-prefix** *<1#>* **default-technology** commands applied to each local zone definition.
- You do not have to associate any zone prefixes for any of the endpoint zones. No matter what the endpoint dials, the gatekeeper should not find a matching zone prefix or technology prefix but should instead route the call to the RASAggregator trunk associated with the zone from which the call originated. To avoid having the gatekeeper accidently match the dialed number to the technology prefix of your MCUs or gateways, mask all MCU and gateway service prefixes with a # character, and then prepend the # character in the route pattern associated with that MCU or gateway.
- Configure zone subnets if any of the H.323 endpoints do not support the ability to specify the Gatekeeper Identifier (name of the zone) with which they wish to register.
- Disable the use of the old MCM Proxy for all zones.
- Set the endpoint registration Time to Live (TTL) to as low of a value as you can without creating undo stress on the gatekeeper. In extreme cases where the gatekeeper is serving hundreds of endpoint registrations, setting the TTL to 60 seconds might cause an unmanageable amount of RAS traffic. In smaller environments, setting it to 60 seconds should work well.

Example 12-4 shows a configuration for an endpoint gatekeeper servicing a single Unified CM cluster in which a single device pool is used to service all H.323 video endpoint types.

Note Some of the commands shown in the following examples are the default values applied in the Cisco IOS Gatekeeper and, therefore, would not have to be configured explicitly, nor would they appear in the running configuration. They are included here for thoroughness but are marked by a **!** at the beginning of the command line.

Example 12-4 Endpoint Gatekeeper Configuration for a Single Cluster and a Single Device Pool

```
gatekeeper
zone local clients domain.com invia clients outvia clients enable-intrazone
zone local MCUs domain.com invia MCUs outvia MCUs enable-intrazone
zone local gateways domain.com invia gateways outvia gateways enable-intrazone
! zone subnet clients default enable
no zone subnet clients [MCUs_IP_addr]/32 enable
no zone subnet clients [gateways_IP_addr]/32 enable
no zone subnet MCUs default enable
zone subnet MCUs [MCUs_IP_addr]/32 enable
zone subnet MCUs [RASAggregators_IP_addr]/32 enable
```

```
no zone subnet gateways default enable
zone subnet gateways [gateways_IP_addr]/32 enable
zone subnet gateways [RASAggregators_IP_addr]/32 enable
no use-proxy clients inbound-to terminals
no use-proxy clients outbound-from terminals
! no use-proxy MCUs inbound-to [MCU | gateway]
! no use-proxy MCUs outbound-from [MCU | gateway]
! no use-proxy gateways inbound-to gateway
! no use-proxy gateways outbound-from gateway
gw-type-prefix 1# default-technology
! no arq reject-unknown-prefix
endpoint ttl 60
no shutdown
```

Applications

Cisco IP Communications provides an expanding portfolio of applications that extend the features of Unified CM and provide advanced capabilities and integration with other communication media. Many of these applications can be used in conjunction with IP Video Telephony devices, even if they do not specifically support video. For instance, Cisco Unified CM does not support the negotiation of video channels for CTI applications using the TAPI/JTAPI protocols, but that does not necessarily preclude using a CTI application in conjunction with a video call. This section reviews some of the Cisco and third-party applications and discusses whether or not they can be used to provide advanced call treatment for video calls.

CTI Applications

The following applications are based on the Computer Telephony Integration (CTI) interface.

Cisco Emergency Responder

Cisco Emergency Responder (ER) routes emergency (911) calls to the correct Public Safety Answering Point (PSAP). It also provides the PSAP with the correct calling line ID of the originating device so that the PSAP can respond to the correct physical location of the incident and call the party back in the event that the call is disconnected. Cisco ER uses JTAPI to integrate with Unified CM. Emergency calls are routed to Cisco ER via a CTI route point, then Cisco ER decides which PSAP to forward the call to and what calling line ID to display. Cisco ER tracks each endpoint on the network to determine its physical location by using Simple Network Management Protocol (SNMP) and Cisco Discovery Protocol (CDP) to discover the physical port and specific Cisco Catalyst Ethernet switch to which the endpoint is connected. If CDP is not available, Cisco ER can be configured to locate endpoints by their IP subnet instead. Cisco ER then correlates this information with the physical location of the switch and stores the information in its database.

Cisco SCCP video devices support CDP for the purpose of Cisco ER discovery. Therefore, if a video telephony user dials 911, Cisco ER is able to route the call to the correct PSAP.

Because third-party SCCP video endpoints do not support CDP, Cisco ER must track these endpoints by their IP subnet. Cisco ER is therefore able to route the call to the correct PSAP.

Because H.323 videoconferencing clients do not support CDP, Cisco ER must track them by their IP subnet. Cisco ER is therefore able to route the call to the correct PSAP. However, the H.323 device must support the Empty Capabilities Set (ECS) procedure in order to have its call routed by Cisco ER. If the H.323 endpoint does not support receiving an ECS from Unified CM, calls to 911 that are handled by Cisco ER will fail.

When Cisco Cius is connected to Unified CM through a 3G or 4G connection, Cisco ER will not be able to locate the appropriate PSAP for the user's location. Cisco Cius users should be instructed to use alternate means to call 911 while roaming with 3G or 4G connections.

Cisco Unified IP Interactive Voice Response and Cisco Unified Contact Center

Cisco Unified IP Interactive Voice Response (Unified IP IVR) and Cisco Unified Contact Center (Unified CC) use JTAPI to integrate with Unified CM. If a video-capable device calls into an IVR application (such as a help desk), the communication is audio-only while the caller is connected to the application server (that is, while the caller browses the IVR menu or waits in queue for a help-desk member to take the call). However, once the IVR application transfers the call to its final destination, video channels can be negotiated at that time. H.323 devices must support the Empty Capabilities Set (ECS) procedure in order to interoperate with Cisco Unified IP IVR and Unified CC. If the H.323 endpoint does not support receiving an ECS from Unified CM, calls that are intercepted by Cisco Unified IP IVR or Unified CC will fail when the application attempts to transfer the caller to the final destination.

IVR applications often use DTMF tones to select options in the IVR menu. An alternative is speech recognition, which enables the caller to speak commands to the IVR server instead of pressing keys on the phone. Because Cisco Unified IP IVR and Unified CC both use JTAPI to integrate with Unified CM, they pass DTMF tones through out-of-band signaling messages. Many H.323 devices on the market today use in-band DTMF tones, and these H.323 clients would not be able to use DTMF to navigate an IP IVR or Unified CC menu. However, these H.323 clients could use speech recognition if the IVR server is enabled for it. SCCP video-capable devices, third-party SCCP video devices, and any H.323 endpoint that uses H.245 alphanumeric out-of-band signaling for DTMF, can navigate the IVR menus using DTMF tones.

Collaboration Solutions

The following technologies are sometimes used to provide video communications between endpoints.

T.120 Application Sharing

Some videoconferencing endpoints use the T.120 protocol to share documents, whiteboards, and text among participants in a conference. Unified CM does not support negotiating a T.120 channel. Instead of T.120, Cisco recommends using web-based collaboration solutions such as Cisco MeetingPlace or other third-party collaboration solutions.

Cisco Unified MeetingPlace

Cisco Unified MeetingPlace combines a high-end audio and video conferencing solution with a web-based front end for scheduling and participating in conferences. For more information, see Cisco Unified MeetingPlace, page 22-21.

Video Interoperability

Video interoperability is the audio and video support for point-to-point calls between Cisco TelePresence System (CTS) endpoints, other Cisco Unified Communications (UC) video endpoints, and third-party video endpoints. Prior to Cisco Unified CM 8.5, video interoperability between the different families of video endpoints was possible only with the insertion of a video component between endpoints, such as a video transcoder or a multipoint control unit (MCU).

Cisco Unified CM 8.5 and later releases not only offer native video interoperability between different video endpoint family types, point-to-point, but also provide better video interoperability in general with H.264 codec negotiation in SIP and H.323 protocols and enable the endpoints to negotiate high definition (HD) resolutions when available. Video interoperability, however, is dependent on he endpoints to support the interoperation.

As stated earlier, video interoperability in Unified CM also enables Cisco TelePresence System (CTS) endpoints to communicate with non-CTS endpoints, provided that the installed CTS software supports such interoperability. For further information, refer to *Interoperability Between CTS Endpoints and Other Cisco Endpoints or Devices*, available at

http://www.cisco.com/en/US/docs/telepresence/interop/endpoint_interop.html

Additionally, Cisco Unified CM 8.6 added scripting support for enhanced interoperability with call agents other than Unified CM. Through scripting, Unified CM has added support for the following features:

- SIP transparency — The ability to pass through known and unknown message components
- SIP normalization — Transformations on inbound and outbound SIP messages and content bodies

The primary motivation for video interoperability support is to facilitate the interaction of a diverse set of video endpoints without the need for deploying an expensive DSP infrastructure that would otherwise be required.

The following sections present general considerations and recommendations for the use of video interoperability:

- Video Interoperability Architecture, page 12-38
- Design Considerations for Video Interoperability, page 12-39

Video Interoperability Architecture

The video interoperability architecture includes the following elements:

- Video interoperability support is available only in Cisco Unified CM 8.5 and later releases.
- Two different video endpoint family types (Cisco TelePresence endpoints, Cisco UC endpoints, or third-party endpoints) engaged in a video call.

The following sections offer further information about the scope of the video interoperability support:

- Video Interoperability Test Cases, page 12-38
- Limitations of Video Interoperability, page 12-39

Video Interoperability Test Cases

In most cases a video endpoint that supports SIP or H.323 without using proprietary signaling would be able to interoperate with a Cisco UC video endpoint that supports video interoperability. For specific information on the scope of the interoperability between common sets of deployed devices and general information about the testing that was conducted to validate these more common examples of interoperability, refer to the *Cisco Unified Communications System Test Results for IP Telephony*, available at

> http://www.cisco.com/en/US/docs/voice_ip_comm/uc_system/unified/communications/system/ucstart.htm

Limitations of Video Interoperability

While video interoperability support attempts to enable any-to-any point-to-point video call interoperability, it is important to note that not all features of an individual video endpoint can be supported when interoperating with another endpoint. There are many reasons for this. For example, incompatibilities between different call control protocols could render a feature unavailable or offer a different representation of that feature. H.264 video media parameters can be represented differently in H.323 than in SIP, as another example. H.323 also does not have support for presence, but presence is quite commonly supported in SIP. Skinny Client Control Protocol (SCCP) does not have any notion of application sharing, which is commonly available in SIP and H.323 endpoint implementations. For instance, an SCCP user trying to share his/her PC screen would be hamper because Binary Flow Control Protocol (BFCP) and H.239 are not available in SCCP.

Design Considerations for Video Interoperability

Because Unified CM video interoperability decreases the reliance on video transcoding to achieve any-to-any video calling, some call flows might change substantially. That is not to say that video transcoding is unnecessary for all call flows, but it is no longer needed for the ones where the video interoperability capabilities of Unified CM can be employed. Therefore, enabling video interoperability reduces the need for DSP resources.

Additionally, the following areas should be considered when implementing the video interoperability capabilities of Unified CM:

- Guideline and Restrictions for Video Interoperability, page 12-39
- Quality of Service (QoS) and Call Admission Control Considerations for Video Interoperability, page 12-40

Guideline and Restrictions for Video Interoperability

The following guidelines and restrictions apply with regard to video interoperability in a Unified CM deployment:

- If H.323 or SCCP protocols are used in conjunction with video interoperability, Unified CM will support only a single H.264 payload and the packetization mode is treated as 0. An example side effect (but not the only one) of this circumstance is the fact that 1080p resolution is not available with these protocols because 1080p requires packetization mode 1.
- If multiple payloads are presented by an H.323 or SCCP endpoint engaged in a video interoperability call, Unified CM will use only the payload with the lowest codec profile. This, in turn, could result in less than the highest supported resolution being selected for the call.
- If a SIP endpoint omits the **level-asymmetry-allowed** parameter in the Session Description Protocol (SDP), Cisco products will assume that the endpoint can support asymmetric resolution transmission. Therefore, different receiving and sending video resolutions could be negotiated during a call.

- If a call is processed with video interoperability while Unified CM is performing protocol interworking with SIP and H.323, the H.323 endpoint must honor the proposed dynamic payload number specified by the SIP side, which means that no re-negotiation to a different payload would be supported.
- Unified CM will not negotiate Real-Time Transport Control Protocol (RTCP) feedback if the video call invokes a media termination point (MTP) or transcoder.

Quality of Service (QoS) and Call Admission Control Considerations for Video Interoperability

There are no changes to the configuration of regions and locations in Unified CM as a result of video interoperability support. However, regions play a significant role in determining the resolution between groups of endpoints, and they can be used to maximize or minimize the resolution that these devices use when interoperating. The **Max Video Call Bit Rate** field in the regions settings is used to determine the amount of bandwidth and, thus, the resolution that endpoints are able to negotiate.

For further information about QoS and call admission control with native video interoperability, see the section on Call Admission Control Design Recommendations for TelePresence Video Interoperability Architectures, page 11-109.

Wireless Networking Solutions

Because video is so bandwidth intensive, Cisco does *not* recommend using a shared wireless medium such as 802.11b/g for video endpoints.

Take care to ensure that video endpoints do not share the wireless bandwidth with any production IP Phones because video will consume much of the bandwidth and make it difficult to support video, audio, and data all on the same wireless medium.

Cisco recommends that you always make the physical Ethernet interface the preferred path. Also, when users connect to the PC port on the back of the IP Phone, instruct them to disable their Aironet Adapter to keep it from accidentally taking precedence.

CHAPTER 13

Gateways

Revised: October 31, 2012; OL-27282-05

Gateways provide a number of methods for connecting an IP telephony network to the Public Switched Telephone Network (PSTN), a legacy PBX, or key systems. Gateways range from specialized, entry-level and stand-alone voice gateways to high-end, feature-rich integrated router and Cisco Catalyst gateways.

This chapter explains important factors to consider when selecting a Cisco gateway to provide the appropriate protocol and feature support for your IP Telephony network. The main topics discussed in this chapter include:

- Traffic Patterns and Gateway Sizing, page 13-2
- TDM and VoIP Trunking Gateways, page 13-7
- Understanding Cisco Gateways, page 13-8
- Gateway Selection, page 13-9
- Fax and Modem Support, page 13-19
- Gateways for Video Telephony, page 13-27

What's New in This Chapter

Table 13-1 lists the topics that are new in this chapter or that have changed significantly from previous releases of this document.

Table 13-1 New or Changed Information Since the Previous Release of This Document

New or Revised Topic	Described in	Revision Date
No changes for Cisco Unified Communications System Release 9.0		June 28, 2012

Traffic Patterns and Gateway Sizing

This section presents a high-level discussion of the differences between various traffic models or patterns and how they can affect voice gateway selection. The emphasis is on traffic patterns and gateway sizing for traffic-intensive deployments.

Definitions and Terminology

This section uses the following terms and definitions:

- Simultaneous calls

  The number of calls that are all active in the system at the same time.

- Maximum simultaneous calls

  The maximum number of simultaneous calls in active (talk) state that the system can handle. The number of calls expected to be active simultaneously during the *busy hour* of the day should not exceed this number.

- Calls per second (cps)

  The call arrival rate, described as the number of calls that arrive (that is, new call setup attempts) in one second. Call arrival rates are also often quoted in calls per hour, but this metric is looser in the sense that 100 calls arriving in the last five seconds of an hour provides an average call arrival rate of 100 calls per hour (which is an extremely low rate for a communications system), while it also provides an arrival rate of 20 calls per second (which is a high rate). Sustaining 20 calls per second for an entire hour would result in 72,000 calls per hour. Therefore calls-per-hour is not a very useful metric for ascertaining a system's ability to handle bursty call arrival traffic patterns.

- Busy Hour Call Attempts (BHCA)

  The number of calls attempted during the busiest hour of the day (the peak hour). This is the same as the calls-per-second rating for the busiest hour of the day, but it is expressed over a period of an hour rather than a second. For example, 10 cps would be equal to 36,000 calls per hour. There is also a metric for Busy Hour Call Completions (BHCC), which can be lower than the BHCA (call attempts) under the assumption that not all calls are successful (as when a blocking factor exists). This chapter assumes 100% call completions, so that BHCA = BHCC.

- Bursty traffic

  Steady arrival means the call attempts are spaced more or less equally over a period of time. For example, 60 calls per hour at a steady arrival rate would present one call attempt roughly every minute (or approximately 0.02 cps). With bursty arrival, the calls arriving over a given period of time (such as an hour) are not spaced equally but are clumped together in one or more spikes. In the worst case, an arrival rate of 60 calls per hour could offer all 60 calls in a single second of the hour, thus averaging 0 cps for most of the hour with a peak of 60 cps for that one second. This kind of traffic is extremely stressful to communications systems.

- Hold time

  This is the period of "talk time" on a voice call; that is, the period of time between call setup and call teardown when there is an open speech path between the two parties. A hold time of 3 minutes (180 seconds) is an industry average used for traffic engineering of voice systems. The shorter the hold time on the average call, the greater the percentage of system CPU time spent on setting up and tearing down calls compared to the CPU time spent on maintaining the speech path.

PSTN Traffic Patterns

Traffic, when used in the context of voice communication systems, refers to the volume of calls being sent and/or received. Of particular importance is the traffic carried by external circuits such as the public switched telephone network (PSTN). Traffic is measured in Erlangs, and an Erlang is defined as one call lasting for one hour. This section does not go into any further detail on Erlangs other than to say that there are mathematical tables (Erlang-B and Erlang-C) that are used to calculate how many circuits are required for a given amount of offered traffic.

The amount of traffic received and generated by your business determines the size of the external circuits required. However; many customers typically continue to use the same number of circuits for their IP-based communications system as they previously used for a TDM-based system. While this sizing method might work if no issues are encountered, the process of ongoing system traffic analysis should be part of any routine maintenance practices. Traffic analysis can show that the system is over-provisioned for the current levels of traffic (and, therefore, the customer is paying for circuits that are not needed) or, conversely, that the system is under-provisioned and may be suffering from occasional blocked and/or lost calls, in which case increasing the number of circuits will remedy the situation.

Normal Business Traffic Profile

Most customers have a normal traffic profile, which means that they typically have two *busy hours* per day, one occurring during the morning from 10:00 to 11:00 and the other in the afternoon from 14:00 to 15:00. These busy-hour patterns can often be attributed to such things as employees starting the work day or returning from a lunch break. The calls themselves tend to have longer hold times and they tend to arrive and leave in a steady manner. A generally accepted industry average holding time to use for traffic calculations is 3 minutes.

Assuming that the communications system is engineered with the busy-hour traffic in mind, no issues should arise. Engineering a system below these levels will result in blocked and/or lost calls, which can have a detrimental effect on business.

Contact Center Traffic Profile

Contact centers present somewhat different patterns of traffic in that these systems typically handle large volumes of calls for the given number of agents or interactive voice response (IVR) systems available to service them. Contact centers want to get the most out of their resources, therefore their agents, trunks, and IVR systems are kept busy all the while they are in operation, which usually is 24 hours a day. Call queuing is typical (when incoming call traffic exceeds agent capacity, calls wait in queue for the next available agent), and the agents are usually dedicated during their work shifts to taking contact center calls.

Call holding times in contact centers are often of a shorter average duration than normal business calls. Contributing to the shorter average call holding time is the fact that many calls interact only with the IVR system and never need to speak to a human agent (also termed self-service calls). A representative holding time for self-service calls is about 30 seconds, while a call that talks to an agent has an average holding time of 3 minutes (the same as normal business traffic), making the overall average holding time in the contact center shorter than for normal business traffic.

The goal of contact centers to optimize resource use (including IVR ports, PSTN trunks, and human agents), combined with the fact that contact centers are systems dedicated to taking telephone calls, also presents the system with higher call arrival rates than in a typical business environment. These call arrival rates can also peak at different times of day and for different reasons (not the usual busy hour) than normal business traffic. For example, when a television advertisement runs for a particular holiday package with a 1-800 number, the call arrival rate for the system where those calls are received will experience a peak of traffic for about 15 minutes after the ad airs. This call arrival rate can exceed the average call arrival rate of the contact center by an order of magnitude.

Gateway Sizing for Contact Center Traffic

Short call durations as well as bursty call arrival rates impact the PSTN gateway's ability to process the traffic. Under these circumstances the gateway needs more resources to process all calls in a timely manner, as compared to gateways that receive calls of longer duration that are presented more uniformly over time. Because gateways have varying capabilities to deal with these traffic patterns, careful consideration should be given to selecting the appropriate gateway for the environment in which it will operate. Some gateways support more T1/E1 ports than others, and some are more able than others to deal with multiple calls arriving at the same time.

For a traffic pattern with multiple calls arriving in close proximity to each other (that is, high or bursty call arrival rates), a gateway with a suitable rating of calls per second (cps) is the best fit. Under these conditions, using calls with 15-second hold times, the Cisco AS5400XM Universal Gateway can maintain 16 cps with 250 calls active at once, the Cisco 3845 Integrated Services Router can maintain 13 cps with 200 calls active at once, and the Cisco 3945 Integrated Services Router can maintain 28 cps with 420 calls active at once. The performance of the Cisco AS5350XM Universal Gateway is identical to that of the AS5400XM in terms of calls per second.

For traffic patterns with a steady arrival rate, the maximum number of active calls that a gateway can handle is generally the more important consideration. Under these conditions, using calls with 180-second hold times, the Cisco AS5400XM Universal Gateway can maintain 600 simultaneously active calls with a call arrival rate of up to 3.3 cps, the Cisco 3845 Integrated Services Router can maintain 450 simultaneously active calls with a call arrival rate of up to 2.5 cps), and the Cisco 3945 Integrated Services Router can maintain 720 simultaneously active calls with a call arrival rate of up to 4 cps).

These numbers assume that all of the following conditions apply:

- CPU utilization does not exceed 75%.
- PSTN gateway calls are made with ISDN PRI trunks using H.323.
- Real Time Control Protocol (RTCP) timer is set to the default value of 5 seconds.
- Voice Activity Detection (VAD) is off.
- G.711 uses 20 ms packetization.
- Cisco IOS Release 15.0.1M is used.
- Dedicated voice gateway configurations are used, with ethernet (GE) egress and no QoS features. (Using QoS-enabled egress interfaces or non-ethernet egress interfaces, or both, will consume additional CPU resources.)
- No supplementary call features or services are enabled – such as general security (for example, access control lists or firewalls), voice-specific security (TLS, IPSec and/or SRTP), AAA lookups, gatekeeper-assisted call setups, VoiceXML or TCL-enabled call flows, call admission control (RSVP), and SNMP polling/logging. Such extra call features will use additional CPU resources.

Voice Activity Detection (VAD)

VAD is a digital signal processing feature that suppresses the creation of most of the IP packets during times when the speech path in a particular direction of the call is perceived as being silent. Typically only one party on a call speaks at a time, so that packets need flow in only one direction, and packets in the reverse (or silent) direction need not be sent except as an occasional keepalive measure. VAD can therefore provide significant savings in the number of IP packets sent for a VoIP call, and thereby save considerable CPU cycles on the gateway platform. While the actual packet savings that VAD can provide varies with the call flow, the application, and the nature of speaker interactions, it tends to use 10% to 30% fewer packets than would be sent for a call made using a VAD-off configuration.

VAD is most often turned off in endpoints and voice gateways deployed in Unified CM networks; VAD is most often turned on in voice gateways in other types of network deployments.

Codec

Both G.711 and G.729A use as their default configuration a 20 ms sampling time, which results in a 50 packets per second (pps) VoIP call in each direction. While a G.711 IP packet (200 bytes) is larger than a G.729A packet (60 bytes), this difference has not proven to have any significant effect on voice gateway CPU performance. Both G.711 and G.729 packets qualify as "small" IP packets to the router, therefore the packet rate is the salient codec parameter affecting CPU performance.

Performance Overload

Cisco IOS is designed to have some amount of CPU left over during peak processing, to handle interrupt-level events. The performance figures in this section are designed with the processor running at an average load of approximately 75%. If the load on a given Cisco IOS gateway continually exceeds this threshold, the following will result:

- The deployment will not be supported by Cisco Technical Assistance Center (TAC).
- The Cisco IOS Gateway will display anomalous behavior, including Q.921 timeouts, longer post-dial delay, and potentially interface flaps.

Cisco IOS Gateways are designed to handle a short burst of calls, but continual overloading of the recommended call rate (calls per second) is not supported.

Note With any gateway, you might be tempted to assign unused hardware ports to other tasks, such as on a CMM gateway where traffic calculations have dictated that only a portion of the ports can be used for PSTN traffic. However, the remaining ports *must* remain unused, otherwise the CPU will be driven beyond supported levels.

Performance Tuning

The CPU utilization of a Cisco IOS Voice gateway is affected by every process that is enabled in a chassis. Some of the lowest level processes such as IP routing and memory defragmentation will occur even when there is no live traffic on the chassis.

Lowering the CPU utilization can help to increase the performance of a Cisco IOS Voice Gateway by ensuring that there are enough available CPU resources to process the real-time voice packets and the call setup instructions. Some of the techniques for decreasing CPU utilization are described in Table 13-2.

Table 13-2 Techniques for Reducing CPU Utilization

Technique	CPU Savings	Description
Enable Voice Activity Detection (VAD)	Up to 20%	Enabling VAD can result in up to 45% fewer voice packets in typical conversations. The difficultly is that, in scenarios where voice recognition is used or there are long delays, a reduction in voice quality can occur. Voice appears to "pop" in at the beginning and "pop" out at the end of talk spurts.
Disable Real Time Control Protocol (RTCP)	Up to 5%	Disabling RTCP results in less out-of-band information being sent between the originating and terminating gateways. This results in lower quality of statistics displayed on the paired gateway. This can also result in the terminating gateway having a call "hang" for a longer period of time if RTCP packets are being used to determine if a call is no longer active.
Disable other non-essential functions such as: Authentication, Authorization, and Accounting (AAA); Simple Network Management Protocol (SNMP); and logging	Up to 2%	Any of these processes, when not required, can be disabled and will result in lower CPU utilization by freeing up the CPU to provide faster processing of real-time traffic.
Change call pattern to increase the length of the call (and reduce the number of calls per second)	Varies	This can be done by a variety of techniques such as including a long(er) introduction prompt played at the beginning of a call or adjusting the call script at the call center.

Additional Information

For more information on Cisco Voice Gateway capabilities and call center traffic analysis, refer to the following sources:

- Cisco Voice Gateway Solutions:

  http://www.cisco.com/en/US/products/sw/voicesw/index.html#~all-prod

- Gateway protocols supported with Cisco Unified Communications Manager (Unified CM):

  http://www.cisco.com/en/US/docs/voice_ip_comm/cucm/admin/8_0_1/ccmsys/a08gw.html

- Interfaces and signaling types supported by the following Cisco Voice Gateways:
  - Cisco 3900 Series Integrated Services Routers

    http://www.cisco.com/en/US/products/ps10536/products_relevant_interfaces_and_modules.html
  - Cisco 2900 Series Integrated Services Routers

    http://www.cisco.com/en/US/products/ps10537/products_relevant_interfaces_and_modules.html
  - Cisco 3800 Series Integrated Services Routers

    http://www.cisco.com/en/US/products/ps5855/products_relevant_interfaces_and_modules.html
  - Cisco 2800 Series Integrated Services Routers

    http://www.cisco.com/en/US/products/ps5854/products_relevant_interfaces_and_modules.html

- Gateway features supported with MGCP, SIP, and H.323:

  http://www.cisco.com/en/US/prod/collateral/routers/ps259/product_data_sheet0900aecd8057f2e0.pdf
- SIP gateway RFC compliance:

  http://www.cisco.com/en/US/prod/collateral/voicesw/ps6790/gatecont/ps6831/product_data_sheet0900aecd804110a2.html
- Skinny Client Control Protocol (SCCP) feature support with FXS gateways:

  http://www.cisco.com/en/US/prod/collateral/voicesw/ps6790/gatecont/ps2250/ps5516/product_data_sheet09186a00801d87f6.html
- Gateway capacities and minimum releases of Cisco IOS and Unified CM required for conferencing, transcoding, media termination point (MTP), MGCP, SIP, and H.323 gateway features:

  http://www.cisco.com/en/US/prod/collateral/routers/ps259/product_data_sheet0900aecd8057f2e0.pdf
- Various voice traffic calculators, including Erlang calculators:

  http://www.erlang.com/calculator/

Considerations for Gateway Redundancy

When deploying a gateway solution, give careful consideration to redundancy when compared with scalability. For example, a Cisco IOS gateway is capable of delivering multiple PRI interfaces on the same platform. However; considering the inherent need for redundancy with PSTN services, multiple smaller gateways delivering the same overall physical quantity of service may be more appropriate. Multiple gateways further allows for placement in different physical locations, thus allowing for another level of redundancy, in this case spatial redundancy.

Gateway deployments involving contact with emergency services must also be considered, and sometimes more than one solution may be necessary. For example, consider a small branch location connected to the PSTN through a SIP trunk located at a remote headquarters. If there is either a WAN or SIP trunk failure, the branch location must still be able to contact the emergency services. In this case the best solution would be either a local analog or PRI service (that is, either a standalone analog service or a PRI service terminated on the branch router).

TDM and VoIP Trunking Gateways

Until approximately 2006, the only choice for an enterprise to connect its internal VoIP network to voice services outside the enterprise was via TDM gateways to the traditional PSTN. Cisco offers a full range of TDM gateways with analog and digital connections to the PSTN as well as to PBXs and key systems. TDM connectivity covers a wide variety of low-density analog (FXS and FXO), low density digital (BRI), and high-density digital (T1, E1, and T3) interface choices.

Starting around 2006, new voice service options to an enterprise started to become available from service providers, most often referred to as SIP trunk services. Using a SIP trunk for connecting to PSTN and other destinations outside the enterprise involves an IP-to-IP connection at the edge of the enterprise's VoIP network. The same functions traditionally fulfilled by a TDM gateway are still needed at this interconnect point, including demarcation, call admission control, ensuring QoS, a troubleshooting boundary, security checks, and so forth. For SIP trunking connections, the Cisco Unified Border Element fulfils these functions as a session border controller (SBC) at the interconnect point between the enterprise and the service provider network. Cisco Unified Border Element also performs protocol translation functions to interconnect H.323 and SIP equipment, or to interconnect SIP equipment using different variations of SIP implementations. Cisco Unified Border Element can also perform transcoding. If used for one of these functions, Cisco Unified Border Element may also be used internal to the enterprise network at interconnect points between equipment that cannot interoperate without a protocol translation or transcoding service.

TDM gateway platforms are discussed in detail in the remainder of this chapter. Cisco Unified Border Element is discussed in greater detail in the chapter on Cisco Unified CM Trunks, page 14-1. Both functions can be enabled on the same Cisco Integrated Services Router (ISR) platform at the same time.

Understanding Cisco Gateways

Cisco access gateways enable Cisco Unified Communications Manager (Unified CM) to communicate with non-IP telecommunications devices. There are two types of Cisco access gateways, analog and digital.

Cisco Access Analog Gateways

There are two categories of Cisco access analog gateways, trunk gateways and station gateways.

- Access analog station gateways

  Analog station gateways connect Unified CM to Plain Old Telephone Service (POTS) analog telephones, interactive voice response (IVR) systems, fax machines, and voice mail systems. Station gateways provide Foreign Exchange Station (FXS) ports.

- Access analog trunk gateways

  Analog trunk gateways connect Unified CM to PSTN central office (CO) or PBX trunks. Trunk gateways provide Foreign Exchange Office (FXO) ports for access to the PSTN, PBXs, or key systems, and E&M (recEive and transMit, or ear and mouth) ports for analog trunk connection to a legacy PBX. Whenever possible, use digital gateways to minimize any answer and disconnect supervision issues. Analog Direct Inward Dialing (DID) and Centralized Automatic Message Accounting (CAMA) are also available for PSTN connectivity.

Cisco Access Digital Trunk Gateways

A Cisco access digital trunk gateway connects Unified CM to the PSTN or to a PBX via digital trunks such as Primary Rate Interface (PRI), Basic Rate Interface (BRI), or T1 Channel Associated Signaling (CAS). Digital T1 PRI trunks may also be used to connect to certain legacy voice mail systems.

Tuning Gateway Gain Settings

Connecting a Cisco Unified Communications network to the PSTN through gateways requires that you properly address voice quality issues arising from echo and signal degradation due to power loss, impedance mismatches, delay, and so forth. For this purpose, you must establish a Network Transmission Loss Plan (NTLP), which provides a complete picture of signal loss in all expected voice paths. Using this plan, you can identify locations where signal strength must be adjusted for optimum loudness and effective echo cancellation. Note that not all carriers use the same loss plan, and that the presence of cellular networks adds further complexity in creating the NTLP. Cisco does not recommend adjusting input gain and output attenuation on gateways without first completing such an NTLP. For more information, refer to *Echo Analysis for Voice Over IP*, available at

http://www.cisco.com/en/US/docs/ios/solutions_docs/voip_solutions/EA_ISD.pdf

Gateway Selection

When selecting an IP telephony gateway, consider the following factors:

- Core Feature Requirements, page 13-9
- Gateway Protocols, page 13-10
- Gateway Protocols and Core Feature Requirements, page 13-10
- Site-Specific Gateway Requirements, page 13-17

Core Feature Requirements

Gateways used in IP telephony applications must meet the following core feature requirements:

- Dual tone multifrequency (DTMF) relay capabilities

  DTMF relay capability, specifically out-of-band DTMF, separates DTMF digits from the voice stream and sends them as signaling indications through the gateway protocol (H.323, SCCP, MGCP, or SIP) signaling channel instead of as part of the voice stream or bearer traffic. Out-of-band DTMF is required when using a low bit-rate codec for voice compression because the potential exists for DTMF signal loss or distortion.

- Supplementary services support

  Supplementary services are typically basic telephony functions such as hold, transfer, and conferencing.

- Fax/modem support

  Fax over IP enables interoperability of traditional analog fax machines with IP telephony networks. The fax image is converted from an analog signal and is carried as digital data over the packet network. For more information, see Fax and Modem Support, page 13-19

- Unified CM redundancy support

  Cisco Unified Communications is based on a distributed model for high availability. Unified CM clusters provide for Unified CM redundancy. The gateways must support the ability to “re-home” to a secondary Unified CM in the event that a primary Unified CM fails. Redundancy differs from call survivability in the event of a Unified CM or network failure.

Refer to the gateway product documentation to verify that any IP Telephony gateway you select for an enterprise deployment can support the preceding core requirements. Additionally, every IP Telephony implementation has its own site-specific feature requirements, such as analog or digital access, DID, and capacity requirements (see Site-Specific Gateway Requirements, page 13-17).

Gateway Protocols

Cisco Unified CM (Release 3.1 and later) supports the following gateway protocols:

- H.323
- Media Gateway Control Protocol (MGCP)

Cisco Unified CM Release 4.0 and later supports Session Initiation Protocol (SIP) on the trunk side. The SIP trunk implementation has been enhanced in Cisco Unified CM releases 5.0 through 7.*x* to support more features.

Protocol selection depends on site-specific requirements and the installed base of equipment. For gateway configuration, MGCP might be preferred over H.323 or SIP due to simpler configuration. On the other hand, H.323 or SIP might be preferred over MGCP because of the robustness of the interfaces supported.

Simplified Message Desk Interface (SMDI) is a standard for integrating voice mail systems to PBXs or Centrex systems. Connecting to a voice mail system via SMDI and using either analog FXS or digital T1 PRI would require either SCCP or MGCP protocol because H.323 or SIP devices do not identify the specific line being used from a group of ports. Use of H.323 or SIP gateways for this purpose means the Cisco Message Interface cannot correctly correlate the SMDI information with the actual port or channel being used for an incoming call.

In addition, the Unified CM deployment model being used can influence gateway protocol selection. (Refer to the chapter on Unified Communications Deployment Models, page 5-1.)

Note Prior to deployment, check the Cisco IOS software release notes to confirm feature or interface support.

Gateway Protocols and Core Feature Requirements

This section describes how each protocol (SCCP, H.323, MGCP, and SIP) supports the following gateway feature requirements:

- DTMF Relay, page 13-10
- Supplementary Services, page 13-12
- Unified CM Redundancy, page 13-15

DTMF Relay

Dual-Tone Multifrequency (DTMF) is a signaling method that uses specific pairs of frequencies within the voice band for signals. A 64 kbps pulse code modulation (PCM) voice channel can carry these signals without difficulty. However, when using a low bite-rate codec for voice compression, the potential exists for DTMF signal loss or distortion. An out-of-band signaling method for carrying DTMF tones across a Voice over IP (VoIP) infrastructure provides an elegant solution for these codec-induced symptoms.

SCCP Gateways

The Cisco VG248 carries DTMF signals out-of-band using Transmission Control Protocol (TCP) port 2002. Out-of-band DTMF is the default gateway configuration mode for the VG248.

H.323 Gateways

The H.323 gateways, such as the Cisco 3800 series products, can communicate with Unified CM using the enhanced H.245 capability for exchanging DTMF signals out-of-band. The following is an example out-of-band DTMF configuration on a Cisco IOS gateway:

```
dial-peer voice 100 voip
destination-pattern 555….
session target ipv4:10.1.1.1
CODEC g729ar8
dtmf-relay h245-alphanumeric
preference 0
```

MGCP Gateway

The Cisco IOS-based platforms use MGCP for Unified CM communication. Within the MGCP protocol is the concept of *packages*. The MGCP gateway loads the DTMF package upon start-up. The MGCP gateway sends *symbols* over the control channel to represent any DTMF tones it receives. Unified CM then interprets these signals and passes on the DTMF signals, out-of-band, to the signaling endpoint. The global configuration command for DTMF relay is:

```
mgcp dtmf-relay VOIP codec all mode out-of-band
```

You must enter additional configuration parameters in the Unified CM MGCP gateway configuration interface.

DTMF relay is enabled by default and does not need additional configuration.

Note Use the **fm-package** command to enable DTMF via RFC 2833, available as of Cisco IOS Release 12.4(6)T.

SIP Gateway

The Cisco IOS-based platforms can use SIP for Unified CM communication. They support various methods for DTMF, but only the following methods can be used to communicate with Unified CM:

- Named Telephony Events (NTE), or RFC 2833
- Unsolicited SIP Notify (UN)
- Key Press Markup Language (KPML)

The following example shows a configuration for NTE:

```
dial-peer voice 100 voip
destination-pattern 555….
session target ipv4:10.1.1.1
session protocol sipv2
dtmf-relay rtp-nte
```

The following example shows a configuration for UN:

```
dial-peer voice 100 voip
destination-pattern 555….
session target ipv4:10.1.1.1
session protocol sipv2
dtmf-relay sip-notify
```

For more details on DTMF method selection, see the chapter on Media Resources, page 17-1.

Supplementary Services

Supplementary services provide user functions such as hold, transfer, and conferencing. These are considered fundamental requirements of any voice installation. Each gateway evaluated for use in an IP telephony network should provide support for supplementary services natively, without the use of a software media termination point (MTP).

SCCP Gateways

The Cisco SCCP gateways provide full supplementary service support. They also support FXS SCCP ports with Cisco IOS Release 12.4.9T. The SCCP gateways use the Gateway-to-Unified CM signaling channel and SCCP to exchange call control parameters.

H.323 Gateways

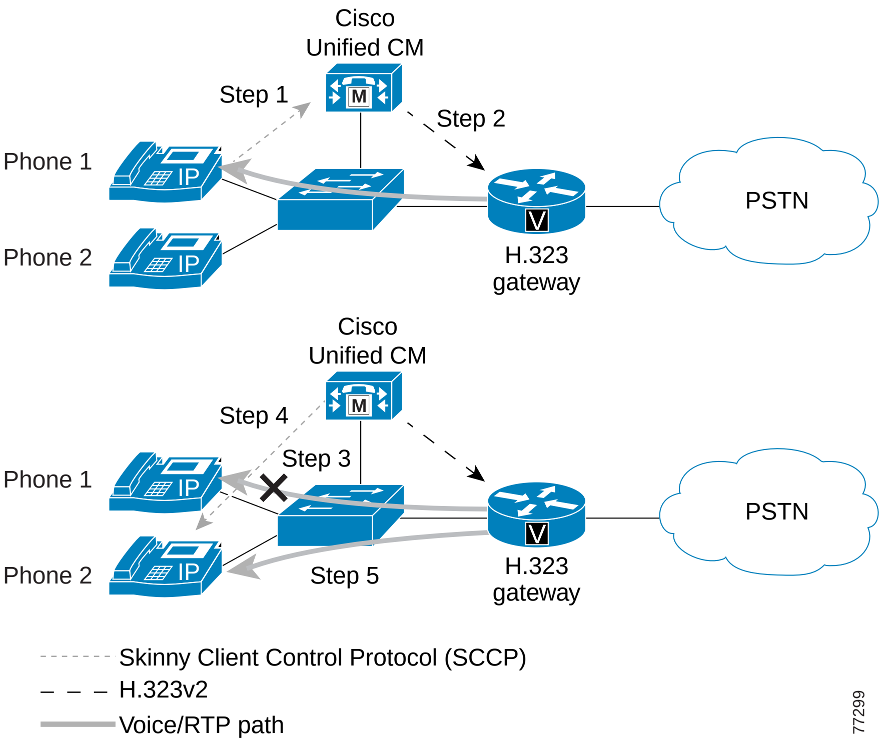
H.323v2 implements Open/Close LogicalChannel and the emptyCapabilitySet features. The use of H.323v2 by H.323 gateways, beginning in Cisco IOS Release 12.0(7)T and Cisco Unified CM Release 3.0 and later, eliminates the requirement for an MTP to provide supplementary services. With Unified CM Release 3.1 and later, a transcoder is allocated dynamically only if required during a call to provide access to G.711-only devices while still maintaining a G.729 stream across the WAN. Full support for H.323v2 is available in Cisco IOS Release 12.1.1T.

Once an H.323v2 call is set up between a Cisco IOS gateway and an IP phone, using the Unified CM as an H.323 proxy, the IP phone can request to modify the bearer connection. Because the Real-Time Transport Protocol (RTP) stream is directly connected to the IP phone from the Cisco IOS gateway, a supported voice codec can be negotiated.

Figure 13-1 and the following steps illustrate a call transfer between two IP phones:

1. If IP Phone 1 wishes to transfer the call from the Cisco IOS gateway to Phone 2, it issues a transfer request to Unified CM using SCCP.
2. Unified CM translates this request into an H.323v2 CloseLogicalChannel request to the Cisco IOS gateway for the appropriate SessionID.
3. The Cisco IOS gateway closes the RTP channel to Phone 1.
4. Unified CM issues a request to Phone 2, using SCCP, to set up an RTP connection to the Cisco IOS gateway. At the same time, Unified CM issues an OpenLogicalChannel request to the Cisco IOS gateway with the new destination parameters, but using the same SessionID.
5. After the Cisco IOS gateway acknowledges the request, an RTP voice bearer channel is established between Phone 2 and the Cisco IOS gateway.

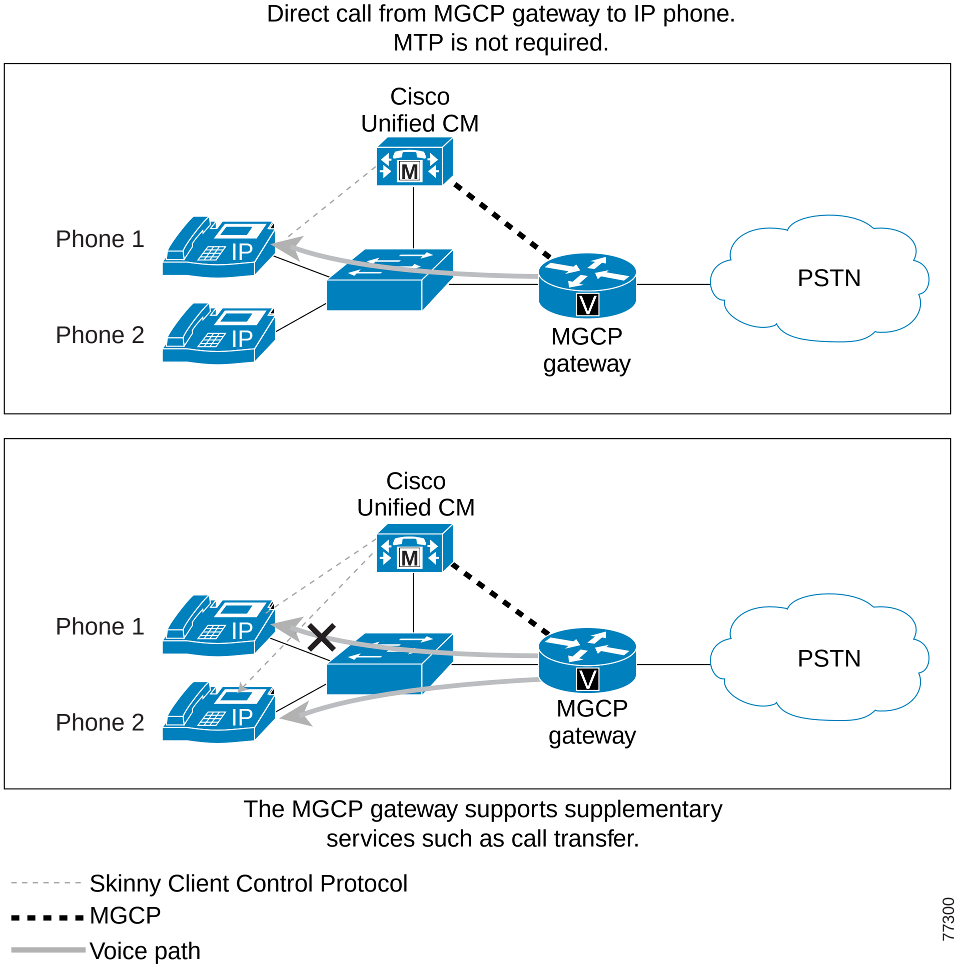
Figure 13-1 H.323 Gateway Supplementary Service Support



MGCP Gateway

The MGCP gateways provide full support for the hold, transfer, and conference features through the MGCP protocol. Because MGCP is a master/slave protocol with Unified CM controlling all session intelligence, Unified CM can easily manipulate MGCP gateway voice connections. If an IP telephony endpoint (for example, an IP phone) needs to modify the session (for example, transfer the call to another endpoint), the endpoint would notify Unified CM using SCCP. Unified CM then informs the MGCP gateway, using the MGCP User Datagram Protocol (UDP) control connection, to terminate the current RTP stream associated with the Session ID and to start a new media session with the new endpoint information. Figure 13-2 illustrates the protocols exchanged between the MGCP gateway, endpoints, and Unified CM.

Figure 13-2 MGCP Gateway Supplementary Service Support


Direct call from MGCP gateway to IP phone.
MTP is not required.
Cisco
Unified CM
M
Phone 1
IP
Phone 2
IP
V
MGCP
gateway
PSTN
Cisco
Unified CM
M
Phone 1
IP
Phone 2
IP
V
MGCP
gateway
PSTN
The MGCP gateway supports supplementary
services such as call transfer.
Skinny Client Control Protocol
MGCP
Voice path
77300

SIP Gateway

The Unified CM SIP trunk interface to Cisco IOS SIP gateways supports supplementary services such as hold, blind transfer, and attended transfer. The support for supplementary services is achieved via SIP methods such as INVITE and REFER. For more details, refer to the following documentation:

- *Cisco Unified Communications Manager System Guide*, available at

  http://www.cisco.com/en/US/products/sw/voicesw/ps556/prod_maintenance_guides_list.html
- *Cisco IOS SIP Configuration Guide*, available at

  http://www.cisco.com/en/US/docs/ios/voice/sip/configuration/guide/12_4t/sip_12_4t_book.html

Unified CM Redundancy

An integral piece of the IP telephony architecture is the provisioning of low-cost, distributed PC-based systems to replace expensive and proprietary legacy PBX systems. This distributed design lends itself to the robust fault tolerant architecture of clustered Unified CMs. Even in its most simplistic form (a two-system cluster), a secondary Unified CM should be able to pick up control of all gateways initially managed by the primary Unified CM.

SCCP Gateways

Upon boot-up, the Cisco VG224, VG248, and ATA 188 gateways are provisioned with Unified CM server information. When these gateways initialize, a list of Unified CMs is downloaded to the gateways. This list is prioritized into a primary Unified CM and secondary Unified CM. In the event that the primary Unified CM becomes unreachable, the gateway registers with the secondary Unified CM.

H.323 VoIP Call Preservation for WAN Link Failures

H.323 VoIP call preservation enhancements for WAN link failures sustain connectivity for H.323 topologies where signaling is handled by an entity that is different from the other endpoint, such as a gatekeeper that provides routed signaling or a call agent (such as the Cisco BTS 10200 Softswitch, Cisco PGW2200 Softswitch, or Cisco Unified CM) that brokers signaling between the two connected parties. Call preservation is useful when a gateway and the other endpoint (typically a Cisco Unified IP Phone) are located at the same site but the call agent is remote and therefore more likely to experience connectivity failures.

H.323 call preservation covers the following types of failures and connections.

Failure Types:

- WAN failures that include WAN links flapping or degraded WAN links.
- Cisco Unified CM software failure, such as when the ccm.exe service crashes on a Unified CM server.
- LAN connectivity failure, except when a failure occurs at the local branch.

Connection Types:

- Calls between two Cisco Unified CM controlled endpoints under the following conditions:
  - During Unified CM reloads.
  - When a Transmission Control Protocol (TCP) connection used for signaling H.225.0 or H.245 messages between one or both endpoints and Unified CM is lost or flapping.
  - Between endpoints that are registered to different Unified CMs in a cluster, and the TCP connection between the two Unified CMs is lost.
  - Between IP phones and the PSTN at the same site.
- Calls between a Cisco IOS gateway and an endpoint controlled by a softswitch, where the signaling (H.225.0, H.245 or both) flows between the gateway and the softswitch and media flows between the gateway and the endpoint:
  - When the softswitch reloads.
  - When the H.225.0 or H.245 TCP connection between the gateway and the softswitch is lost, and the softswitch does not clear the call on the endpoint.
  - When the H.225.0 or H.245 TCP connection between softswitch and the endpoint is lost, and the softswitch does not clear the call on the gateway.

- Call flows involving a Cisco Unified Border Element (formerly, Cisco Multiservice IP-to-IP Gateway) running in media flow-around mode that reload or lose connection with the rest of the network.

Note that, after the media is preserved, the call is torn down later when either one of the parties hangs up or media inactivity is detected. In cases where there is a machine-generated media stream, such as music streaming from a media server, the media inactivity detection will not work and the call might hang. Cisco Unified CM addresses such conditions by indicating to the gateway that such calls should not be preserved, but third-party devices or the Cisco Unified Border Element would not do this.

Flapping is defined for this feature as the repeated and temporary loss of IP connectivity, which can be caused by WAN or LAN failures. H.323 VoIP calls between a Cisco IOS gateway and Cisco Unified CM may be torn down when flapping occurs. When Unified CM detects that the TCP connection is lost, it clears the call and closes the TCP sockets used for the call by sending a TCP FIN, without sending an H.225.0 Release Complete or H.245 End Session message. This is called *quiet clearing*. The TCP FIN sent from Unified CM could reach the gateway if the network comes up for a short duration, and the gateway will tear down the call. Even if the TCP FIN does not reach the gateway, the TCP keepalives sent from the gateway could reach Unified CM when the network comes up. Unified CM will send TCP RST messages in response to the keepalives because it has already closed the TCP connection. The gateway will tear down H.323 calls if it receives the RST message.

Configuration of H.323 VoIP call preservation enhancements for WAN link failures involves configuring the **call preserve** command. If you are using Cisco Unified Communications Manager, you must enable the Allow Peer to Preserve H.323 Calls parameter from the Service Parameters window.

The **call preserve** command causes the gateway to ignore socket closure or socket errors on H.225.0 or H.245 connections for active calls, thus allowing the socket to be closed without tearing down calls using those connections.

Example of H.323 VoIP Call Preservation for All Calls

The following configuration example enables H.323 VoIP call preservation for all calls:

```
voice service voip
 h323
  call preserve
```

MGCP Gateway

MGCP gateways also have the ability to fail over to a secondary Unified CM in the event of communication loss with the primary Unified CM. When the failover occurs, active calls are preserved.

Within the MGCP gateway configuration file, the primary Unified CM is identified using the **call-agent <hostname>** command, and a list of secondary Unified CM is added using the **ccm-manager redundant-host** command. Keepalives with the primary Unified CM are through the MGCP application-level keepalive mechanism, whereby the MGCP gateway sends an empty MGCP notify (NTFY) message to Unified CM and waits for an acknowledgement. Keepalive with the backup Unified CMs is through the TCP keepalive mechanism.

If the primary Unified CM becomes available at a later time, the MGCP gateway can "re-home," or switch back to the original Unified CM. This re-homing can occur either immediately, after a configurable amount of time, or only when all connected sessions have been released. This is enabled through the following global configuration commands:

```
ccm-manager redundant-host <hostname1 | ipaddress1 > <hostname2 | ipaddress2>
[no] call-manager redundancy switchback [immediate|graceful|delay <delay_time>]
```

SIP Gateway

Redundancy with Cisco IOS SIP gateways can be achieved similarly to H.323. If the SIP gateway cannot establish a connection to the primary Unified CM, it tries a second Unified CM defined under another dial-peer statement with a higher preference.

By default the Cisco IOS SIP gateway transmits the SIP INVITE request 6 times to the Unified CM IP address configured under the dial-peer. If the SIP gateway does not receive a response from that Unified CM, it will try to contact the Unified CM configured under the other dial-peer with a higher preference.

Cisco IOS SIP gateways wait for the SIP 100 response to an INVITE for a period of 500 ms. By default, it can take up to 3 seconds for the Cisco IOS SIP gateway to reach the backup Unified CM. You can change the SIP INVITE retry attempts under the **sip-ua** configuration by using the command **retry invite** <*number*>. You can also change the period that the Cisco IOS SIP gateway waits for a SIP 100 response to a SIP INVITE request by using the command **timers trying** <*time*> under the **sip-ua** configuration.

One other way to speed up the failover to the backup Unified CM is to configure the command **monitor probe icmp-ping** under the **dial-peer** statement. If Unified CM does not respond to an Internet Control Message Protocol (ICMP) echo message (ping), the dial-peer will be shut down. This command is useful only when the Unified CM is not reachable. ICMP echo messages are sent every 10 seconds.

The following commands enable you to configure Unified CM redundancy on a Cisco IOS SIP gateway:

```
sip-ua
 retry invite <number>
 timers trying <time>

dial-peer voice 101 voip
 destination-pattern 2...
 session target ipv4:10.1.1.101
 preference 0
 monitor probe icmp-ping
 session protocol sipv2


dial-peer voice 102 voip
 destination-pattern 2...
 session target ipv4:10.1.1.102
 preference 1
 monitor probe icmp-ping
 session protocol sipv2
```

Site-Specific Gateway Requirements

Each IP Telephony implementation has its own site-specific requirements. The following questions can help you with IP Telephony gateway selection:

- Is the PSTN (or PBX) access analog or digital?
- What type of analog (FXO, FXS, E&M, DID, CAMA) or digital (T1, E1, CAS, CCS) interface is required for the PSTN or PBX?
- If the PSTN access is digital, what type of signaling is required (T1 CAS, Q.931 PRI, E1 CAS, or R2)?

- What type of signaling does the PBX currently use?
  - FXO or FXS: loop start or ground start
  - E&M: wink-start, delay-start, or immediate-start
  - E&M: type I, II, III, IV, or V
  - T1: CAS, Q.931 PRI (User-Side or Network-Side), QSIG, DPNSS, or Proprietary d-channel (CCS) signaling
  - E1: CAS, R2, Q.931 PRI (User-Side or Network-Side), QSIG, DPNSS, Proprietary d-channel (CCS) signaling
- What type of framing (SF, ESF, or G.704) and line encoding (B8ZS, AMI, CRC-4, or HDB3) does the PBX currently use?
- Does the PBX require passing proprietary signaling? If so, which time slot is the signaling passed on, and is it HDLC-framed?
- What is the required capacity of the gateway; that is, how many channels are required? (Typically, if 12 or more voice channels are required, then digital is more cost effective than an analog solution.)
- Is Direct Inward Dialing (DID) required? If so, specify analog or digital.
- Is Calling Line ID (CLID) needed?
- Is Calling Name needed?
- What types of fax and modem support are required?
- What types of voice compression are required?
- What types of supplementary services are required?
- Will the PBX provide clocking, or will it expect the Cisco gateway to provide clocking?
- Is rack space available for all needed gateways, routers, and switches?

Note Direct Inward Dial (DID) refers to a private branch exchange (PBX) or Centrex feature that permits external calls to be placed directly to a station line without use of an operator.

Note Calling Line Identification (CLI, CLID, or ANI) refers to a service available on digital phone networks to display the calling number to the called party. The central office equipment identifies the phone number of the caller, enabling information about the caller to be sent along with the call itself. CLID is synonymous with Automatic Number Identification (ANI).

Cisco Unified Communications gateways are able to inter-operate with most major PBX vendors, and they are EIA/TIA-464B compliant.

The site-specific and core gateway requirements are a good start to help narrow the possible choices. Once you have defined the required features, you can make a gateway selection for each of the pertinent configurations, whether they are single-site enterprise deployments of various sizes and complexities or multisite enterprise deployments.

The following tables summarize the features and interface types supported by the various Cisco gateway models.

Note In the following tables, the Cisco IOS and Unified CM release numbers refer to the minimum release that can support the listed feature on a particular gateway platform. For more information about Cisco IOS features, refer to the Cisco Feature Navigator located at http://tools.cisco.com/ITDIT/CFN/jsp/index.jsp.

Fax and Modem Support

This section describes the fax and modem support available with Unified CM and Cisco voice gateways. This section first presents brief overviews of fax and modem support on Cisco voice gateways, followed by a listing of supported platforms and example configuration files.

Gateway Support for Fax Passthrough and Fax Relay

Fax over IP enables interoperability of traditional analog fax machines with IP Telephony networks. The fax image is converted from an analog signal and is carried digitally over the packet network.

In its original form, fax data is digital and is contained in High-Level Data Link Control (HDLC) frames. However, to transmit across a traditional PSTN, these digital HDLC frames are modulated onto an analog carrier. While this analog carrier is necessary for effective faxing in PSTN environments, it is not ideal for the type of digital transport used by IP packet networks. Therefore, specific transport methods have been devised for successful transport of fax transmissions over an IP infrastructure.

The two main methods for transporting fax over IP are passthrough and relay. Passthrough is the simplest method, and it works by sampling and digitizing the analog fax signal just like a voice codec does for human speech. While there are a number of codecs available, passthrough on Cisco voice gateways always uses the G.711 codec for carrying fax information because it offers the least distortion of the analog fax signals. If a high-compression codec is being used by the original voice call, then passthrough uses an upspeed feature to change the codec to G.711. Passthrough is also commonly referred to as Voice Band Data (VBD), and Cisco provides two versions of passthrough: modem passthrough and fax pass-through. The names of these two passthrough versions are derived from how they are configured in the Cisco IOS command line interface (CLI). Additional differences between these passthrough versions center around their switchovers and triggering tones, and these are discussed in more detail in the following paragraphs.

Modem passthrough typically uses Cisco proprietary Named Signaling Event (NSE) packets to switch the call from voice mode to passthrough mode in what is commonly termed an NSE-based switchover. This switchover from voice mode to passthrough is an important concept for fax pass-through and relay as well. Every call on a Cisco voice gateway starts as a voice call, and the proper switchover occurs only when the gateway determines that the call is truly a fax call.

The modem passthrough feature is triggered by a 2100 Hz CED or ANSam tone at the beginning of a fax or modem call. The CED tone is associated with G3 faxes and low-speed modems, while the ANSam tone is used by SG3 faxes and high-speed modems. Historically, when the ANSam or CED tone was detected, modem passthrough used Cisco proprietary NSE packets to signal the remote voice gateway of the switchover from voice mode to modem passthrough. Now, however, in addition to an NSE-based switchover, modem passthrough also supports a protocol-based switchover using the H.323 or SIP call control protocols. When modem passthrough is configured to handle the switchover using H.323 or SIP, it also will use a standards-based NTE message to optionally signal the remote voice gateway to disable its echo cancellers. These enhancements to modem passthrough allow for increased interoperability with third-party devices and are found in Cisco IOS Release 12.4(24)T and higher.

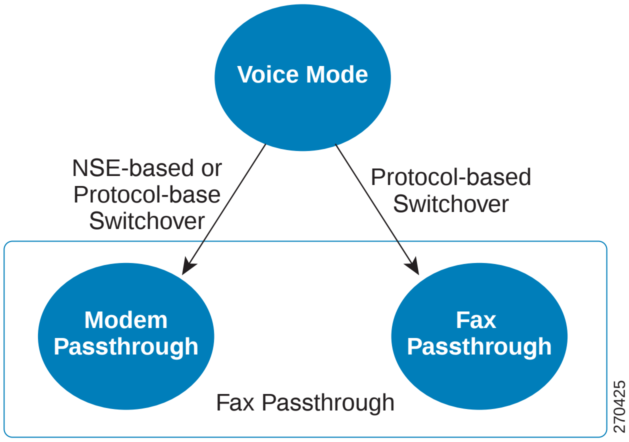
Despite its name, modem passthrough is also widely used for fax calls. You can activate it in the Cisco IOS command line interface (CLI) by using the **modem passthrough** command for H.323, SIP, and SCCP voice gateways or the **mgcp modem passthrough** command for MGCP voice gateways.

Fax pass-through does not support an NSE-based switchover as modem passthrough does. Instead, it always relies on the underlying call control protocol to switch the call from voice mode to fax pass-through. Fax pass-through supports only a protocol-based switchover using the call control protocols of H.323 and SIP. Because fax pass-through utilizes the call control protocol for its switchover, it will typically interoperate with third-party devices.

Fax pass-though is triggered by the detection of V.21 flags associated with G3 fax calls. Therefore, this transport method does not work for modems or SG3 fax calls. The command to enable fax pass-through on H.323 and SIP voice gateways is **fax protocol pass-through**.

Figure 13-3 highlights the two different passthrough implementations employed by Cisco voice gateways for fax calls.

Figure 13-3 Cisco Passthrough Implementations for Fax Calls

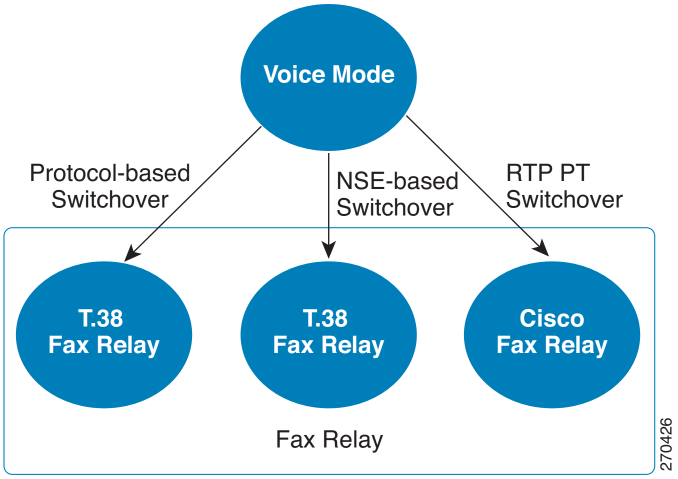


Relay is the other main method for transporting fax over IP, and its implementation is a bit more complicated than passthrough. Relay strips off the analog carrier from the fax signal in a process known as *demodulation* to expose the fax HDLC data frames. The pertinent information in these HDLC frames is then removed and efficiently packaged in a fax relay protocol to be transported to the gateway on the other side. When received on the other side, the fax information is pulled from the relay protocol, reconstructed back into fax HDLC frames, and modulated onto an analog carrier for transmission to a fax machine.

Cisco supports two versions of fax relay, T.38 and Cisco fax relay. An ITU standard, T.38 allows Cisco gateways to interoperate with third-party devices that also support the T.38 specification. In most scenarios, T.38 fax relay uses the call control protocol to switch from voice mode to T.38 fax relay mode, and this is referred to as protocol-based or standards-based T.38 fax relay. However, it is also possible to configure T.38 fax relay to switch over using Cisco proprietary NSEs in what is termed NSE-based T.38 fax relay. To ensure third-party interoperability, protocol-based T.38 must be utilized.

Cisco fax relay is a pre-standard implementation, and it is proprietary to Cisco voice gateways. It is also the default fax transport configuration on nearly all Cisco voice gateways. Unlike the NSE or protocol-based methods used by T.38 fax relay and passthrough, Cisco fax relay transitions from voice to relay mode utilizing specific RTP dynamic payload types (PT). Figure 13-4 illustrates the Cisco fax relay methods.

Figure 13-4 Cisco Relay Implementations for Fax Calls



Fax relay mode, and more specifically T.38, is the preferred method to transport fax traffic. However, if T.38 fax relay is not supported, then Cisco fax relay or passthrough can be used as an alternative.

Best Practices

The following recommendations and guidelines can assist you in best implementing fax support on Cisco voice gateways:

- When using QoS, make every effort to minimize the following:
  - Packet loss
  - Delay
  - Delay variation (jitter)

  All transmissions of fax over IP are extremely sensitive to packet loss. Even minimal packet loss can cause fax failures. If packet loss is a problem in your network, then you should use the redundancy feature in T.38 fax relay. Also, ensure that constant packet delay on the network does not exceed 1 second and that delay variation (jitter) does not exceed 300 milliseconds for T.38 and Cisco fax relay. When passthrough is used, the jitter should follow VoIP design best practices and not exceed 30 ms. For detailed information about implementing QoS in a Cisco Unified Communications network, refer to the *Enterprise QoS Solution Reference Network Design Guide*, available at

  http://www.cisco.com/go/designzone

- The following tips can help ensure the integrity of the fax calls:
  - Use call admission control (CAC) to ensure that calls are not admitted if they exceed the specified total bandwidth limit. The following table lists approximate fax call bandwidth usage for the common fax transport methods.

Fax Transport Method	Bandwidth[1]
Modem Passthrough or Fax Pass-through (G.711)	83 kbps
Modem Passthrough with Redundancy	170 kbps
T.38 (no redundancy)	25 kbps

Fax Transport Method	Bandwidth[1]
T.38 (high-speed redundancy level set to 1)	41 kbps
T.38 (high-speed redundancy level set to 2)	57 kbps
Cisco Fax Relay	48 kbps

[1] Bandwidth values are approximate with Ethernet or Frame Relay L2 headers. T.38 and Cisco fax relay bandwidth values are peak and occur only during the sending of a fax page at 14.4 kbps.

  – Disable call waiting on all dedicated modem and fax ports.

- T.38 fax relay provides the best fax performance based on network considerations and is the recommended transport method for fax traffic.

  To insure interoperability with other vendor's T.38 products, use protocol-based T.38.

  NSE-based T.38 must be used for communicating with certain Cisco voice gateways, such as the Cisco VG248 and any Cisco IOS SCCP gateways. For older versions of Unified CM with limited support for protocol-based T.38, NSE-based T.38 fax relay is a valid alternative.

  In Unified CM scenarios where T.38 is to be deployed among gateways running a variety of call signaling protocols, protocol-based T.38 should be the first choice. The latest release of Cisco Unified CM supports protocol-based T.38 with H.323, SIP, and MGCP call control protocols. If protocol-based T.38 is not supported in your installed version of Cisco Unified CM or if SCCP gateways are involved, then NSE-based T.38 should be used. To verify if your version of Unified CM supports protocol-based T.38, refer to the Cisco Unified Communications Manager release notes available at

  http://www.cisco.com/en/US/products/sw/voicesw/ps556/prod_release_notes_list.html

- T.38 fax relay is supported on most of the current Cisco voice gateways, especially those running Cisco IOS. For details, refer to the product data sheets for your specific gateway models.

- Error Correction Mode (ECM) is a negotiated feature on fax calls, and it ensures that fax pages are received error-free. However, in its effort to retransmit any errors, ECM can lead to increased fax transmission times and call failures. If desired, you can disable ECM on the gateway itself rather than disabling it on multiple fax machines. However, if packet drops or other IP or PSTN impairments occur, the fax image quality might deteriorate. Therefore, you should disable ECM only after considering whether you want to risk compromising image quality rather than experiencing longer call durations or dropped calls. You should also monitor and evaluate the network to identify and resolve the impairments that are causing the fax page errors.

Super Group 3 Fax Support

Commonly referred to as "high-speed" fax or V.34 fax, the Super Group 3 (SG3) classification uses V.34 modulation to increase the maximum fax page transmission speed to 33.6 kbps. SG3 fax machines are backward compatible with standard G3 fax machines that support a maximum page transmission speed of 14.4 kbps.

Cisco IOS gateways with Cisco IOS Release 12.4.4T and later offer support for Super Group 3 (SG3) fax transmissions when T.38 or Cisco fax relay are configured; however, only Group 3 speeds are negotiated. For more information on this feature, refer to *Cisco IOS Fax, Modem, and Text Support over IP Configuration Guide, Cisco IOS Release 15.1M&T*, available at

http://www.cisco.com/en/US/docs/ios/voice/fax/configuration/guide/15_1/vf_15_1_book.html

If it is necessary to transport SG3 high-speed faxes at their native speeds, then modem passthrough must be used. With the release of Cisco IOS version 15.1.1T, a new feature will provide native support for SG3 faxes by T.38 fax relay.

Gateway Support for Modem Passthrough and Modem Relay

In general, there are three mechanisms for supporting modem sessions over an IP network using voice gateways:

- Modem passthrough
- Cisco Modem Relay
- Secure Modem Relay (Secure Communication Between STE Endpoints)

Each of these mechanisms can transport modem calls, but the relay methods are restrictive in that only certain modem modulations are supported. Modem passthrough, on the other hand, can usually handle any modulation.

An important concept to understand when dealing with the transport of modem signals across IP networks is the switchover that must occur on the gateway. Every call on a Cisco gateway begins as a voice call initially. Even if the call is between modems, the call will be set up as a voice call first. Then, once the gateway is sure that the call is truly a modem call, a switchover occurs that converts the gateway from voice call mode to a modem passthrough or modem relay mode. There are various switchover methods to transition a call from voice mode to modem passthrough or relay.

As discussed previously in the section on Gateway Support for Fax Passthrough and Fax Relay, page 13-19, modem passthrough uses proprietary NSE packets or the H.323/SIP protocol stack to switch a voice call into passthrough mode. When modem signals are detected, the gateways can use these NSE messages to inform each other of the impending modem call. Special messages within the H.323 or SIP call control protocols can also be used. The gateways then make adjustments to better handle the transport of the modem signals. These adjustments include up-speeding the voice codec to G.711, disabling Voice Activity Detection (VAD), and disabling the echo cancellers if necessary. Because modem passthrough simply samples the analog modem signal using the G.711 codec, it should handle any modem modulation, but not always at the highest speeds.

Cisco Modem Relay is a proprietary implementation that efficiently transports V.34 modem calls over an IP network. V.90 calls are also supported, but they are forced to train down to V.34 speeds. As with modem passthrough, NSE packets are used to handle the switchover to Cisco Modem Relay from voice mode.

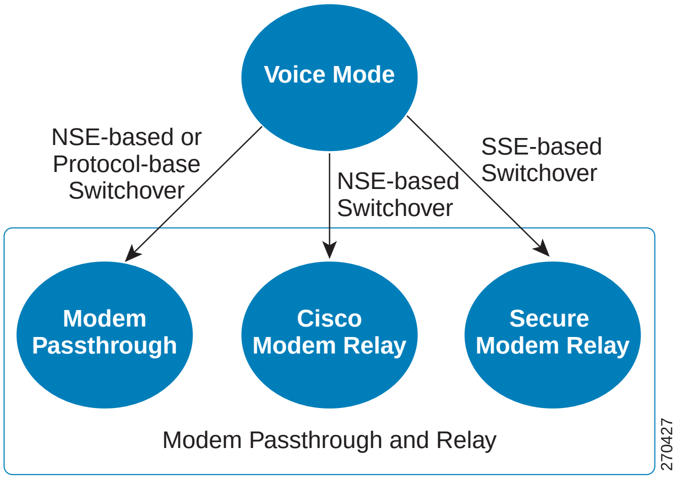
Secure modem relay, which is also referred to as Secure Communication Between STE Endpoints, allows for the transport of secure telephone calls over an IP infrastructure. Special devices known as Secure Terminal Equipment (STE) transmit encrypted voice using the V.32 modulation. Secure modem relay is designed to handle the transport of information between STEs in Unified CM environments with SCCP and MGCP gateways. Secure modem relay is not compatible with Cisco Modem Relay. One of the main reasons is that the switchover for secure modem relay does not use NSEs but instead uses V.150.1-based State Signaling Event (SSE) messages.

Secure modem relay is designed specifically for transporting STE signals and is almost never used outside of government or defense-related deployments. In most cases, Cisco Modem Relay or modem passthrough should be used for transporting modem calls. For more information on secure modem relay, refer to *Secure Communication Between IP-STE Endpoint and Line-Side STE Endpoint*, available at

http://www.cisco.com/en/US/docs/ios/12_4t/12_4t4/htv1501.html

Figure 13-5 summarizes the Cisco modem transport implementations. Modem relay should be used whenever possible because it offers the most bandwidth efficiency and tolerance for network impairments when compared to modem passthrough. The disadvantage of modem relay is that it is quite restrictive on the modulations supported, while modem passthrough can handle any modem modulation.

Figure 13-5 Cisco Passthrough and Relay Implementations for Modem Calls



Best Practices

Observe the following recommended best practices to ensure optimum performance of modem traffic transported over an IP infrastructure:

- Ensure that the IP network is enable for Quality of Service (QoS) and that you adhere to all of the recommendations for providing QoS in the LAN, MAN, and WAN environments. Every effort should be made to minimize the following parameters:
  - Packet loss — Fax and modem traffic requires an essentially loss-free transport. A single lost packet can result in retransmissions.
  - Delay
  - Delay variation (jitter)

  For more information, refer to the *Enterprise QoS Solution Reference Network Design Guide*, available at

  http://www.cisco.com/go/designzone
- Use call admission control (CAC) to ensure that calls are not admitted if they exceed the specified total bandwidth limit. For planning purposes, assume that modem passthrough calls will always consume approximately 83 kbps of bandwidth, or 170 kbps with redundancy enabled, regardless of the modem modulation being transported. Modem relay bandwidth is sporadic because of the nature of modem communications, but plan for peaks of about 45 kbps for the maximum V.34 connection speed of 33.6 kbps. The bandwidth values cited here are approximations and assume Ethernet or frame relay as the L2 transport.
- Use modem relay whenever possible. Modem passthrough should be used for any modulations not supported by modem relay.

- Do not use the IP network to connect modems that will be used to troubleshoot or diagnose problems with the IP network. In this case, the modems used to troubleshoot the devices that compose the IP infrastructure should be connected to a plain old telephone service (POTS).
- Because of the NSE switchover utilized by Cisco modem relay and modem passthrough, gateways using different call control protocols can easily communicate with one another. For example, an MGCP gateway and an H.323 gateway connected to Unified CM can successfully negotiate Cisco modem relay or modem passthrough because the NSE switchover occurs within the RTP voice media stream that has already been set up by Unified CM.
- Disable call waiting on all dedicated modem and fax ports.

V.90 Support

Currently, Cisco equipment supports only V.34 modems. Although V.90 modems will function on existing hardware, and speeds higher than V.34 speeds can be achieved, full V.90 support cannot be guaranteed.

Supported Platforms and Features

The following Cisco platforms support fax and modem features:

- Cisco IOS Gateways support:
  - Modem passthrough
  - Fax passthrough for the H.323 and SIP protocols
  - T.38 fax relay. Both NSE and protocol-based switchovers for T.38 are supported, except in the case of SCCP where only NSE-based T.38 fax relay is supported.
  - Cisco fax relay. The Cisco AS5350, AS5400, and AS5850 using Nextport DSP cards do not support Cisco fax relay. The PVDM3 DSP modules also do not support Cisco fax relay.
  - Cisco modem relay
- Cisco non-IOS gateways:
  - The Cisco VG248 supports modem passthrough, NSE-based T.38 fax relay, and Cisco fax relay.
  - The Cisco 6608 and 6624 support only modem passthrough and Cisco fax relay.
  - The Cisco ATAs support modem passthrough for fax calls only. Using modem passthrough with an ATA for modem calls is not officially supported.

Note The fax and modem support information presented here is valid beginning with Cisco IOS Release 12.4(9)T for the Cisco IOS gateways and Release 1.3.1 of the Cisco VG248 Analog Phone Gateway.

Platform Protocol Support

Common call control protocols used today in enterprise solutions include H.323, Session Initiation Protocol (SIP), Media Gateway Control Protocol (MGCP), and Skinny Client Control Protocol (SCCP). Not all Cisco voice platforms support all of these protocols or all of the fax and modem features, thus raising interoperability issues. Additional interoperability issues occur when mixing Cisco IOS gateways such as the Cisco 2800 Series or the Cisco 3800 Series with non-IOS gateways such as the VG248. This section lists the combinations of gateways that provide support for interoperability of fax, modem, and protocol features.

Some of the common combinations of protocols in a network include: MGCP and H.323; SCCP and H.323; SCCP and SIP; MGCP and SIP; H.323 and SIP; and SCCP and MGCP.

Table 13-3 lists the protocol combinations that currently support fax and modem interoperability.

Table 13-3 Fax and Modem Features Supported with Various Combinations of Call Control Protocols

Protocol Combinations	Modem Relay	Modem Passthrough[1]	T.38 Fax Relay	Cisco Fax Relay	Fax Passthrough
Unified CM using MGCP combined with Unified CM using H.323 or SIP	Yes	Yes	Yes[2]	Yes	No
Unified CM using MGCP combined with Unified CM using MGCP	Yes	Yes	Yes[2]	Yes	No
SCCP combined with Unified CM using H.323 or SIP	Yes	Yes	Yes[3]	Yes	No
SCCP combined with Unified CM using MGCP	Yes	Yes	Yes[3]	Yes	No
Unified CM using H.323 combined with H.323 or SIP	Yes	Yes	Yes[2]	Yes	Yes
Unified CM using SIP combined with H.323 or SIP	Yes	Yes	Yes[2]	Yes	Yes

[1] Modem passthrough works for both modem and fax passthrough calls.

[2] NSE-based T.38 fax relay works, but protocol-based T.38 fax relay depends on the version of Unified CM. For version information, refer to the Cisco Unified Communications Manager release notes available at http://www.cisco.com/en/US/products/sw/voicesw/ps556/prod_release_notes_list.html.

[3] SCCP protocol works only with NSE-based T.38 fax relay.

Note Table 13-3 is a general reference. You should be aware that specific products might have limitations that are not listed in this table. For example, the Cisco ATA supports H.323, SIP, and SCCP call control protocols, but only modem passthrough is supported no matter which call control protocol is used.

Gateway Configuration Examples

For detailed configuration information about fax and modem support on Cisco gateways, refer to the *Cisco IOS Fax, Modem, and Text Support over IP Configuration Guide*, which is available at

http://www.cisco.com/en/US/docs/ios/voice/fax/configuration/guide/12_4t/vf_12_4t_book.html

Gateways for Video Telephony

Video gateways terminate video calls into an IP telephony network or the PSTN. Video gateways are different from voice gateways because they have to interact with the ISDN trunks that support video and convert that call to a video call on the IP network using protocols such as H.323 or SIP. Enterprises can consider separate gateways for voice calls and video calls, or they can have integrated gateways that route both voice and video calls.

The following key considerations can help you decide if you need separate gateways for voice and video or an integrated gateway:

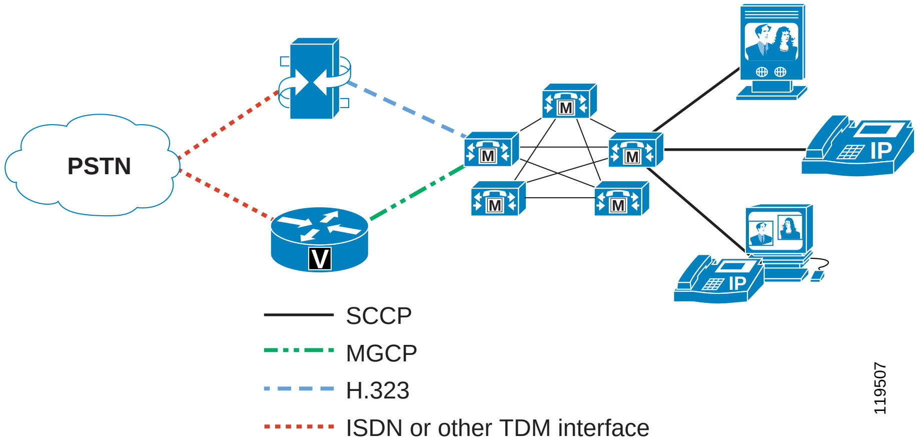
- Dial plan — If the enterprise has the flexibility of a separate dial plan for video users, it can use separate video gateways that allow it to keep existing enterprise dial plans.
- Video users — If the enterprise has a large number of users who primarily use voice rather than video, then Cisco recommends using separate video gateways to service the video call users.
- Locations — If the enterprise has a large number of distributed locations with video users at many locations, then Cisco recommends using an integrated gateway to reduce total cost of ownership (TCO).
- Additional video capabilities such as video IVR, auto attendant, and bonding across trunks — Dedicated video gateways might support advanced features that integrated gateways do not support.
- Protocol — Gateway protocol can be an important factor to align with enterprise policies and standards.
- Capacity — Dedicated gateways might support lower simultaneous call volumes, while integrated gateways should have higher capacity because they can support voice calls in addition to video.
- Device management — Ease of maintenance, management, and troubleshooting can be an important factor. Dedicated gateways provide a better user interface (GUI) for management and configuration, while integrated gateways can provide better troubleshooting. However, these factors are dependent on the respective products.

Dedicated Video Gateways

Enterprises that have an extensive voice infrastructure with voice gateways can add video gateways so that users can make video calls through them to the PSTN. The Cisco Unified Videoconferencing 3500 and 5200 Series Video Gateways can be used for that purpose.

Figure 13-6 shows an enterprise deployment that can use existing protocols for its voice gateways and add video gateways so that Unified CM users can make voice and video calls to the PSTN.

Figure 13-6 Unified CM System with Separate PSTN Lines for Voice and IP Video Telephony



The Unified Videoconferencing gateways, while excellent for video calls, do not support all of the features that Cisco Voice Gateways offer. The Unified Videoconferencing gateways have the following characteristics:

- They support only H.323 and H.320.
- They are standalone devices that cannot be integrated into Cisco IOS Routers or Cisco Catalyst Switches.
- They support only T1/E1-PRI and ISDN BRI.
- They support H.261, H.263, and H.264 video codecs
- They support only G.711, G.722, G.722.1, G.723.1 and G.728; they do not support G.729 audio.
- They support the H.245 Empty Capabilities Set (ECS).
- They support T.120 and H.239 data sharing protocol
- They support H.235 encryption
- They do not support many of the manageability and troubleshooting capabilities inherent in Cisco Voice Gateways.

As a result of these differences in the products, Cisco Unified Videoconferencing 3500 Series gateways are not recommended as replacements for Cisco Voice Gateways. IP Telephony customers who want to add video to their communications environment should deploy both types of gateways and use the Cisco Voice Gateways for all voice calls and use the Cisco Unified Videoconferencing 3500 Series gateways for video calls only. Customers might also have to procure separate circuits for voice and video from their PSTN service provider, depending on which model of Cisco IOS Gateway is deployed.

With separate voice and video gateways (see Figure 13-6), the route plans must also be separate for both inbound and outbound calls. For inbound calls, there is no way to have a single Direct Inward Dial (DID) extension for a user who wants to be able to receive both voice and video calls. Typically, each user will already have a DID for voice calls. When you introduce video into the scenario, users will have to be dialed some other way, such as via a second DID number or by dialing the main number of the video gateway and then entering the users video extension when prompted by the Interactive Voice Response (IVR). For outbound calls, there is no way to have a single PSTN access code for both voice and video calls. Typically, users will already have a well-known access code for voice (such as 9 in most US enterprises), but when you introduce video into the scenario, they will have to dial some other access code to place outbound video calls.

Another consideration for deploying two types of gateways is the placement of those gateways. Typically, enterprises have many PSTN gateway resources consolidated at their central site(s), and each branch office has some local gateway resources as well. For instance, Cisco Catalyst 6500 gateways may be deployed at the central site with several T1/E1 circuits connected to them, while Cisco Integrated Services Routers (ISRs) may be deployed at each branch office with either analog or digital trunks to the local CO. When video is introduced into this scenario, the customer must also determine the number of PSTN circuits they will need for video and where the video gateways will be placed. For instance, will they deploy only a few Cisco Unified Videoconferencing video gateways at the central site, or will they also deploy them at each branch office?

Finally, consider how calls will be routed across the IP network to a remote gateway for the purpose of providing toll bypass, and how calls will be re-routed over the PSTN in the event that the IP network is unavailable or does not have enough bandwidth to complete the call. More specifically, do you want to invoke automated alternate routing (AAR) for video calls?

Integrated Video Gateways

Enterprises may consider an integrated device for voice and video gateway functionality. This provides the enterprise the advantages of managing fewer devices and keeping the dial plan simple. The gateway processes the call as a voice call if it is voice and as a video call if it is video.

The Cisco IOS Integrated Video Gateway has the following characteristics:

- Supports Cisco ISO-13871 bonding
- Provides H.320, H.323, and SIP support
- Supports existing voice codec and H.254 video codec
- Provides extensive called and calling transformation capabilities
- Provides extensive logging and troubleshooting capabilities

The following considerations apply for deploying Cisco IOS Integrated Video gateways:

- Consider the capacity needed on PSTN links for additional video calls.
- Consider the need of devices to use data applications such as T.120 and the additional bandwidth that will be used on the IP network.
- Consider if users need features such far-end camera control or DTMF that is used for conferences that the H.320 gateway needs to support.

Routing Inbound Calls from the PSTN

Use one of the following methods to route inbound calls from the PSTN:

- Assign at least two different directory numbers to each video-enabled device in the Unified CM cluster, with one line for audio and another line for video. With this method, the outside (PSTN) caller must dial the correct number to enable video.
- For video calls, have outside callers dial the main number of the video gateway.
  Cisco Unified Videoconferencing gateways offer an integrated IVR that prompts the caller to enter the extension number of the party they are trying to reach. Unified CM will then recognize that it is

a video call when ringing the destination device. This method relieves the caller from having to remember two different DID numbers for each called party, but it adds an extra step to dialing an inbound video call.

Note The outside video endpoints must support DTMF in order to enter the extension of the called party at the IVR prompt.

The following example illustrates the second method:

> A user has a Cisco Unified IP Phone 7960 attached to a PC running Cisco Unified Video Advantage. The extension of the IP Phone is 51212, and the fully qualified DID number is 1-408-555-1212. To reach the user from the PSTN for a voice-only call, people simply dial the DID number. The CO sends calls to that DID number through T1-PRI circuit(s) connected to a Cisco Voice Gateway. When the call is received by the gateway, Unified CM knows that the gateway is capable of audio only, so it negotiates only a single audio channel for that call. Conversely, for people to reach the user from the PSTN for a video call, they must dial the main number of the video gateway and then enter the user's extension. For example, they might dial 1-408-555-1000. The CO would send calls to that number through the T1-PRI circuit(s) connected to a Cisco Unified Videoconferencing 3500 Series video gateway. When the call is received by the gateway, an IVR prompt asks the caller to enter the extension of the person they are trying to reach. When the caller enters the extension via DTMF tones, Unified CM knows that the gateway is capable of video, so it negotiates both audio and video channels for that call.

Gateway Digit Manipulation

The Cisco Unified Videoconferencing 3500 Series Gateways cannot manipulate digits for calls received from the PSTN. It takes the exact number of digits passed to it in the Q.931 Called Party Number field and sends them all to Unified CM. Therefore, Unified CM must manipulate the digits in order to match the directory number (DN) of the destination device. For instance, if the circuit from the CO switch to the gateway is configured to pass 10 digits but the extension of the called party is only five digits, Unified CM must strip off the leading five digits before attempting to find a matching DN. You can implement this digit manipulation in one of the following ways:

- By configuring the Significant Digits field on the H.323 gateway device or on the H.225 gatekeeper-controlled trunk that carries the incoming calls from the Cisco Unified Videoconferencing gateway

  This method enables you to instruct Unified CM to pay attention to only the least-significant N digits of the called number. For example, setting the Significant Digits to 5 will cause Unified CM to ignore all but the last 5 digits of the called number. This is the easiest approach, but it affects all calls received from that gateway. Thus, if you have variable-length extension numbers, this is not the recommended approach.

- By configuring a translation pattern and placing it in the calling search space of the H.323 gateway device or of the H.225 gatekeeper-controlled trunk that carries the incoming calls from the Cisco Unified Videoconferencing gateway

  This method enables Unified CM to match calls to the full number of digits received, to modify the called number, and then to continue performing digit analysis on the resulting modified number. This approach is slightly more complex than the preceding method, but it is more flexible and enables you to use a finer granularity for matching calls and for specifying how they will be modified.

Routing Outbound Calls to the PSTN

Use one of the following methods to route outbound calls to the PSTN:

- Assign different access codes (that is, different route patterns) for voice and video calls. For example, when the user dials 9 followed by the PSTN telephone number they are trying to reach, it could match a route pattern that directs the call out a voice gateway. Similarly, the digit 8 could be used for the route pattern that directs calls out a video gateway.
- Assign at least two different directory numbers on each video-enabled device in the Unified CM cluster, with one line for audio and another line for video. The two lines can then be given different calling search spaces. When users dial the access code (9, for example) on the first line, it could be directed out a voice gateway, while dialing the same access code on the second line could direct the call out a video gateway. This method alleviates the need for users to remember two different access codes but requires them to press the correct line on their phones when placing calls.

Gateway Service Prefixes

The Cisco Unified Videoconferencing Gateways use service prefixes to define the speed for outbound calls. When you configure a service prefix in the gateway, you must choose one of the following speeds:

- Voice-only
- 128 kbps
- 256 kbps
- 384 kbps
- 768 kbps
- Auto (dynamically determined; supports any call speed in the range of 128 kbps to 768 kbps)

Note Each of the above speeds represents a multiple of 64 kbps. For 56-kbps dialing, there is a check-box on the service prefix configuration page to restrict each channel to 56 kbps. Therefore, a 128-kbps service with restricted mode enabled would result in a 112-kbps service; a 384 kbps service with restricted mode enabled would result in a 336-kbps service; and so on.

Calls from an IP endpoint toward the PSTN must include the service prefix at the beginning of the called number in order for the gateway to decide which service to use for the call. Optionally, you can configure the default prefix to be used for calls that do not include a service prefix at the beginning of the number. This method can become quite complex because users will have to remember which prefix to dial for the speed of the call they wish to make, and you would have to configure multiple route patterns in Unified CM (one for each speed). Fortunately, the Auto speed enables you to minimize this effort. If the majority of your calls are made using 64 kbps per channel (for example, 128 kbps, 384 kbps, 512 kbps, 768 kbps, and so on), you could use the Auto service in that case. You would then need to create only one other service for the rare case in which someone makes a call using 56 kbps per channel (for example, 112 kbps, 336 kbps, and so on).

Cisco recommends that you always use a # character in your service prefixes because the gateway recognizes the # as an end-of-dialing character. By placing this character in the service prefix, you block people from attempting to use the gateway for toll fraud by dialing the main number of the gateway, reaching the IVR, and then dialing out to an off-net number. The # can either be at the beginning (recommended) or the end of the service prefix. For example, if your access code to reach the PSTN is 8 for video calls, Cisco recommends that you configure the service prefix as #8 or 8#. Or, if you have two service prefixes as described above, you might use #80 for the Auto 64-kbps service and #81 for the Auto 56-kbps service.

The ramification of using a service prefix is that Unified CM must prepend the service prefix to the called number when sending calls to the Cisco Unified Videoconferencing gateway. Because forcing users to dial the # would not be very user-friendly, Cisco recommends that you configure Unified CM to prepend the # to the dialed number. For example, if the access code to dial a video call to the PSTN is 8, you could configure a route pattern as 8.@ in Unified CM, and in the route pattern configuration you would configure the called number translation rule to prepend #8 whenever that route pattern is dialed. Or, if you have two service prefixes as described above, you might use 80.@ for the Auto 64-kbps service (prefixing # to the called number) and 81.@ for the Auto 56-kbps service (prefixing # to the called number).

Automated Alternate Routing (AAR)

When the IP network does not have enough bandwidth available to process a call, Unified CM uses its call admission control mechanism to determine what to do with the call. As described in the chapter on IP Video Telephony, page 12-1, Unified CM performs one of the following actions with the call, depending on how you have configured it:

- Fail the call, playing busy tone to the caller and displaying a Bandwidth Unavailable message on the caller’s screen
- Retry the video call as an audio-only call
- Use automated alternate routing (AAR) to re-route the call over an alternative path, such as a PSTN gateway

The first two options are covered in the chapter on IP Video Telephony, page 12-1, and this section covers the AAR option.

To provide AAR for voice or video calls, you must configure the calling and called devices as members of an AAR group and configure an External Phone Number Mask for the called device. The External Phone Number Mask designates the fully qualified E.164 address for the user’s extension, and the AAR group indicates what digits should be prepended to the External Phone Number Mask of the called device in order for the call to route successfully over the PSTN.

For example, assume that user A is in the San Jose AAR group and user B is in the San Francisco AAR group. User B's extension is 51212, and the External Phone Number Mask is 6505551212. The AAR groups are configured to prepend 91 for calls between the San Jose and San Francisco AAR groups. Thus, if user A dials 51212 and there is not enough bandwidth available to process the call over the IP WAN between those two sites, Unified CM will take user B's External Phone Number Mask of 6505551212, prepend 91 to it, and generate a new call to 916505551212 using the AAR calling search space for user A.

This same logic applies to video calls as well, with one additional step in the process. For video-capable devices, there is field called Retry Video Call as Audio. As described in the chapter on IP Video Telephony, page 12-1, if this option is enabled (checked), Unified CM does not perform AAR but retries the same call (that is, the call to 51212) as a voice-only call instead. If this option is disabled (unchecked), Unified CM performs AAR. By default, all video-capable devices in Unified CM have the Retry Video Call as Audio option enabled (checked). Therefore, to provide AAR for video calls, you must disable (uncheck) the Retry Video Call as Audio option. Additionally, if a call admission control policy based on Resource Reservation Protocol (RSVP) is being used between locations, the RSVP policy must be set to Mandatory for both the audio and video streams.

Furthermore, Unified CM looks at only the called device to determine whether the Retry Video Call as Audio option is enabled or disabled. So in the scenario above, user B's phone would have to have the Retry Video Call as Audio option disabled in order for the AAR process to take place.

Finally, devices can belong to only one AAR group. Because the AAR groups determine which digits to prepend, AAR groups also influence which gateway will be used for the rerouted call. Depending on your choice of configuration for outbound call routing to the PSTN, as discussed in the previous section, video calls that are rerouted by AAR might go out a voice gateway instead of a video gateway. Therefore, carefully construct the AAR groups and the AAR calling search spaces to ensure that the correct digits are prepended and that the correct calling search space is used for AAR calls.

While these considerations can make AAR quite complex to configure in a large enterprise environment, AAR is easier to implement when the endpoints are strictly of one type or the other (such as IP Phones for audio-only calls and systems such as the Tandberg T-1000 dedicated for video calls). When endpoints are capable of both audio and video calls (such as Cisco Unified Video Advantage or a Cisco IP Video Phone 7985G), the configuration of AAR can quickly become unwieldy. Therefore, Cisco recommends that large enterprise customers who have a mixture of voice and video endpoints give careful thought to the importance of AAR for each user, and use AAR only for select video devices such as dedicated videoconference rooms or executive video systems. Table 13-4 lists scenarios when it is appropriate to use AAR with various device types.

Table 13-4 When to Use AAR with a Particular Device Type

Device Type	Device is used to call:	Enable AAR?	Comments
IP Phone	Other IP Phones and video-capable devices	Yes	Even when calling a video-capable device, the source device is capable of audio-only, thus AAR can be configured to route calls out a voice gateway.
IP Phone with Cisco Unified Video Advantage, or Cisco IP Video Phone 7985G	Other video-capable devices only	Yes	Because the device is used strictly for video calls, you can configure the AAR groups accordingly.
	IP Phones and other video-capable devices	No	It will be difficult to configure the AAR groups to route audio-only calls differently than video calls.
Sony or Tandberg SCCP endpoint	Other video-capable devices only	Yes	Because the device is used strictly for video calls, you can configure the AAR groups accordingly.
	IP Phones and other video-capable devices	No	It will be difficult to configure the AAR groups to route audio-only calls differently than video calls.
H.323 or SIP client	Other video-capable devices only	Yes	Because the device is used strictly for video calls, you can configure the AAR groups accordingly.
	IP Phones and other video-capable devices	No	It will be difficult to configure the AAR groups to route audio-only calls differently than video calls

Least-Cost Routing

Least-cost routing (LCR) and tail-end hop-off (TEHO) are very popular in VoIP networks and can be used successfully for video calls as well. In general, both terms refer to a way of configuring the call routing rules so that calls to a long-distance number are routed over the IP network to the gateway closest to the destination, in an effort to reduce toll charges. (For Cisco Unified CM Release 4.1, LCR basically means the same thing as TEHO.) Unified CM supports this feature through its rich set of digit analysis and digit manipulation capabilities, including:

- Partitions and calling search spaces
- Translation patterns
- Route patterns and route filters
- Route lists and route groups

Configuring LCR for video calls is somewhat more complicated than for voice calls, for the following reasons:

- Video calls require their own dedicated gateways, as discussed previously in this chapter
- Video calls require much more bandwidth than voice calls

With respect to dedicated gateways, the logic behind why you might or might not decide to use LCR for video calls is very similar to that explained in the section on Automated Alternate Routing (AAR), page 13-32. Due to the need to have different types of gateways for voice and video, it can become quite complex to configure all the necessary partitions, calling search spaces, translation patterns, route patterns, route filters, route lists, and route groups needed for LCR to route voice calls out one gateway and video calls out another.

With respect to bandwidth requirements, the decision to use LCR depends on whether or not you have enough available bandwidth on your IP network to support LCR for video calls to/from a given location. If the current bandwidth is not sufficient, then you have to determine whether the benefits of video calls are worth the cost of either upgrading your IP network to make room for video calls or deploying local gateways and routing calls over the PSTN. For example, suppose you have a central site with a branch office connected to it via a 1.544-Mbps T1 Frame Relay circuit. The branch office has twenty video-enabled users in it. A 1.544-Mbps T1 circuit can handle at most about four 384-kbps video calls. Would it really make sense in this case to route video calls up to the central site in order to save on toll charges? Depending on the number of calls you want to support, you might have to upgrade your 1.544-Mbps T1 circuit to something faster. Is video an important enough application to justify the additional monthly charges for this upgrade? If not, it might make more sense to deploy a Cisco Unified Videoconferencing gateway at the branch office and not bother with LCR. However, placing local Cisco Unified Videoconferencing gateways at each branch office is not inexpensive either, so ultimately you must decide how important video-to-PSTN calls are to your business. If video is not critical, perhaps it is not worth upgrading the bandwidth or buying video gateways but, instead, using the Retry Video Call as Audio feature to reroute video calls as voice-only calls if they exceed the available bandwidth. Once a call is downgraded to voice-only, local gateway resources and bandwidth to perform LCR become more affordable and easier to configure.

ISDN B-Channel Binding, Rollover, and Busy Out

With Cisco IOS Release 12.4.20T or later releases, Cisco IOS H.320 gateways support the ISO-13871 bonding technique, which supports video calls at speeds up to 1 Mbps for video calls. With this functionality the Cisco IOS router can be used as an integrated gateways for both voice and video calls.

H.320 video uses multiple ISDN channels bound together to achieve the speeds needed to pass full-motion video. One of the problems with this bonding mechanism is that, when an inbound ISDN video call is received, the gateway does not know how many channels will be requested for that call until after it accepts the call and the source device indicates how many additional channels are required. If there are not enough B-Channels to satisfy the request, the call is disconnected. Therefore, careful traffic engineering is required to minimize the possibility that this situation will occur. Essentially, you want to ensure that there are always enough B-Channels available to handle the next call that might come in.

This B-Channel issue occurs in two cases:

- Inbound calls from the PSTN to the IP network
- Outbound calls from the IP network to the PSTN

Inbound Calls

For inbound calls, consider the following scenario:

> A company has a Cisco Unified Videoconferencing 3527 Gateway with an ISDN PRI circuit connecting it to a central office (CO) switch. The ISDN PRI circuit in this case offers 23 B-Channels. A video call is received from the PSTN at 384 kbps. This call takes six B-Channels, leaving 17 available. A second and third 384-kbps call are received on the line while the first one is still active. These each take an additional six channels, leaving five channels available. When the fourth 384-kbps call is received, the gateway will answer the call but, recognizing that it does not have enough B-Channels available (it only has five left but the call requires six), it will disconnect (by sending a Q.931 RELEASE COMPLETE with "16: Normal Call Clearing" as the reason). The caller attempting to make the fourth call will not know why the call failed and might redial the number repeatedly, trying to make the call work.

On Cisco Unified Videoconferencing gateways, you can minimize your chances of running into this issue by configuring the gateway to send a request to the CO to busy-out the remaining B-Channels (in this example, five channels) whenever the gateway reaches a certain threshold of utilization (configured as a percentage of total bandwidth).

In addition, you can have the CO provision multiple ISDN circuits in a trunk group. When the first circuit reaches the busy-out threshold, calls will roll over to the next PRI in the group. The Cisco Unified Videoconferencing 3500 Series Gateway offers two ISDN PRI connections and supports bonding channels across both ports. For example, port 1 might have only five channels available while port 2 is sitting idle and, therefore, has 23 channels available. By taking the five channels from port 1 and one channel from port 2 and bonding them together, the fourth 384-kbps call can succeed. This leaves 22 channels available on controller 2, and at some point additional inbound calls would reach the busy-out threshold again. At that point the remaining channels on port 2 will be busied out, and all further inbound calls will be rejected with cause code "Network Congestion." Cisco Unified Videoconferencing gateways cannot bound channels across different gateways or across different Cisco 3500 Series gateway models in the same Cisco 3545 chassis, so two ports is the maximum that you can bond together. The CO switch can still roll calls over to a third or forth PRI in the trunk group (most COs support trunk groups of up to 6 circuits), but you cannot bond channels between PRI number one and PRI number three, for example, as you can between PRI number one and PRI number two.

The busy-out logic described above depends on the assumption that all calls take place at the same speed. Suppose, for example, that two 384-kbps calls are active on a port and a 128-kbps call came in. This call would take only two channels, using a total of 14 channels for the three calls (6+6+2 = 14) and leaving nine channels available on the circuit. However, if the busy-out threshold is set at 18 channels (assuming that all calls would take place at 384-kbps), only four channels are still available under this busy-out threshold. If another 384 kbps call comes in at this point, the call will fail because the remaining four channels are not enough to support the call. Also, because the busy-out threshold of 18 channels has not been reached yet (only 14 channels are used), the circuit is not busied out and calls will not roll over to the next circuit. This condition will persist until one of the existing calls is disconnected. To avoid such situations, it is important to try to standardize on a single call speed for all calls.

Outbound Calls

Outbound calls encounter the same potential situations as inbound calls, but the way in which the busy-out occurs is different. The Cisco Unified Videoconferencing 3500 Series Gateways support messages called Resource Availability Indicator and Resource Availability Confirm (RAI/RAC). The RAI/RAC messages are defined under the H.225 RAS specification and are used by the gateways to tell the gatekeeper that they are full and to no longer route any more calls to them. When the gateway reaches the busy-out threshold, it sends an RAI message with a status of True to the gatekeeper. True means "Do not send me any more calls;" False means "I am available." The gateway sends an RAI=False as soon as it is no longer at its busy-out threshold. The busy-out threshold for outbound calls is separate from the busy-out threshold for inbound calls, and you can configure them differently so that inbound calls will roll over to the next available circuit but outbound calls will still be accepted, or vice versa. For example, you could configure the RAI threshold to 12 channels but the ISDN busy-out threshold to 18 channels. When two 384 kbps are active, outbound calls will roll over to the next available gateway, but a third 384-kbps inbound call could still be received. An equally efficient method of achieving outbound call busy-out failover is to use Unified CM's route group and route list construct, as described in the following section, instead of the RAI/RAC method.

Configuring the Gateways in Unified CM

You can configure a Unified Videoconferencing gateway in either of the following ways in Unified CM:

- Configure it as an H.323 gateway, and Unified CM will route calls directly to the gateway.
- Configure an H.225 gatekeeper-controlled trunk to the gatekeeper, and route calls to the gateway through the gatekeeper.

If you have only one gateway, it is probably easier to configure it directly in Unified CM instead of going through a trunk to get to it. If you have multiple gateways for load balancing and redundancy, you can either configure them all in Unified CM and place them into a route group(s) and route list, or configure an H.225 trunk to the gatekeeper and rely on RAI/RAC between the gateways and the gatekeeper to tell Unified CM which gateway it should send a given call to.

For inbound calls from the PSTN to Unified CM, the Cisco Unified Videoconferencing gateways can either register with a gatekeeper or be configured with the IP addresses of up to three Unified CM servers to which they should send all inbound call requests. This method is known as peer-to-peer mode. Either way, the goal is have all inbound calls received by the gateways sent to Unified CM so that Unified CM can decide how to route the calls. See Gatekeepers, page 12-19, for more details on how to configure the gatekeeper to route calls from the gateways to Unified CM.

Call Signaling Port Numbers

By default, the Cisco Unified Videoconferencing Gateways listen on TCP port 2720 instead of the well-known port 1720. However, also by default, Unified CM sends H.323 calls to port 1720. You can change the port that the gateway listens on or you can change the port that Unified CM sends to in the H.323 gateway device configuration in Unified CM. Either way, both sides have to match in order for outbound calls to the gateway to succeed.

In the inbound direction, when configured to operate in peer-to-peer mode, the Cisco Unified Videoconferencing Gateways will send the call to Unified CM on port 1720. When configured to register with a gatekeeper, Unified CM uses a randomly generated port number for all gatekeeper-controlled trunks. This method enables Unified CM to have multiple trunks to the same gatekeeper. This port number is included in the Registration Request (RRQ) from Unified CM to the gatekeeper, so the inbound H.225 setup message from the gateway to Unified CM will be sent to this port number. However, if the gateway is configured directly in Unified CM as an H.323 gateway device, Unified CM will ignore the fact that the call came in on the TCP port of the H.225 trunk and will instead match the source IP address to the H.323 gateway device configured in its database. If it does not find a matching device, Unified CM will treat the call as if it came in on the trunk.

In the outbound direction, if Unified CM uses a gatekeeper-controlled H.225 trunk to reach the gateway, the gatekeeper will tell Unified CM which TCP port to use to reach the gateway. If the gateway is configured in Unified CM as an H.323 gateway device (that is, peer-to-peer mode), then Unified CM must be configured to send calls either to port 2720 (default) or to 1720 (if the listening port on the gateway has been modified).

Call Signaling Timers

Due to the delay inherent in H.320 bonding, video calls can take longer to complete than voice calls. Several timers in Unified CM are tuned, by default, to make voice calls process as fast as possible, and they can cause video calls to fail. Therefore, you must modify the following timers from their default values in order to support H.320 gateway calls:

- H.245TCSTimeout
- Media Exchange Interface Capability Timer
- Media Exchange Timer

Cisco recommends that you increase each of these timers to 25 by modifying them under the Service Parameters in Unified CM Administration. Note that these are cluster-wide service parameters, so they will affect calls to all types of H.323 devices, including voice calls to existing H.323 Cisco Voice Gateways.

Bearer Capabilities of Voice Gateways

H.323 calls use the H.225/Q.931 Bearer Capabilities Information Element (bearer-caps) to indicate what type of call is being made. A voice-only call has its bearer-caps set to "speech" or "3.1 KHz Audio" while a video call has its bearer-caps set to "Unrestricted Digital Information." Some devices do not support Unrestricted Digital Information bearer-caps. Calls to these devices might fail if Unified CM attempts the call as a H.323 video call.

Unified CM decides which bearer-caps to set, based on the following factors:

- Whether the calling and/or called devices are video-capable
- Whether the region in Unified CM is configured to allow video for calls between those devices

Unified CM supports retrying the video call as audio, and this feature can be enabled through configuration. When Unified CM makes a video call with bearer-caps set to "Unrestricted Digital" and the call fails, Unified CM then retries the same call as an audio call with the bearer-caps set to "speech."

When using H.323, Cisco IOS gateways can service calls as voice or video, based on the bearer capabilities it receives in the call setup. When using SIP, the gateway translates the ISDN capabilities into SDP for call negotiations.

If the Cisco voice gateway uses MGCP to communicate with Unified CM, the problem will not occur because Unified CM does not support video on its MGCP protocol stack and because, in MGCP mode, Unified CM has complete control over the D-Channel signaling to the PSTN.

CHAPTER 14

Cisco Unified CM Trunks

Revised: August 31, 2012; OL-27282-05

A trunk is a communications channel on Cisco Unified Communications Manager (Unified CM) that enables Unified CM to connect to other servers. Using one or more trunks, Unified CM can receive or place voice, video, and encrypted calls, exchange real-time event information, and communicate in other ways with call control servers and other external servers.

Trunks are an integral and a crucial part of a Cisco Unified Communications deployment, hence it is important to understand the types of trunks available, their capabilities, and design and deployment considerations such as resiliency, capacity, load balancing, and so forth.

There are two basic types of trunks that can be configured in Unified CM:

- SIP and H.323 trunks, both of which can be used for external communications
- Intercluster trunks (ICTs)

This chapter describes the general capabilities and functions of these trunks. Additional discussion on specific applications of Unified CM trunks can be found in other pertinent chapters of this document.

This chapter discusses the following topics:

- A Comparison of SIP and H.323 Trunks, page 14-3
- SIP Trunks Overview, page 14-6
- H.323 Trunks Overview, page 14-38
- General SIP and H.323 Trunk Design Considerations, page 14-57
- IP PSTN and IP Trunks to Service Provider Networks, page 14-63
- Trunk Aggregation Platforms, page 14-64

For more details on the applications of Unified CM trunks, refer to their respective sections in the following chapters:

- Unified Communications Deployment Models, page 5-1
- Media Resources, page 17-1
- Call Admission Control, page 11-1
- IP Video Telephony, page 12-1
- Cisco IM and Presence, page 23-1

What's New in This Chapter

Table 14-1 lists the topics that are new in this chapter or that have changed significantly from previous releases of this document.

Table 14-1 New or Changed Information Since the Previous Release of This Document

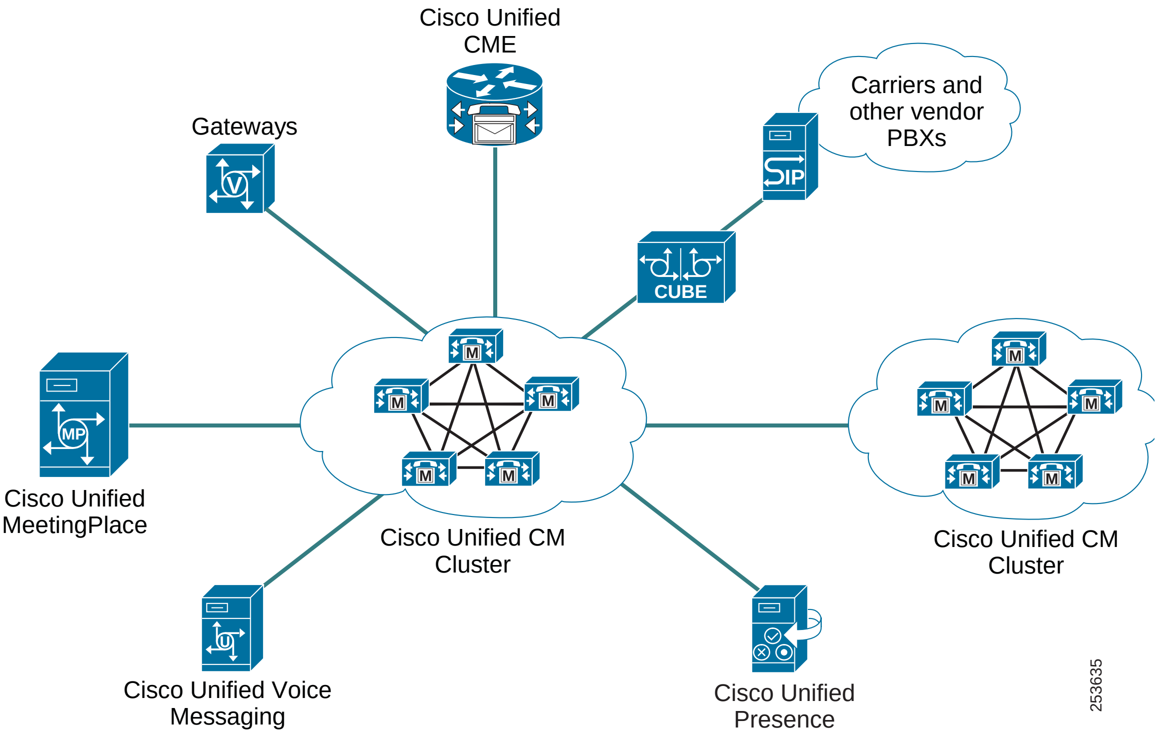
New or Revised Topic	Described in:	Revision Date
Clustering over the WAN with Unified CM Session Management Edition Cluster Trunks	Design Guidance for Clustering over the WAN with Unified CM Session Management Edition Cluster Trunks, page 14-37	August 31, 2012
Audio codec preference lists	Audio Codec Preference (Accept Audio Codec Preference in Received Offer), page 14-11 Accept Audio Codec Preferences in Received Offer, page 14-60	June 28, 2012
Normalization and transparency scripts	Normalization and Transparency Scripts in Unified CM, page 14-15	June 28, 2012
Other minor updates for Cisco Unified Communications System Release 9.0	Various sections throughout this chapter	June 28, 2012

Unified CM Trunks Solution Architecture

Unified CM uses the mechanism of IP trunks to exchange call-related information with other components of a Unified Communications solution. Given their importance in this respect, it is important to develop the system architecture of the IP trunks with proper regard to the protocol, feature and service expectations, performance requirements, and so forth.

Figure 14-1 illustrates the role of IP trunks in system connectivity. The illustration does not show all possible connections from the Unified CM cluster.

Figure 14-1 IP Trunks Provide Connections to Unified CM



Calls are directed toward trunks as defined by the dial plan using the route pattern construct. A route pattern can use a trunk either directly or through a route list. The route list, if used, consists of one or more route groups, each of which contains one or more trunks. An individual trunk within a route group may be configured to be selected in either a top-down or circular fashion. For outgoing calls, Unified CM selects one of the trunks associated in this fashion with the route pattern. Before it accepts an incoming call, Unified CM verifies whether a trunk is defined to the remote address from which the call is received.

A Comparison of SIP and H.323 Trunks

Cisco Unified CM trunk connections support both SIP and H.323. In many cases, the decision to use SIP or H.323 is driven by the unique feature(s) offered by each protocol. There are also a number of external factors that can affect the choice of trunk protocol, such as customer preference or the protocol's maturity and degree of interoperability offered between various vendors' products.

For trunk connections between Cisco devices, this decision is relatively straightforward. For trunk connections to other vendors' products and to service provider networks, it is important to understand which features are required by the customer and the extent of interoperability between any two vendors' products.

Table 14-2 compares some of the features offered over SIP and H.323 trunks between Unified CM clusters.

Table 14-2 Comparison of SIP and H.323 Features on Cisco Unified CM Trunks

Feature	SIP	QSIG over SIP	H.323	QSIG over H.323
Calling Line (Number) Identification Presentation	Yes	Yes	Yes	Yes
Calling Line (Number) Identification Restriction	Yes	Yes	Yes	Yes
Calling Name Identification Presentation	Yes	Yes	Yes	Yes
Calling Name Identification Restriction	Yes	Yes	Yes	Yes
Connected Line (Number) Identification Presentation	Yes	Yes	Yes	Yes
Connected Line (Number) Identification Restriction	Yes	Yes	Yes	Yes
Connected Name Identification Presentation	Yes	Yes	Yes	Yes
Connected Name Identification Restriction	Yes	Yes	Yes	Yes
Alerting Name	Yes	Yes	No	Yes
Call Transfer (Blind/Attended)	Yes/Yes	Yes/Yes	Yes/Yes	Yes/Yes
Call Forward All	Yes	Yes	Yes	Yes
Call Forward Busy	Yes	Yes	Yes	Yes
Call Forward No Reply	Yes	Yes	Yes	Yes
Call Completion to Busy Subscriber	No	Yes	No	Yes
Call Completion No Reply	No	Yes	No	Yes
Subscribe/Notify, Publish – Presence	Yes	Yes	No	No
Message Waiting Indication (MWI: lamp ON, lamp OFF)	Yes	Yes	No	Yes
Path Replacement	No	Yes	No	Yes
Call Hold/Resume	Yes	Yes	Yes	Yes
Music On Hold (unicast and multicast)	Yes	Yes	Yes	Yes
DTMF-relay	RFC 2833, KPML (OOB), Unsolicited Notify (OOB)	RFC 2833, KPML (OOB), Unsolicited Notify (OOB)	H.245 Out Of Band (OOB)[1]	H.245 Out Of Band (OOB)[1]
SIP Early Offer	Yes – MTP may be required	Yes – MTP may be required	N/A	N/A
SIP Delayed Offer	Yes	Yes	N/A	N/A
H.323 Fast Start	N/A	N/A	Yes – MTP always required for Outbound Fast Start	Yes – MTP always required for Outbound Fast Start
H.323 Slow Start	N/A	N/A	Yes	Yes
Audio codecs	G.711, G.722, G.723, G.729, iLBC, AAC, iSAC	G.711, G.722, G.723, G.729, iLBC, AAC, iSAC	G.711, G.722, G.723, G.729	G.711, G.722, G.723, G.729

Table 14-2 Comparison of SIP and H.323 Features on Cisco Unified CM Trunks (continued)

Feature	SIP	QSIG over SIP	H.323	QSIG over H.323
Accept Audio Codec Preference in Received Offer	Yes	Yes	No	No
Codecs with MTP	All codecs supported when **Early Offer support for voice and video calls (insert MTP if needed)** is checked G.711, G.729 when **MTP Required** is checked	All codecs supported when **Early Offer support for voice and video calls (insert MTP if needed)** is checked G.711, G.729 when **MTP Required** is checked	G.711, G.723, G.729	G.711, G.723, G.729
Video	Yes	Yes	Yes	Yes
Video codecs	H.261, H.263, H.263+, H.264 AVC	H.261, H.263, H.263+, H.264 AVC	H.261, H.263, H.263+, H.264 AVC	H.261, H.263, H.263+, H.264 AVC
T.38 Fax	Yes	Yes	Yes	Yes
Signaling Authentication	Digest, TLS	Digest, TLS	No	No
Signaling Encryption	TLS	TLS	No	No
Media Encryption (audio)	SRTP	SRTP	SRTP	SRTP
RSVP-based QoS and call admission control	Yes	Yes	No	No
Support for + character	Yes	Yes	No	No
Inbound Calls — Called Party: Significant Digits, Prefix-Digits	Yes	Yes	Yes	Yes
Incoming Calling Party Settings: Strip Digits, Prefix-Digits based on Number Type	SIP does not support Number Type - "Unknown" used for all calls	SIP does not support Number Type - "Unknown" used for all calls	Unified CM, Unknown, National, International, Subscriber	Unified CM, Unknown, National, International, Subscriber
Incoming Called Party Settings: Strip Digits, Prefix-Digits based on Number Type	N/A	N/A	Unified CM, Unknown, National, International, Subscriber	Unified CM, Unknown, National, International, Subscriber
Connected Party Transformation	Yes	Yes	No	No
Outbound Calling Party Transformations	Yes	Yes	Yes	Yes
Outbound Called Party Transformations	Yes	Yes	Yes	Yes

Table 14-2 Comparison of SIP and H.323 Features on Cisco Unified CM Trunks (continued)

Feature	SIP	QSIG over SIP	H.323	QSIG over H.323
Outbound Calling/Called Party Number Type Setting	SIP does not support Number Type	SIP does not support Number Type	Unified CM, Unknown, National, International, Subscriber	Unified CM, Unknown, National, International, Subscriber
Outbound Called/Called Party Numbering Plan Setting	SIP does not support Number Plan	SIP does not support Number Plan	Unified CM, ISDN, National Standard, Private, Unknown	Unified CM, ISDN, National Standard, Private, Unknown
Trunk destination — State detection mechanism	OPTIONS Ping	OPTIONS Ping	Per call attempt	Per call attempt

[1] H.323 trunks support signaling of RFC 2833 for certain connection types.

SIP Trunks Overview

SIP trunks provide connectivity to other SIP devices such as gateways, Cisco Unified CM Session Management Edition, SIP proxies, Unified Communications applications, and other Unified CM clusters. Today, SIP is arguably the most commonly chosen protocol when connecting to service providers and Unified Communications applications. Cisco Unified CM provides the following SIP trunk and call routing enhancements:

- Can run on all Unified CM nodes
- Up to 16 destination IP addresses per trunk
- SIP OPTIONS ping keepalives
- SIP Early Offer support for voice and video calls (insert MTP if needed)
- Audio codec preference (Accept Audio Codec Preference in Received Offer)
- QSIG over SIP
- SIP trunk normalization and transparency
- SIP REFER transparency
- Supports the use of route lists on all Unified CM nodes

The SIP trunk features available in the current release of Unified CM make SIP the preferred choice for new and existing trunk connections. The QSIG over SIP feature provides parity with H.323 intercluster trunks and can also be used to provide QSIG over SIP trunk connections to Cisco IOS gateways (and on to QSIG-based TDM PBXs). The ability to run on all Unified CM nodes and to handle up to 16 destination IP addresses improves outbound call distribution from Unified CM clusters and reduces the number of SIP trunks required between clusters and devices. SIP OPTIONS ping provides dynamic reachability detection for SIP trunk destinations, rather than per-call reachability determination. SIP Early Offer support for voice and video calls (insert MTP if needed) can reduce or eliminate the need to use MTPs and allows voice, video, and encrypted calls to be made over SIP Early Offer trunks.

SIP trunk normalization and transparency improve native Unified CM interoperability with and between third-party unified communications systems. Normalization allows inbound and outbound SIP messages and SDP information to be modified on a per-SIP-trunk basis. Transparency allows Unified CM to pass SIP headers, parameters, and content bodies from one SIP trunk call leg to another, even if Unified CM does not understand or support the parts of the message that are being passed through.

These features are discussed in detail later in this section.

For the complete list of new enhancements for SIP trunks, refer to the Cisco Unified Communications Manager product release notes available at

http://www.cisco.com/en/US/products/sw/voicesw/ps556/prod_release_notes_list.html

General Deployment Considerations

Unified CM SIP trunks offer a greater set of features in comparison with H.323 intercluster trunks, thus making SIP the protocol of choice for intercluster trunk connections (although H.323 Annex M1 may still be preferred for intercluster trunk connections to Unified CM clusters using earlier software versions). Also, given the wide support of SIP in the industry, SIP trunks are usually a good choice for connectivity to third-party applications and service providers.

SIP Trunk Features and Operation

This section explains how Unified CM SIP trunks operate and describes several key SIP trunk features that should be taken into account when designing and deploying Unified CM SIP trunks.

SIP Trunks Can Run on All Active Unified CM Nodes

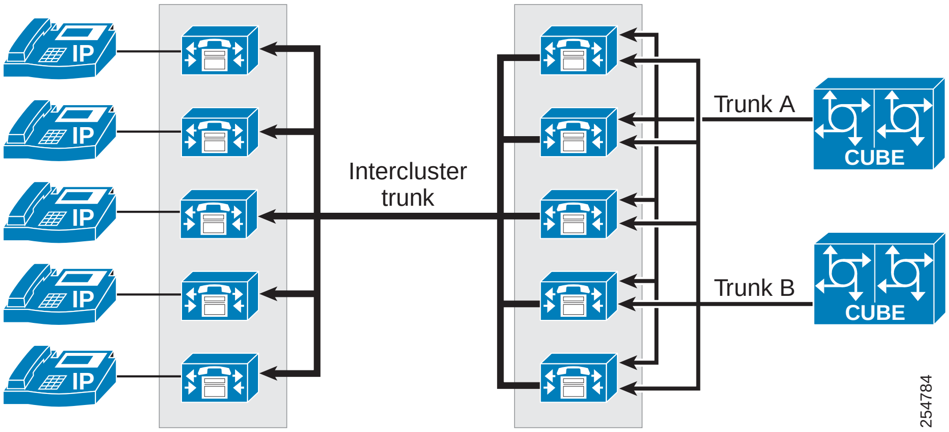
When the **Run on all Active Unified CM Nodes** option is checked on a SIP trunk, Unified CM creates an instance of the SIP trunk daemon on every call processing subscriber within the cluster, thus allowing SIP trunk calls to be made or received on any call processing subscriber. (Prior to this feature, up to three nodes could be selected per trunk by using Unified CM Groups.) With **Run on all Active Unified CM Nodes** enabled, outbound SIP trunk calls originate from the same node on which the inbound call (for example, from a phone or trunk) is received. As with all Unified CM SIP trunks, the SIP daemons associated with the trunk will accept inbound calls only from end systems with IP addresses that are defined in the trunk's destination address fields. Running SIP trunks on all nodes is recommended where the SIP trunk is required to process a large number of calls so that outbound and inbound call distribution can be evenly spread across all call processing subscribers within a cluster. Also, when multiple SIP trunks to the same destination(s) are using the same subscriber, a unique incoming and destination port number must be defined per trunk to allow each trunk to be identified uniquely.

Up to 16 SIP Trunk Destination IP Addresses

SIP trunks can be configured with up to 16 destination IP addresses, 16 fully qualified domain names, or a single DNS SRV entry. Support for additional destination IP addresses reduces the need to create multiple trunks associated with route lists and route groups for call distribution between two Unified Communications systems, thus simplifying Unified CM trunk design. (See Figure 14-2.) This feature can be used in conjunction with the **Run on all Active Unified CM Nodes** feature or with a SIP trunk that uses standard Unified CM Groups to create a SIP daemon on up to three nodes within the cluster.

Bear in mind, however, that the SIP daemons associated with a Unified CM SIP trunk will accept inbound calls only from end systems with IP addresses that are defined in the trunk's destination address fields.

Figure 14-2 SIP Trunks with Multiple Destination IP Addresses Running on All Active Nodes



SIP OPTIONS Ping

The SIP OPTIONS Ping feature can be enabled on the SIP Profile associated with a SIP trunk to dynamically track the state of the trunk's destination(s). When this feature is enabled, each node running the trunk's SIP daemon will periodically send an OPTIONS Request to each of the trunk's destination IP addresses to determine its reachability and will send calls only to reachable nodes. A destination address is considered to be "out of service" if it fails to respond to an OPTIONS Request, if it sends a Service Unavailable (503) response or Request Timeout (408) response, or if a TCP connection cannot be established. The trunk state is considered to be "in service" when at least one node receives a response (other than a 408 or 503) from a least one destination address. SIP trunk nodes can send OPTIONS Requests to the trunk's configured destination IP addresses or to the resolved IP addresses of the trunk's DNS SRV entry. Enabling SIP OPTIONS Ping is recommended for all SIP trunks because it allows Unified CM to dynamically track trunk state rather than determining trunk state on a per-call and timeout basis.

SIP Early Offer Support over Unified CM SIP Trunks

SIP negotiates media exchange by means of the Session Description Protocol (SDP), where one side offers a set of capabilities to which the other side answers, thus converging on a set of media characteristics. SIP allows the initial offer to be sent either by the caller in the initial INVITE message (Early Offer) or, if the caller chooses not to, the called party can send the initial offer in the first reliable response (Delayed Offer).

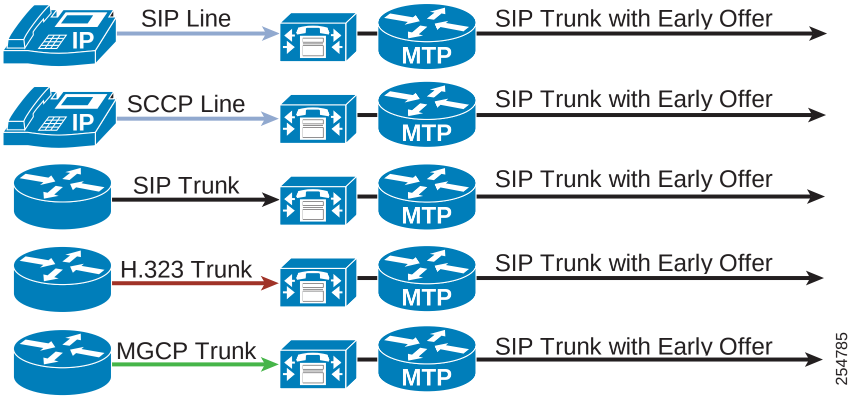
By default, Unified CM SIP trunks send the INVITE without an initial offer (Delayed Offer). In general SIP Delayed Offer is preferred for Unified CM SIP trunks because MTPs are not needed to establish a Delayed Offer call for voice, video, or encrypted media. If SIP Early Offer is desired, Unified CM has two configurable options to enable a SIP trunk to send the offer in the INVITE:

- Media Termination Point Required, page 14-9
- Early Offer Support for Voice and Video Calls (Insert MTP If Needed), page 14-9

Media Termination Point Required

Enabling the **Media Termination Point Required** option on the SIP trunk assigns an MTP from the trunk's media resources group (MRG) to every outbound call. (See Figure 14-3.) This statically assigned MTP supports only the G.711 or G.729 codecs, thus limiting media to voice calls only.

Figure 14-3 SIP Early Offer with Media Termination Point Required

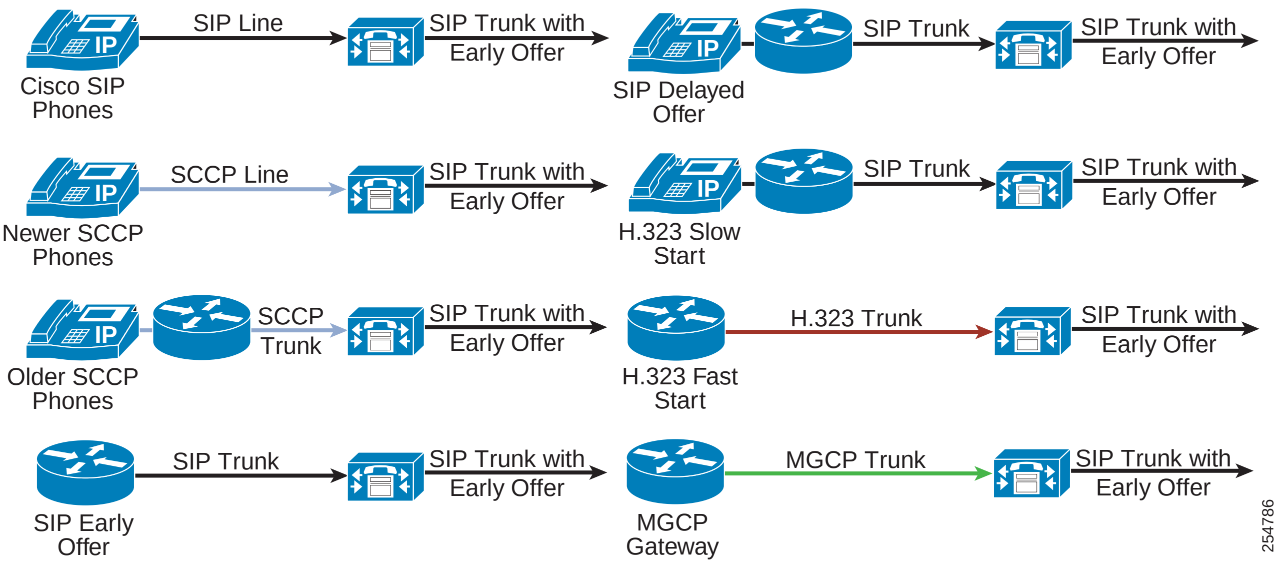

IP
SIP Line
MTP
SIP Trunk with Early Offer
IP
SCCP Line
MTP
SIP Trunk with Early Offer
SIP Trunk
MTP
SIP Trunk with Early Offer
H.323 Trunk
MTP
SIP Trunk with Early Offer
MGCP Trunk
MTP
SIP Trunk with Early Offer
254785

Early Offer Support for Voice and Video Calls (Insert MTP If Needed)

Enabling **Early Offer support for voice and video calls (insert MTP if needed)** on the SIP Profile associated with the SIP trunk inserts an MTP only if the calling device cannot provide Unified CM with the media characteristics required to create the Early Offer. In general, **Early Offer support for voice and video calls (insert MTP if needed)** is recommended over **Media Termination Point Required** because this configuration option reduces MTP usage (see Figure 14-4). Calls from older SCCP-based phones registered to Unified CM over SIP Early Offer trunks configured with this option will use an MTP to create the Offer SDP, and these calls support voice, video, and encrypted media. Inbound calls to Unified CM from SIP Delayed Offer trunks or H.323 Slow Start trunks that are extended over an outbound SIP Early Offer trunk will use an MTP to create the Offer SDP; however, these calls support audio only in the initial call set up but can be escalated mid-call to support video and SRTP if the call media is renegotiated (for example, after hold/resume). For guidance on when to use **Early Offer support for voice and video calls (insert MTP if needed)**, see Design Considerations for SIP Trunks, page 14-28.

Note MTP resources are not required for incoming INVITE messages, whether or not they contain an initial offer SDP.

Figure 14-4 Early Offer Support for Voice and Video Calls



Unified CM does not need to insert an MTP to create an outbound Early Offer call over a SIP trunk if the inbound call to Unified CM is received by any of the following means:

- On a SIP trunk using Early Offer
- On an H.323 trunk using Fast Start
- On an MGCP trunk
- From a SIP-based IP phone registered to Unified CM
- From newer SCCP-based Cisco Unified IP Phone models registered to Unified CM

For the above devices, Unified CM uses the media capabilities of the endpoint and applies the codec filtering rules based on the region-pair of the calling device and outgoing SIP trunk to create the offer SDP for the outbound SIP trunk call. In most cases, the offer SDP will have the IP address and port number of the endpoint initiating the call. This is assuming that Unified CM does not have to insert an MTP for other reasons such as a DTMF mismatch, TRP requirements, or a transcoder requirement when there is no common codec between the regions of the calling device and the SIP trunk.

When **Early Offer support for voice and video calls (insert MTP if needed)** is configured on a trunk's SIP Profile, calls from older SCCP-based phones, SIP Delayed Offer trunks, and H.323 Slow Start trunks will cause Unified CM to allocate an MTP if an MTP or transcoder has not already been allocated for that call for another reason. The MTP is used to generate an offer SDP with a valid media port and IP address. The MTP will be allocated from the media resources associated with the calling device rather than from the outbound SIP trunk's media resources. (This prevents the media path from being anchored to the outbound SIP trunk's MTP). If the MTP cannot be allocated from the calling device's media resource group list (MRGL), then the MTP allocation is attempted from the SIP trunk's MRGL.

For calls from older SCCP phones registered to Unified CM, some of the media capabilities of the calling device (for example, supported voice codecs, video codecs, and encryption keys if supported) are available for media exchange through the Session Description Protocol (SDP). Unified CM will create a superset of the endpoint and MTP codec capabilities and apply the codec filtering based on the applicable region-pair settings. The outbound Offer SDP will use the MTP's IP address and port number and can support voice, video, and encrypted media. Note that the MTP should be configured to support the pass-through codec.

Note Older SCCP-based IP phones such as the Cisco Unified IP Phone 7902, 7905, 7910, 7912, 7920, 7935, 7940, and 7960 require the use of an MTP when they make calls over a SIP trunk with the **Early Offer for voice and video (insert MTP if needed)** feature enabled. If you have a significant number of these phone types deployed in a cluster, provision MTP resources in the cluster equivalent to the number of busy hour calls over those SIP trunks using the **Early Offer for voice and video (insert MTP if needed)** feature.

When Unified CM receives an inbound call on an H.323 Slow Start or SIP Delayed Offer trunk, the media capabilities of the calling device are not available when the call is initiated. In this case, Unified CM must insert an MTP and will use its IP address and UDP port number to advertise all supported audio codecs (after region pair filtering) in the Offer SDP of the initial INVITE sent over the outbound SIP trunk. When the Answer SDP is received on the SIP trunk, if it contains a codec that is supported by the calling endpoint, then no additional offer-answer transaction is needed. In case of codec mismatch, Unified CM can either insert a transcoder to address the mismatch or send a reINVITE or UPDATE to trigger media negotiation. Calls from H.323 Slow Start or SIP Delayed Offer trunks support audio only in the initial call setup, but they can be escalated mid-call to support video and SRTP if the call media is renegotiated (for example, after Hold/Resume).

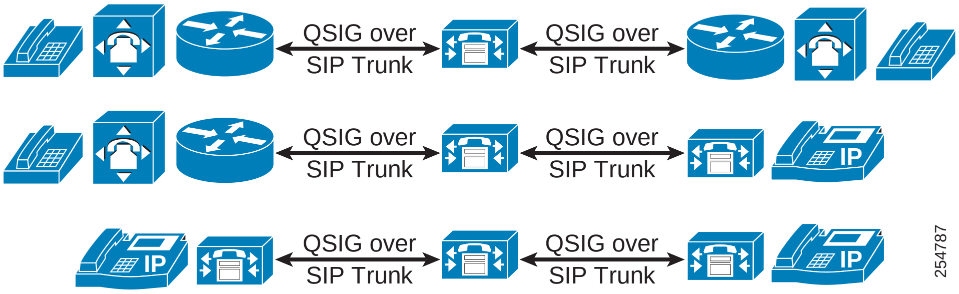
Audio Codec Preference (Accept Audio Codec Preference in Received Offer)

Cisco Unified CM 9.*x* provides configurable audio codec preference lists, which can be used to prioritize codec preferences for calls within regions, between regions, and between clusters. For calls over SIP trunks, the configurable SIP Profile option **Accept Audio Codec Preference in Received Offer** allows the trunk to override the codec preferences configured for the trunk's region or region pair and to use the codec preferences received in an Offer from an off-cluster device. This feature is particularly useful in deployments where a SIP call passes through two or more Unified CM clusters (for example, in SME deployments where the end codec preference needs to be preserved for a call). Codec preference is discussed in detail in the section on Codec Selection Over IP Trunks, page 14-58.

QSIG over SIP Trunks

Unified CM can encapsulate QSIG content in SIP messages, thus allowing features such as Call Back, MWI, and Path Replacement to be invoked over SIP QSIG intercluster trunks and over SIP QSIG trunks to Cisco IOS gateways. (See Figure 14-5.) QSIG over SIP trunks provides parity with the QSIG feature set on H.323 Annex M1 intercluster trunks and MGCP QSIG trunks. (ISO and ECMA variants of QSIG are supported on a per-trunk basis.)

Figure 14-5 QSIG over SIP Trunks



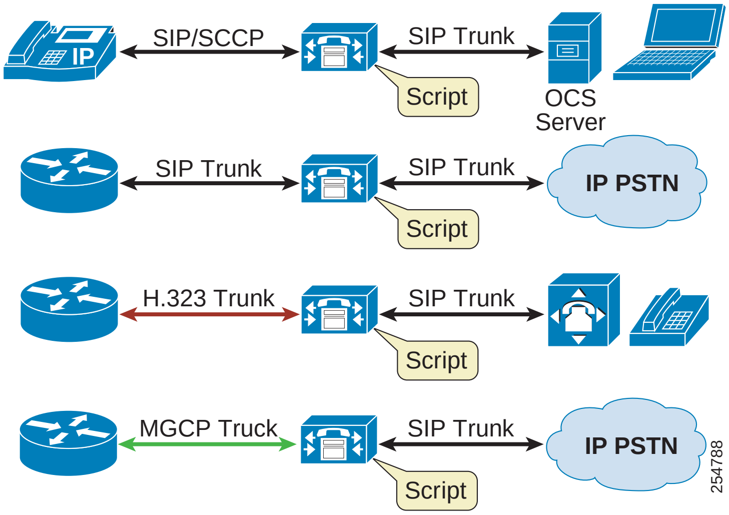
SIP Trunk Message Normalization and Transparency

Normalization and transparency provide powerful script-based functionality for SIP trunks that can be used to transparently forward and/or modify SIP messages and message body contents as they traverse Unified CM. Normalization and transparency scripts are designed to address SIP interoperability issues, allowing Unified CM to interoperate with SIP-based third-party PBXs, applications, and IP PSTN services.

SIP Trunk Normalization

Normalization allows incoming and outgoing SIP messages to be modified on their way through Unified CM. Normalization applies to all calls that traverse a SIP trunk with an associated script, regardless of what protocol is being used for the other endpoint involved in the call. For example, a SIP trunk normalization script can operate on a call from a SIP line device to a SIP trunk, from an SCCP-based device to a SIP trunk, from MGCP to SIP trunk, from H.323 to SIP trunk, and so forth. (See Figure 14-6.) Normalization does not require end-to-end SIP.

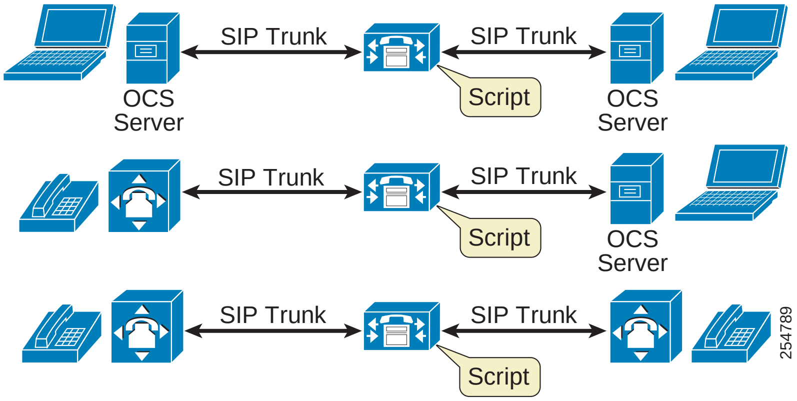
Figure 14-6 SIP Trunk Normalization



SIP Trunk Transparency

Transparency allows Unified CM to pass SIP headers, parameters, and content bodies from one SIP trunk call leg to another, even if Unified CM does not understand or support the parts of the message that are being passed through. Transparency (or transparent pass-through) is applicable only when the call through Unified CM is from SIP trunk to SIP trunk, as illustrated in Figure 14-7.

Figure 14-7 SIP Trunk Transparency



Normalization and transparency scripts use Lua, a powerful, fast, lightweight, embeddable scripting language to modify SIP messages and SDP body content on SIP trunks. (For more information on Lua, refer to the documentation available at http://lua-users.org/wiki/LuaOrgGuide.)

Cisco has created a library of Lua-based SIP Message APIs that allow specified information in the SIP message and SDP body to be retrieved, modified, replaced, removed, passed through, ignored, appended to, transformed, and so on. The underlying Lua language allows retrieved information to be stored as variables and operated on using a series of operations such as: If, elseif, while, do, <, >, =, and so forth. The scripting approach naturally supports multiple variables and state-specific contexts for making script decisions. The combination of Cisco's SIP Message Library APIs and the functionality underlying the Lua language creates a very powerful scripting environment that allows almost any SIP message and/or its SDP body content to be modified.

For inbound messages on a SIP trunk, normalization and transparency script processing occurs immediately after receiving the message from the network. For outbound messages, script processing occurs immediately before sending the message to the network.

Within a Lua script, callback functions (also known as message handlers) are used to request message types of interest. The Cisco Lua environment constructs the name of the message handler based on the message direction and method for requests (for example, inbound_INVITE) and based on the message direction, response code, and method (from the CSeq header) for responses (for example, outbound_180_INVITE). A message object (for example, msg) is passed to the message handler, thereby allowing the script to modify the message (for example, inbound_INVITE(msg)).

Callback Function (message Handler) examples:

inbound_INVITE()	outbound_INVITE()
inbound_UPDATE()	outbound_SUBSCRIBE()
inbound_3XX_INVITE()	outbound_180_INVITE()

The Lua script then uses APIs defined in the Cisco SIP Message library to access and manipulate message parameters. For example:

- **getHeader**(*header-name*) returns header-value or ""
- **getHeaderValues**(*header-name*) returns a table of header values
- **addHeaderValueParameter**(*header-name*, *parameter-name*, [*parameter-value*])
- **getUri**(*header-name*) retrieves the URI from the specified header

- **block()** blocks the specified SIP message
- **applyNumberMask**(*header-name, mask*) retrieves the specified header and applies the specified number mask to the URI
- **getSdp()** returns the SDP content
- **sdp:getLine(start of line, line contains**_)_ returns line in SDP that starts with "start of line" and also has string "line contains"
- **sdp:modifyLine(start of line, line contains,** *new-line***)** finds the in SDP that starts with "start of line", the line matching "line contains" is replaced with the *new-line* parameter

The following examples illustrate the use of SIP Message API scripts.

Example 14-1 SIP Message API — getRequestLine

getRequestLine() returns the method, request-uri, and version.

This method returns three values:

- The method name
- The request-uri
- The protocol version

Example script:

Line 1	M = {}
Line 2	function M.outbound_INVITE(message)
Line 3	local method, ruri, ver = message:getRequestLine()
Line 4	end
Line 5	return M

Line 1 initializes the set of callback functions to an empty value. This set of callback functions, named M, is essentially a Lua table.

Lines 2 to 4 define a message handler. This callback function is executed when an outbound INVITE is sent from Unified CM. The script then gets the method, request-uri, and version from the request line and stores these values.

The script can define multiple message handlers. The name of the message handler dictates which message handler is invoked (if any) for a given SIP message.

The last line returns the set of callbacks. This line is absolutely required.

Message:

```
INVITE sip:1234@10.10.10.1 SIP/2.0
```

Output and result:

```
method == "INVITE"
ruri == "sip:1234@10.10.10.1"
version == "SIP/2.0"
```

Example 14-2 A script that simply removes the "Cisco-Guid" header in an outbound INVITE

Line 1	M = {}
Line 2	function M.outbound_INVITE(message)
Line 3	message:removeHeader("Cisco-Guid")
Line 4	end
Line 5	return M

Line 1 initializes the set of callback functions to an empty value. This set of callback functions, named M, is essentially a Lua Table.

Lines 2 to 4 define a message handler. This callback function is executed when an outbound INVITE is sent from Unified CM. The script can define multiple message handlers. The name of the message handler dictates which message handler is invoked (if any) for a given SIP message.

The last line returns the set of callbacks. This line is absolutely required.

Message:

```
INVITE sip:1234@10.10.10.1 SIP/2.0
.
P-Asserted-Identity: "1234" <1234@10.10.10.1>
Cisco-Guid: 1234-4567-1234
Session-Expires: 1800
```

Output and results:

```
INVITE sip:1234@10.10.10.1 SIP/2.0
.
P-Asserted-Identity: "1234"
```

For more information on SIP trunk normalization and transparency scripts, refer to the *Developer Guide for SIP Transparency and Normalization*, available at

http://www.cisco.com/en/US/docs/voice_ip_comm/cucm/sip_tn/8_5_1/sip_t_n.html

Normalization and Transparency Scripts in Unified CM

A number of normalization and transparency scripts are pre-loaded into Unified CM, and the following scripts are a representative sample of them:

- Refer-passthrough script — This script allows Unified CM to be removed from the call signaling path when a blind transfer (using an in-dialog REFER) is invoked between two SIP trunks.
- ContactHeader script — This script removes the audio and video attributes from the contact header in an inbound Delayed Offer mid-call re-invite.
- HCS-PCV-PAI-passthrough script — This script is used for integration with IMS-based networks and passes through/ adds the P-Charging-Vector header in INVITE, UPDATE & 200 OK messages.
- Diversion-Counter script — This script provides the capability to adjust the diversion counter for various Call Forward scenarios.
- VCS-interop script — This script provides interoperability for endpoints registered to the Video Communication Server (VCS).

Route Lists Run on All Active Unified CM Nodes

Although this is not specifically a SIP trunk feature, running route lists on all nodes provides benefits for trunks in route lists and route groups. Running route lists on all nodes improves outbound call distribution by using the "route local" rule to avoid unnecessary intra-cluster traffic.

For route lists, the route local rule operates as follows:

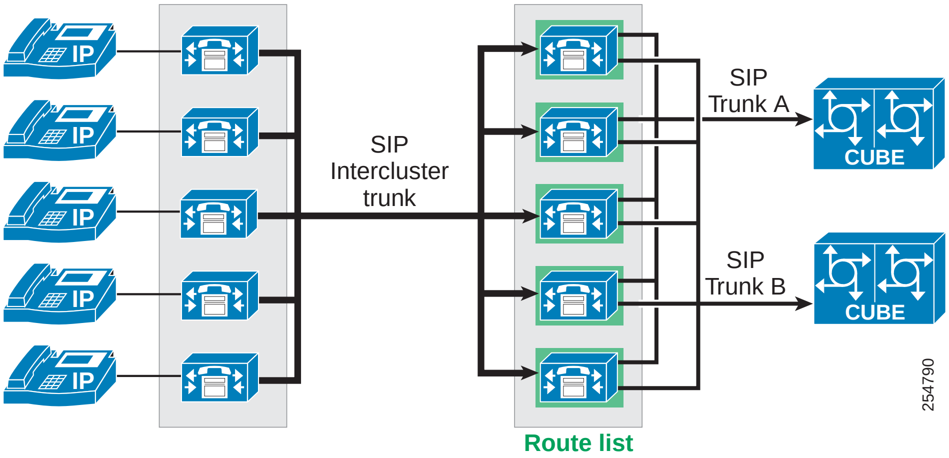
> For outbound calls that use route lists (and associated route groups and trunks), when a call from a registered phone or inbound trunk arrives at the node with the route list instance, Unified CM checks to see if an instance of the selected outbound trunk exists on the same node as the route list. If so, Unified CM will use this node to establish the outbound trunk call.

If both the route list and the trunk have **Run on all Active Unified CM Nodes** enabled, outbound call distribution will be determined by the node on which the inbound call arrives. When the selected outbound trunk uses Unified CM Groups instead of running on all nodes, Unified CM will apply the route local rule if an instance of the selected outbound trunk exists on the same node on which the inbound call arrived. If an instance of the trunk does not exist on this node, then Unified CM will forward the call (within the cluster) to a node where the trunk is active.

If the route list does not have **Run on all Active Unified CM Nodes** enabled, an instance of the route list will be active on one node within the cluster (the primary node of the route list's Unified CM Group) and the route local rule will be applied on this node. With this configuration, Cisco recommends that you do not use the primary node of the route list's Unified CM Group in the Unified CM Group of any of the trunks associated with the route list because this can result in sub-optimal outbound call distribution.

As a general recommendation, **Run on all Active Unified CM Nodes** should be enabled for all route lists. (See Figure 14-8.)

Figure 14-8 Route Lists Running on All Active Unified CM Nodes



SIP Trunks Using DNS

Using a DNS SRV entry as the destination of a SIP trunk might be preferable to defining multiple destination IP addresses in certain situations such as the following:

- SRV host prioritization is required
- SRV host weighting is required
- More than 16 destination IP addresses are required
- DNS SRV resolution is a requirement of the destination Unified Communications system

Note If the configuration option **Destination Address is an SRV** is selected, only a single SRV entry can be added as the trunk destination. (For example, Destination Address = cluster1.cisco.com. Port = 0.)

Figure 14-9 shows the call flow for a SIP trunk using DNS SRV to resolve the addresses to a destination Unified CM cluster. However, this destination could also be a third-party unified communications system.

Figure 14-9 Call Flow for Intercluster SIP Trunk Using DNS SRV

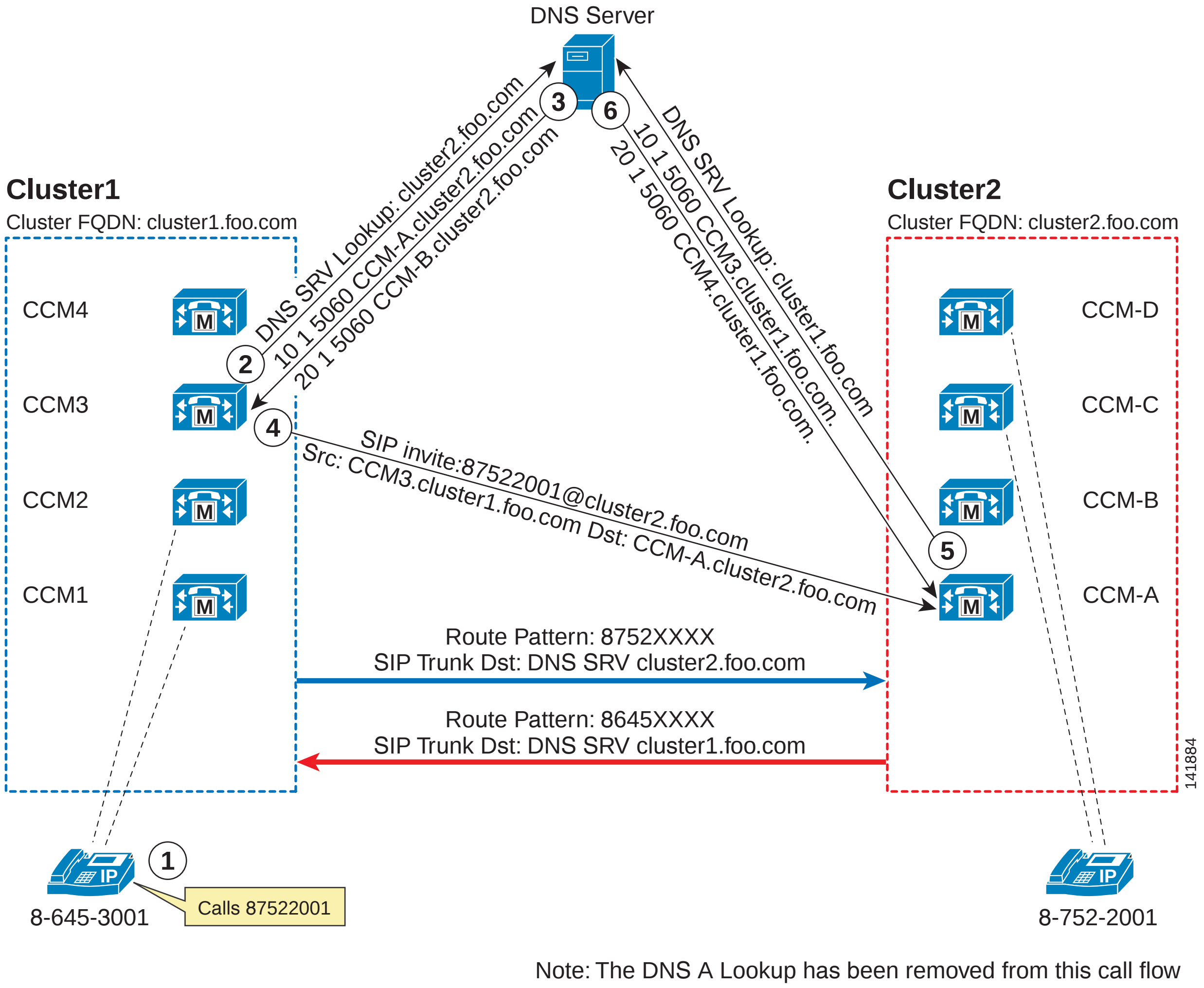

DNS Server
Cluster1
Cluster FQDN: cluster1.foo.com
CCM4
CCM3
CCM2
CCM1
Cluster2
Cluster FQDN: cluster2.foo.com
CCM-D
CCM-C
CCM-B
CCM-A
2
DNS SRV Lookup: cluster2.foo.com
3
10 1 5060 CCM-A.cluster2.foo.com
20 1 5060 CCM-B.cluster2.foo.com
6
DNS SRV Lookup: cluster1.foo.com
10 1 5060 CCM3.cluster1.foo.com.
20 1 5060 CCM4.cluster1.foo.com.
5
4
SIP invite:87522001@cluster2.foo.com
Src: CCM3.cluster1.foo.com Dst: CCM-A.cluster2.foo.com
Route Pattern: 8752XXXX
SIP Trunk Dst: DNS SRV cluster2.foo.com
Route Pattern: 8645XXXX
SIP Trunk Dst: DNS SRV cluster1.foo.com
1
Calls 87522001
8-645-3001
8-752-2001
141884
Note: The DNS A Lookup has been removed from this call flow

Figure 14-9 illustrates the following steps in the call flow:

1. The IP phone in Cluster1 calls 87522001.
2. The call matches a route pattern of 8752XXXX that is pointing to the SIP trunk with DNS SRV of cluster2.foo.com. CCM3 in Cluster1 is the node handling this call because the SIP trunk is registered to it. CCM3 sends a DNS SRV lookup for cluster2.foo.com
3. The DNS server replies with two records: CCM-A.cluster2.foo.com and CCM-B.cluster2.foo.com. Because CCM-A.cluster2.foo.com has a higher priority, the call is attempted to this Unified CM. Before sending the SIP Invite, another DNS lookup is done for CCM-A.cluster2.foo.com.
4. CCM3 sends a SIP Invite to 87522001@cluster2.foo.com, with destination address set to the IP address of CCM-A.
5. Unified CM interprets this call as a local call because the host portion of the uniform resource identifier (URI) matches the Cluster FQDN enterprise parameter. Cluster2 does not have any SIP trunk configured with a destination of CCM3, so it does a DNS SRV lookup for all domains configured under the SIP trunks with DNS SRV. In this case, the example shows a single trunk with a DNS SRV destination of cluster1.foo.com

6. The DNS server returns two entries, and one of them matches the source IP address of the Invite. The cluster accepts the call and extends it to extension 87522001.

High Availability for SIP Trunks

A variety of Unified CM options is available for configuring high availability with SIP trunks, all of which can be combined to provided redundancy and resiliency for both the source and destination servers of SIP trunks. These options can be categorized as follows:

- Multiple Source Unified CM Servers for Originating SIP Trunk Calls, page 14-19
- Multiple Destination IP Addresses per SIP Trunk, page 14-20
- Multiple SIP Trunks Using Route Lists and Route Groups, page 14-20
- SIP OPTIONS Ping, page 14-20

Multiple Source Unified CM Servers for Originating SIP Trunk Calls

Using Standard Unified CM Groups

The nodes defined in the Unified CM Group associated with an individual trunk make up the set of servers that can place or receive calls over that trunk. Up to three nodes can be defined in a Unified CM Group, thus ensuring high availability of the trunk itself.

Using Run on All Active Unified CM Nodes

The **Run on all Active Unified CM Nodes** feature creates and enables a SIP trunk instance on each call processing subscriber within the cluster, thus allowing these nodes to place or receive calls over the trunk.

The Unified CM Route Local Feature And Its Effect on Subscriber Selection for Outbound SIP Trunk Calls

The Route Local feature in Unified CM is designed to reduce intra-cluster traffic. The feature operates as illustrated by the following example:

> When a device such as a phone is making an outbound call over SIP Trunk 1, if an instance of SIP Trunk 1 is active on the same node as the one to which the phone is registered, then always use this co-located SIP Trunk 1 instance rather than internally routing the call to another SIP Trunk 1 instance on another node within the cluster.

The effect of the Route Local feature on node selection depends on whether Unified CM Groups or **Run on all Active Unified CM Nodes** is configured on the trunk. For trunks with **Run on all Active Unified CM Nodes** configured, the node to which the calling device is registered is used to make the outbound SIP trunk call. When Unified CM Groups are used on the trunk, if the calling device is registered to one of the nodes in the trunk's Unified CM Group, then the Route Local rule applies. If the calling device is not registered to one of the nodes in the trunk's Unified CM Group, then Unified CM will randomly distribute the call over the nodes in the trunk's Unified CM Group.

Using **Run on all Active Unified CM Nodes** is the recommended approach for SIP trunks because it allows call distribution across nodes to be determined by the calling device and it minimizes intra-cluster traffic.

Multiple Destination IP Addresses per SIP Trunk

A single SIP trunk can be configured with up to 16 destination IP addresses. Unified CM uses random distribution to the configured destination IP addresses when placing calls over a SIP trunk. Using multiple IP addresses on a SIP trunk can help to reduce the need to deploy multiple trunks with route lists and route groups.

Design Considerations When Using Run on All Active Unified CM Nodes

When using **Run on All Active Unified CM Nodes** in conjunction with multiple destination addresses, be aware that to accept inbound calls, the inbound source IP address received on the SIP trunk must match with a configured destination IP address on the inbound trunk. For example, if **Run on all Active Unified CM Nodes** is configured on the SIP intercluster trunk in each cluster, then each trunk must be configured with the corresponding destination address of every active node in the destination cluster. Where clustering over the WAN designs are deployed and geographic call distribution and failover are required, use standard Unified CM Groups on multiple intercluster trunks (each with up to three destination IP addresses) in conjunction with route lists and route groups.

Multiple SIP Trunks Using Route Lists and Route Groups

Multiple prioritized SIP trunks are often required to address failure scenarios in Unified Communications designs. These trunks should be configured in route groups in a single route list and associated with a route pattern. If Unified CM is not able to place a call over the selected trunk in the list, it will try the next trunk in the list. As a general recommendation, enable **Run on all Active Unified CM Nodes** for all route lists.

SIP OPTIONS Ping

SIP OPTIONS Ping can be enabled on the SIP Profile associated with a SIP trunk to dynamically track the state of the trunk's destination(s). When this option is enabled, each node running the trunk's SIP daemon will periodically send an OPTIONS Request to each of the trunk's destination IP addresses to determine its reachability. Enabling SIP OPTIONS Ping is recommended for all SIP trunks that require high availability because it allows Unified CM to dynamically track trunk state rather than determining trunk state on a per-call and timeout basis.

Load Balancing for SIP Trunks

When designing load balancing for SIP trunks, consider both the node that sources the call and its destination. With Unified CM SIP trunks, the node used to originate the call is determined by the Route Local rule, the number of nodes on which the outbound trunk is active, and whether a route list is used in conjunction with multiple outbound trunks. These considerations are discussed in the following sections.

Outbound Calls over a Single SIP Trunk

A single SIP trunk can run on up to three Unified CM nodes in a Unified CM Group, or it can run on all active Unified CM nodes in the cluster. To select the source node for outbound calls, Unified CM applies the following decision processes:

- Where an instance of the trunk runs on all nodes, the Route Local rule applies and the node used for each outbound call is determined by the node on which the call arrives (for example, the node to which the calling phone is registered or the node on which the inbound trunk call arrives).
- Where Unified CM Groups are used, the Route Local rule still applies for those calling devices that are registered to the same node as the nodes in the trunk's Unified CM Group. For calling devices that are registered to other servers within the cluster, Unified CM will randomly distribute calls across the nodes in the trunk's Unified CM Group. Unified CM uses round-robin call distribution across the trunk's configured destination addresses. SIP trunks may be configured with up to 16 destination IP addresses.

Outbound Calls over Multiple SIP Trunks

Because SIP trunks can run on all active Unified CM nodes and have up to 16 destination addresses, multiple SIP trunks typically do not need to be used to provide even call distribution between two Unified Communications systems. Where multiple trunks are used with route lists and route groups, route lists should be enabled to run on all active Unified CM nodes. Multiple SIP trunks are often used in conjunction with route lists to provide failover to the PSTN or to a group of Unified CM servers in a different site as part of a cluster deployed over the WAN. The selection of the Unified CM node used to initiate an outbound SIP trunk call, and the distribution of calls over the trunk's configured destination IP addresses, are determined in the same way as described for single trunks. Where clustering over the WAN designs are deployed and geographic call distribution and failover are required, use multiple intercluster trunks (each with up to three destination IP addresses) with standard Unified CM Groups in conjunction with route lists and route groups.

SIP OPTIONS Ping

Use OPTIONS Ping to dynamically track the state of each destination IP address on each SIP trunk and the collective state of the trunk as a whole. If a destination address is unreachable, Unified CM will not extend calls to this device. When all destinations are unreachable, the SIP trunk is considered to be out-of-service.

SIP Delayed Offer and Early Offer

Cisco Unified CM uses the SIP Offer/Answer model for establishing SIP sessions, as defined in RFC 3264. In this context, an Offer is contained in the Session Description Protocol (SDP) fields sent in the body of a SIP message. The Offer typically defines the media characteristics supported by the device (media streams, codecs, directional attributes, IP address, and ports to use). The device receiving the Offer sends an Answer in the SDP fields of its SIP response, with its corresponding matching media streams and codec, whether accepted or not, and the IP address and port on which it wants to receive the media streams. Unified CM uses this Offer/Answer model to establish SIP sessions as defined in the key SIP standard, RFC 3261.

RFC 3261 defines two ways that SDP messages can be sent in the Offer and Answer. These methods are commonly known as Delayed Offer and Early Offer, and support for both methods by User Agent Client/Servers is a mandatory requirement of the specification. In the simplest terms, an initial SIP Invite sent with SDP in the message body defines an Early Offer, whereas an initial SIP Invite without SDP in the message body defines a Delayed Offer.

In an Early Offer, the session initiator (calling device) sends its capabilities (for example, codecs supported) in the SDP contained in the initial Invite (thus allowing the called device to choose its preferred codec for the session). In a Delayed Offer, the session initiator does not send its capabilities in the initial Invite but waits for the called device to send its capabilities first (for example, the list of codecs supported by the called device, thus allowing the calling device to choose the codec to be used for the session).

Delayed Offer and Early Offer are the two options available to all standards-based SIP switches for media capabilities exchange. Most vendors have a preference for either Delayed Offer or Early Offer, each of which has its own set of benefits and limitations. For Unified CM SIP trunks, both SIP Delayed Offer and SIP Early Offer are supported. Unified CM Delayed Offer trunks do not require an MTP for the SIP Offer/Answer exchange. For Unified CM Early Offer trunks, SIP **Early Offer for voice and video (insert MTP if needed)** is the preferred configuration option rather than SIP Early Offer using **MTP Required** because MTP resources are typically not required to establish calls over an Early Offer trunk.

Note Unified CM can support Delayed Offer in one direction and Early Offer in the other direction over a SIP trunk. This capability can be useful in situations where a SIP switch connected to Unified CM by a SIP trunk wishes to control the codecs offered and selected for both inbound and outbound calls.

Early Media

In certain circumstances, a SIP session might require that a media path be set up prior to the finalization of the media capabilities exchange between the two SIP endpoints. To this end, the SIP protocol allows the establishment of Early Media after the initial Offer has been received by an endpoint. Some reasons for using Early Media include:

- The called device might want to establish an Early Media RTP path to reduce the effects of audio cut-through delay (clipping) for calls experiencing long signaling delays or to provide a network-based voice message to the caller.
- The calling device might want to establish an Early Media RTP path to access a DTMF or voice-driven interactive voice response (IVR) system.

The requirement to send one-way Early Media is one reason why service providers offering IP PSTN services accept only Early Offer SIP calls. With Early Offer, the service provider can stream one-way media to the caller after receipt of the initial INVITE with its SDP body containing the caller's media characteristics. An example of the need to stream Early Media to a caller would be where a caller calls a non-existent number and the service provider needs to send a network announcement to the caller without billing them for the attempted call (typically billing commences after two-way media has been established and when the final ACK is received). Unified CM supports the receipt of one-way media with SIP Early Offer calls.

Early two-way media cut-through can also be achieved by enabling Provisional Reliable Acknowledgement (PRACK) on each SIP Unified Communications system. PRACK allows the SIP Offer and Answer to be sent reliably in provisional responses (for example, 1XX responses), thus reducing the number of messages that need to be exchanged before two-way media can be established.

Unified CM supports PRACK-based Early Media for both Early Offer and Delayed Offer calls.

For a SIP trunk to support Early Media cut-through, you must enable PRACK through the **SIP Rel1XX Options** feature in the SIP Profile associate with the trunk.

Note The terms Early Offer and Early Media are often confused, but they are not the same.

Media Termination Points

MTPs are used by Unified CM for the following purposes:

- To deliver a SIP Early Offer over SIP trunks
- To address DTMF transport mismatches
- To act as an RSVP agent
- To act as a Trusted Relay Point (TRP)
- To provide conversion between IPv4 and IPv6 for voice RTP streams

Either of the following methods can be used to enable Early Offer on SIP trunks:

- Check the **MTP Required** checkbox on the SIP trunk

  In this case an MTP is used for every outbound call, and only voice calls using a single codec are supported.

- Check the **Early Offer support for voice and video calls (insert MTP if needed)** checkbox on the SIP Profile associated with the SIP trunk

  With this method an MTP is inserted only if the calling device or trunk cannot send all of the information about its media capabilities in the initial SIP Invite (for example, an inbound call to Unified CM from a SIP Delayed Offer or H.323 Slow Start trunk). In this case, when an MTP is used, additional voice codecs can be supported in the initial call setup by using the MTP's pass-through codec. Once established, this audio call can be escalated to support video and encryption if the call's media is renegotiated (for example, after hold/resume). When an MTP is not needed, all calls support voice, video, and encrypted media.

Unified CM SIP Delayed Offer and Early Offer Recommendations

Cisco Unified CM SIP trunks support Delayed Offer (Invite without SDP) by default. Media termination points (MTPs) are generally not required for Delayed Offer calls from Unified CM SIP trunks and therefore voice, video, and encrypted calls are all supported.

In cases where SIP Early Offer is required on Unified CM SIP trunks, Cisco recommends **Early Offer support for voice and video calls (insert MTP if needed)** because fewer MTP resources are required in comparison with **MTP Required**. When MTPs are used with **Early Offer support for voice and video calls (insert MTP if needed)**, they can provide support for voice, video, and encrypted media.

For IP PSTN SIP trunk connections, SIP Early Offer is generally required by the Service Provider. For designs where the IP PSTN needs to support large numbers of concurrent calls, the Cisco Unified Border Element’s SIP Delayed Offer to Early Offer feature can be used as an alternative to using Early Offer on the Unified CM SIP trunk, thereby potentially eliminating MTP usage.

For calls inbound and outbound from Unified CM, endpoints can negotiate the use of RFC 2833 or an out-of-band DTMF method (for example, KPML) end-to-end. If a common DTMF method cannot be negotiated between the endpoints, Unified CM will insert an MTP dynamically.

MTPs are available in three forms:

- Software MTPs in Cisco IOS gateways — Available with any Cisco IOS T-train software release and scaling up to 5,000 sessions (calls) on the Cisco ASR 1000 Series Aggregation Services Routers with Route Processor RP2.
- Hardware MTPs in Cisco IOS gateways — Available with any Cisco IOS T-train software release, hardware MTPs use on-board DSP resources and scale calls according to the number of DSPs supported on the Cisco router platform.
- Cisco Unified CM software MTPs using the Cisco IP Voice Media Streaming Application on a Cisco Media Convergence Server (MCS).

In general, Cisco IOS MTPs are recommended over Unified CM MTPs because Cisco IOS MTPs provide additional functionality such as support for additional codec types and the pass-through codec. (For details, see Media Termination Point (MTP), page 17-12.)

The following example configuration is for a Cisco IOS software-based MTP:

```
!
sccp local Vlan5
sccp ccm 10.10.5.1 identifier 5 version 5.0.1
! Communications Manager IP address (10.10.5.1)
sccp
!
sccp ccm group 5
 bind interface Vlan5
 associate ccm 5 priority 1
 associate profile 5 register MTP000E83783C50
! MTP name (MTP000E83783C50) ... must match the Unified CM MTP name.
!
dspfarm profile 5 mtp
 description software MTP
 codec g711ulaw
 codec pass-through
 maximum sessions software 500
 associate application SCCP
```

DTMF Transport

There are several methods of transporting DTMF information between SIP endpoints. In general terms, these methods can be classified as out-of-band (OOB) and in-band signaling. In-band DTMF transport methods send either raw or signaled DTMF tones within the RTP stream, and they need to be handled and interpreted by the endpoints that generate and/or receive them. Out-of-band signaling methods transport DTMF tones outside of the RTP path, either directly to and from the endpoints or through a call agent such as Cisco Unified CM, which interprets and/or forwards these tones as required.

Out-of-band (OOB) SIP DTMF signaling methods include Unsolicited Notify (UN), Information (INFO), and Key Press Markup Language (KPML). While KPML (RFC 4730) is the OOB signaling method preferred by Cisco, KPML is not widely used in the market place at this time. Currently, the only known products supporting KPML are Cisco Unified CM, Cisco IOS Gateways (Release 12.4 and later), and some models of Cisco IP Phones. INFO is not supported by Unified CM.

In-band DTMF transport methods send DTMF tones as either raw tones in the RTP media stream or as signaled tones in the RTP payload using RFC 2833. Among SIP product vendors, RFC 2833 has become the predominant method of sending and receiving DTMF tones and is supported by the majority of Cisco voice products.

Because in-band signaling methods send DTMF tones in the RTP media stream, the SIP endpoints in a session must either support the transport method used (for example, RFC 2833) or provide a method of intercepting this in-band signaling and converting it. If the two endpoints are using a back-to-back user agent (B2BUA) server for the call control (for example, Cisco Unified CM) and the endpoints negotiate different DTMF methods between each device and call control box, then the call agent determines how to handle the DTMF differences, either through MTP insertion or by OOB methods. With Unified CM, a DTMF transport mismatch (for example, in-band to out-of-band DTMF) is resolved by inserting a media termination point (MTP), which terminates the RTP stream with in-band DTMF signaling (RFC 2833), extracts the DTMF tones from the RTP stream, and forwards these tones out-of-band to Unified CM, where they are then forwarded to the endpoint supporting out-of-band signaling. In this case, the MTP is always in the media path between the two endpoints because there is no MTP codec dependency for DTMF translation.

In-band DTMF tones can also be transported as raw (audible) tones in the RTP media stream. This transport method is not widely supported by Cisco products and, in general, is not recommended as an end-to-end DTMF transport mechanism. In-band audio DTMF tones can generally be reproduced reliably only when using G.711 a-law or mu-law codecs, and they are not suitable for use with low-bandwidth codecs. In cases where in-band audio is the only available DTMF transport mechanism, the Cisco Unified Border Element can be used to translate the in-band audio DTMF signaling into RFC 2833 signaling.

Over Unified CM SIP trunks, Cisco recommends configuring the DTMF Signaling Method to **No Preference**. This setting allows Unified CM to make an optimal decision for DTMF and to minimize MTP allocation.

SIP Trunk Transport Protocols

SIP trunks can use TCP, TLS (which runs over TCP), or UDP as a message transport protocol. As a reliable, connection-orientated protocol that maintains the connection state, TCP is recommended within Cisco Enterprise Unified Communications networks. UDP is not connection-orientated or reliable (message delivery is not guaranteed), and it relies on the SIP Invite Retry count and SIP Trying timers to detect and respond to far-end device failures. Use SIP OPTIONS Ping to dynamically track the state of each destination IP address on each SIP trunk and the collective state of the trunk as a whole.

For more information on SIP trunk timer tuning, refer to the configuration example and technical notes at:

http://www.cisco.com/en/US/products/sw/voicesw/ps556/products_configuration_example09186a008082d76a.shtml

Note Although TCP is the recommended transport protocol within a Cisco Enterprise Unified Communications network, most service providers prefer to use UDP because it has a lower processing overhead than TCP. Cisco Unified Border Element can be used to provide TCP-based SIP trunk connections to the Enterprise Unified Communications network and UDP-based SIP trunk connections to service provider networks.

Secure SIP Trunks

Securing SIP trunks involves two processes:

- Configuring the trunk to encrypt media (see Media Encryption, page 14-26)
- Configuring the trunk to encrypt signaling (see Signaling Encryption, page 14-26)

Media Encryption

Media encryption can be configured on SIP trunks by checking the trunk's **SRTP allowed** check box. It is important to understand that enabling **SRTP allowed** causes the media for calls to be encrypted, but the trunk signaling will not be encrypted and therefore the session keys used to establish the secure media stream will be sent in the clear. It is therefore important that you ensure that signaling between Unified CM and its destination SIP trunk device is also encrypted so that keys and other security-related information do not get exposed during call negotiations.

Signaling Encryption

SIP trunks use TLS for signaling encryption. TLS is configured on the SIP Security Profile associated with the SIP trunk, and it uses X.509 certificate exchanges to authenticate trunk devices and to enable signaling encryption.

Certificates can be either of the following:

- Imported to each Unified CM node from every device that wishes to establish a TLS connection to that node's SIP trunk daemon
- Signed by a Certificate Authority (CA), in which case there is no need to import the certificates of the remote devices; only the CA certificate needs to be imported.

Unified CM provides a bulk certificate import and export facility. However, for SIP trunks using **Run on all Active Unified CM Nodes** and up to 16 destination addresses, using a Certificate Authority provides a centralized and less administratively burdensome approach to setting up signaling encryption on SIP trunks.

For more information on TLS for SIP trunks, refer to the latest version of the *Cisco Unified Communications Manager Security Guide*, available at

http://www.cisco.com/en/US/products/sw/voicesw/ps556/prod_maintenance_guides_list.html

For information on certificate authorities, refer to the Certificate Authority (CA) information in the latest version of the *Cisco Unified Communications Operating System Administration Guide*, available at

http://www.cisco.com/en/US/products/sw/voicesw/ps556/prod_maintenance_guides_list.html

If the system can establish a secure media or signaling path and if the end devices support SRTP, the system uses a SRTP connection. If the system cannot establish a secure media or signaling path or if at least one device does not support SRTP, the system uses an RTP connection. SRTP-to-RTP fallback (and vice versa) may occur for transfers from a secure device to a non-secure device or for conferencing, transcoding, music on hold, and so on.

For SRTP-configured devices, Unified CM classifies a call as encrypted if the **SRTP Allowed** check box is checked for the device and if the SRTP capabilities for the devices are successfully negotiated for the call. If these criteria are not met, Unified CM classifies the call as non-secure. If the device is connected to a phone that can display security icons, the phone displays the lock icon when the call is encrypted.

Note MTPs that are statically assigned to a SIP trunk by means of the **MTP Required** checkbox do not support SRTP because they do not support the pass-through codec.

To ensure that SRTP is supported for all calls, configure the SIP trunk for Delayed Offer.

Where **Early Offer support for voice and video calls (insert MTP if needed)** is configured, for devices that support encryption, all calls that do not need to use MTPs can support SRTP. When an MTP is inserted into the call path, this dynamically inserted MTP supports the pass-through codec, and encrypted calls are supported in the following cases:

- If the calling device is an older SCCP-based phone registered to Unified CM, SRTP can be negotiated in the initial call setup.
- If the call arrives inbound to Unified CM on a Delayed Offer SIP trunk or an H.323 Slow Start trunk, SRTP will not be negotiated in the initial call setup because no security keys are available, but the call can be escalated mid-call to support SRTP if the call media is renegotiated (for example, after hold/resume).

If Unified CM dynamically inserts an MTP for reasons other than Early Offer, such as for a Trusted Relay Point or as an RSVP agent, then SRTP will be supported with an MTP that supports the pass-through codec.

Note that **dtmf-relay** using an MTP (where the MTP needs to convert between in-band and out-of-band DTMF signals) will not function for SRTP because it will be unable to decrypt the DTMF packets in the media stream.

Note SRTP is not supported over SAF-enabled SIP trunks.

Calling Party Number Transformation and SIP Trunks

Unified CM provides the capability to transform calling party numbers of calls inbound over gateways and trunks to a normalized format. Typically, you would want this format to be the globally routable international representation of the number according to E.164 specifications.

The process of normalization relies on receiving the number and the associated number-type of the incoming call. The number-type parameter can be used to select the appropriate digits to prefix to the calling number. Number-types can be one of four types: Unknown, Subscriber, National, or International. For more details and examples on how these number-types are used, refer to the chapter on Dial Plan, page 9-1.

You can specify the prefix digits for each of the four number types in the H.323 trunk and H.323 gateway configuration pages in Unified CM. H.323 can transport these number types in its signaling. SIP, on the other hand, is unable to transport the number-type information in its signaling. Thus, a call coming in through a SIP gateway across a SIP trunk to Unified CM will not have any indication of whether the calling-party number is local, national, or international. Without the number-type information, Unified CM is unable to apply the correct prefix to the calling-party number.

The inability of the SIP trunk to transport the number type implies that the normalization of the calling number must be performed before the call is presented to Unified CM. One place where the transformation can be performed is on the ingress SIP gateway. The following example configuration shows the translation rules that can be defined on a Cisco IOS gateway to accomplish this transformation:

```
voice translation-rule 1
 rule 1 // /+4940/ type subscriber subscriber
 rule 2 // /+49/ type national national
 rule 3 // /+/ type international international
...
voice translation-profile 1
 translate calling 1
...
dial-peer voice 300 voip
```

```
 translation-profile outgoing 1
 destination-pattern .T
 session protocol sipv2
 session target ipv4:9.6.3.12
...
```

When configured as in the example above, a Cisco IOS gateway using SIP to communicate with Unified CM will send calling party information digits normalized to the E.164 format, including the + sign. The Unified CM configuration will receive all calls from this gateway with a numbering type of "unknown" and would not need to add any prefixes.

For more details on configuring translation rules, refer to the document *Voice Translation Rules*, available at

http://www.cisco.com/en/US/tech/tk652/tk90/technologies_tech_note09186a0080325e8e.shtml

Unified CM can set the calling party number of outgoing calls to the normalized global format. The number-type in outgoing calls from the SIP trunk will be “unknown,” and the Cisco IOS gateway should change it to International if no stripping is done, or perform a combination of stripping and numbering type change if required by the connected service provider.

SIP Trunk Service Types

Most SIP trunks are general-purpose trunks capable of connecting to a wide variety of SIP servers such as other Cisco Unified CMs, Cisco Unified Border Elements, Cisco Unified Gateways, and so forth. In addition to these all-purpose trunks, Unified CM provides SIP trunks dedicated for specific services. These special-purpose trunks enable technologies such as the following:

- Cisco Intercompany Media Engine (IME)

  See Cisco Intercompany Media Engine, page 5-75.
- Cisco Unified Communications Call Control Discovery (CCD) through the Cisco IOS Service Advertisement Framework (SAF)

  See Service Advertisement Framework (SAF), page 3-69.
- Cisco Extension Mobility Cross Cluster (EMCC)

  See Extension Mobility Cross Cluster (EMCC), page 19-10.

Design Considerations for SIP Trunks

Considerations for SIP Intercluster Trunks

For intercluster trunk connections, the SIP trunk configured in each cluster may be using standard Unified CM Groups or the **Run on all Active Unified CM Nodes** feature. The reasons for using each type of feature will typically be determined by the Unified CM version used in the cluster, or if clustering over the WAN has been deployed and geographically based call distribution is required.

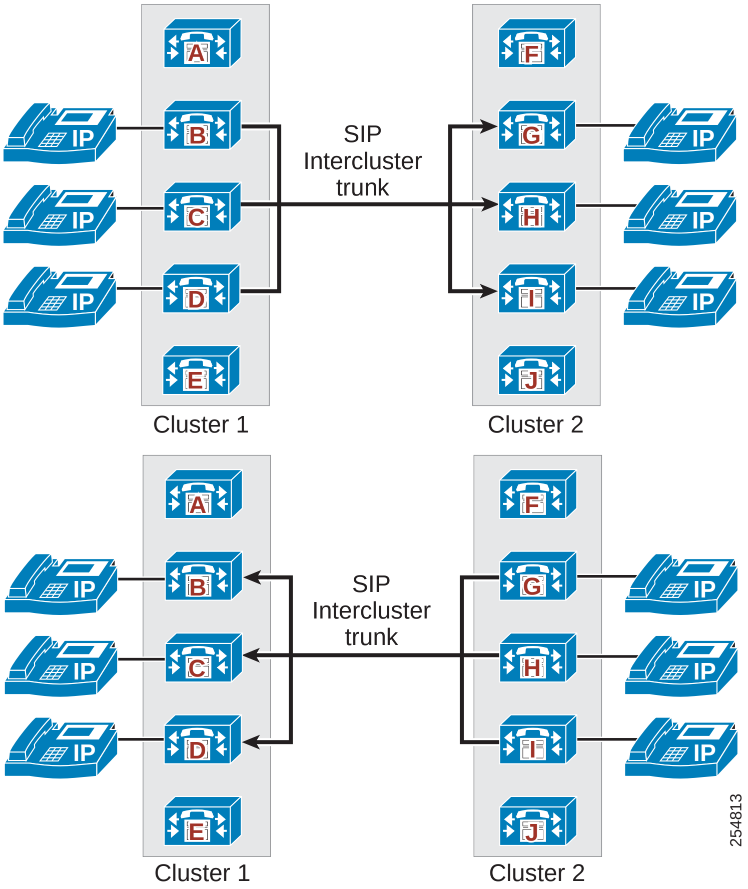
Using Standard Unified CM Groups with SIP Intercluster Trunks

In this type of deployment standard Unified CM Groups are used by SIP intercluster trunks in each cluster. When defining this type of trunk with standard Unified CM Groups, you should define a maximum of three remote Unified CM servers as destination IP addresses in the remote cluster. The trunk will automatically load-balance across all defined remote Unified CM servers. In the remote cluster, it is important to configure a corresponding SIP intercluster trunk that has the same Unified CM nodes in its Unified CM Group as those defined as remote destination Unified CM servers in the first cluster.

For example, if Cluster 1 has a SIP trunk to Cluster 2 and Cluster 2 has a SIP trunk to Cluster 1, the following configurations would be needed (see Figure 14-10):

- Cluster 1
  - Servers B, C, and D are configured as members of the Unified CM Group defined in the device pool associated with the SIP trunk to Cluster 2.
  - The SIP trunk has Cluster 2's remote servers G, H, and I configured as destinations.
- Cluster 2
  - Servers G, H, and I are configured as members of the Unified CM Group defined in the device pool associated with the SIP trunk to Cluster 1.
  - The SIP trunk has Cluster 1's remote servers B, C, and D configured as destinations.

Figure 14-10 SIP Intercluster Trunks with Unified CM Groups



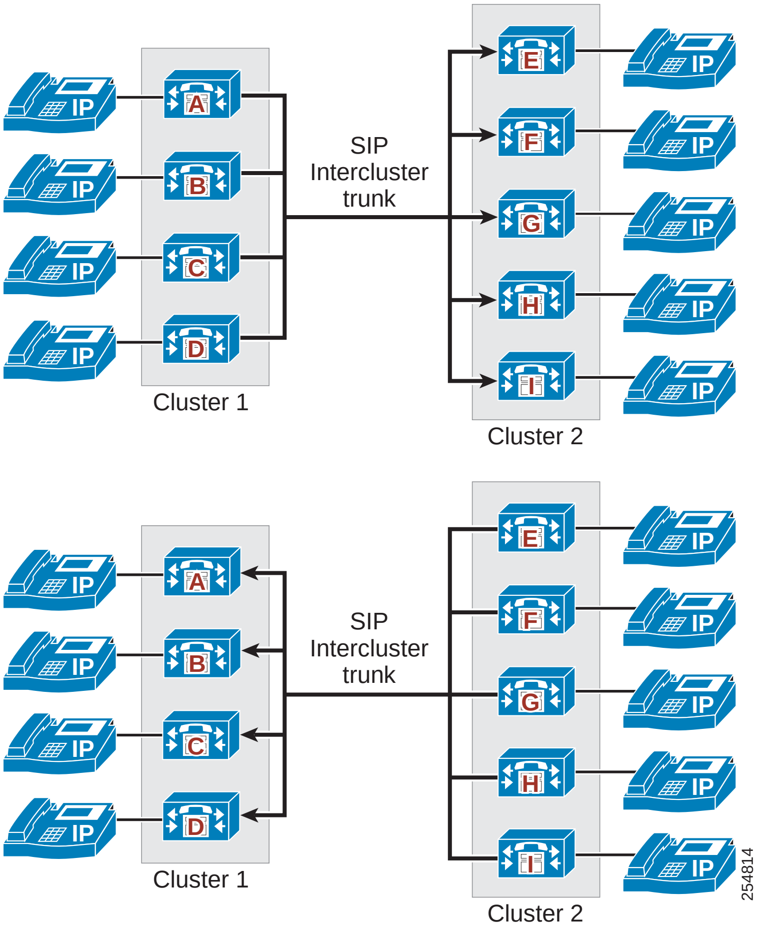
Using Run on All Active Unified CM Nodes with SIP Intercluster Trunks

In this type of deployment, **Run on all Active Unified CM Nodes** is used by SIP intercluster trunks in each cluster. When defining this type of trunk you may define up to 16 remote Unified CM servers in the destination cluster. (The number of remote servers that you need to define will depend on the number of active Unified CM nodes in the destination cluster.) The trunk will automatically load-balance calls across all defined remote destination servers. In the remote cluster, it is important to configure a corresponding SIP intercluster trunk that has **Run on all Active Unified CM Nodes** configured, where these nodes are defined as the remote destination Unified CM servers in the first cluster.

For example, if Cluster 1 (with four active nodes) has a SIP trunk to Cluster 2, and Cluster 2 (with five active nodes) has a SIP trunk to Cluster 1, the following configurations would be needed (see Figure 14-11):

- Cluster 1 has four active Unified CM nodes (A, B, C, and D).
  - Enabling **Run on all Active Unified CM Nodes** causes servers A, B, C, and D to have active SIP trunk daemons associated with the SIP trunk to Cluster 2.
  - The SIP trunk has Cluster 2's remote servers E, F, G, H, and I configured as destinations.
- Cluster 2 has five active Unified CM nodes (E, F, G, H, and I).
  - Enabling **Run on all Active Unified CM Nodes** causes servers E, F, G, H, and I to have active SIP trunk daemons associated with the SIP trunk to Cluster 1.
  - The SIP trunk has Cluster 1's remote servers A, B, C, and D configured.

Figure 14-11 SIP Intercluster Trunks Running on All Active Unified CM Nodes



Using Standard Unified CM Groups and Run on All Active Unified CM Nodes with SIP Intercluster Trunks

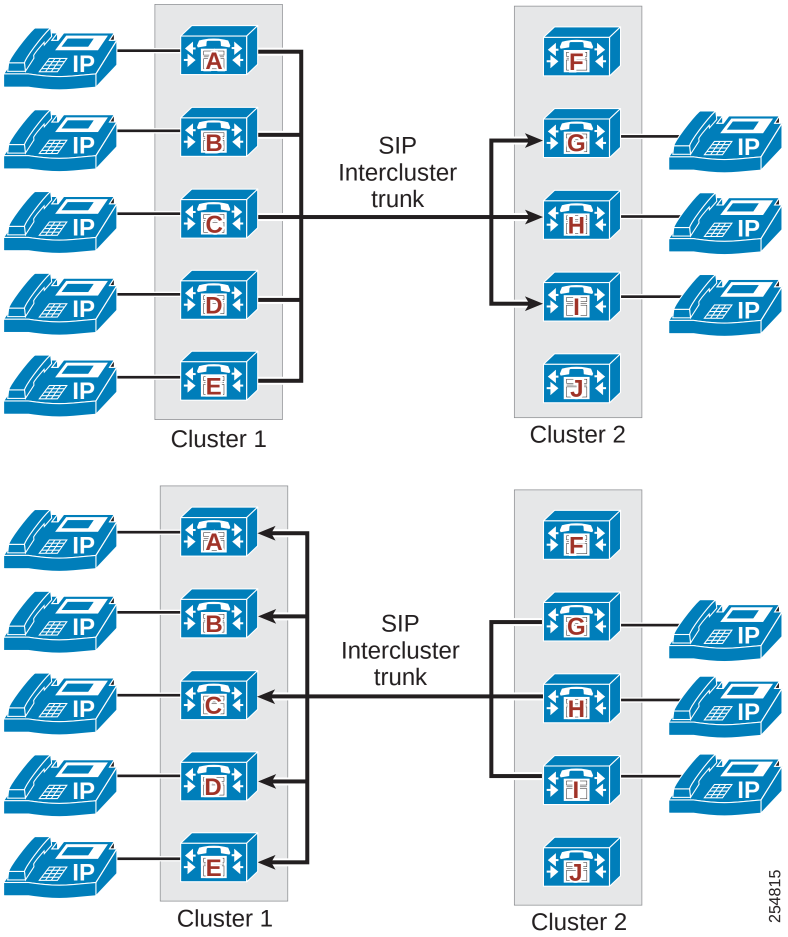
In this type of deployment, **Run on all Active Unified CM Nodes** is used by the SIP intercluster trunk in one cluster and standard Unified CM Groups are used by the SIP intercluster trunk in the other cluster. When configuring these trunks, the number of remote Unified CM server destinations that you define should match the number of active Unified CM nodes associated with the corresponding trunk in the destination cluster. The trunk will automatically load-balance calls across all defined remote destination Unified CM servers. In the remote cluster, it is important to configure a corresponding SIP intercluster trunk that has Unified CM nodes with active SIP daemons where these nodes are defined as remote destination Unified CM servers in the first cluster.

For example, if Cluster 1 has a trunk to Cluster 2, and Cluster 2 has a trunk to Cluster 1, the following configurations would be needed (see Figure 14-12):

- Cluster 1 has five active Unified CM nodes (A, B, C, D, and E).
  - Enabling **Run on all Active Unified CM Nodes** causes servers A, B, C, D, and E to have active SIP trunk daemons associated with the SIP trunk to Cluster 2.

  - The SIP trunk has Cluster 2's remote servers G, H, and I configured as destinations.
- Cluster 2 has five active Unified CM nodes and uses an intercluster trunk with a Unified CM Group containing nodes G, H, and I.
  - Servers G, H, and I are configured as members of the Unified CM Group defined in the device pool associated with the SIP trunk to Cluster 1.
  - The SIP trunk has Cluster 1's remote servers A, B, C, D, and E configured as destinations.

Figure 14-12 SIP Intercluster Trunks Using Unified CM Groups and Run on All Active Unified CM Nodes



Trunk Type and Feature Recommendations for Multi-Cluster Deployments

Multiple Clusters All Running Unified CM 8.5 or Later Releases

Where all clusters are running Unified CM 8.5 or later releases, the following SIP trunk features should be used where applicable (see Figure 14-13):

- SIP OPTIONS Ping
- SIP Delayed Offer
- Early Offer support for Voice and Video (insert MTP if needed)
- Run on All Active Unified CM Nodes
- Multiple destination IP addresses
- Audio codec preference lists
- QSIG over SIP
- SIP Normalization and Transparency

Deploying these features reduces MTP usage and provides high availability, even call distribution, and dynamic SIP trunk failure detection. For Unified CM SIP trunks, both SIP Delayed Offer and SIP Early Offer can be used. Unified CM Delayed Offer trunks do not require an MTP for the SIP Offer/Answer exchange. For Unified CM Early Offer trunks, **SIP Early Offer for voice and video (insert MTP if needed)** is the preferred configuration option rather than SIP Early Offer using **MTP Required** because MTP resources are typically not required to establish calls over an Early Offer trunk.

In general, for outbound calls from Unified CM SIP trunks, Delayed Offer is preferred because MTPs are not required to establish a Delayed Offer call. For inbound calls to Unified CM SIP trunks, Early Offer or Delayed Offer (or a mixture of both Early Offer and Delayed Offer) can be used.

SIP intercluster trunks support voice; video, and encrypted media between Unified CM clusters, and all of the above features can be used. If multiple trunks are used with route lists, enable the **Run on All Active Unified CM Nodes** feature on the route lists.

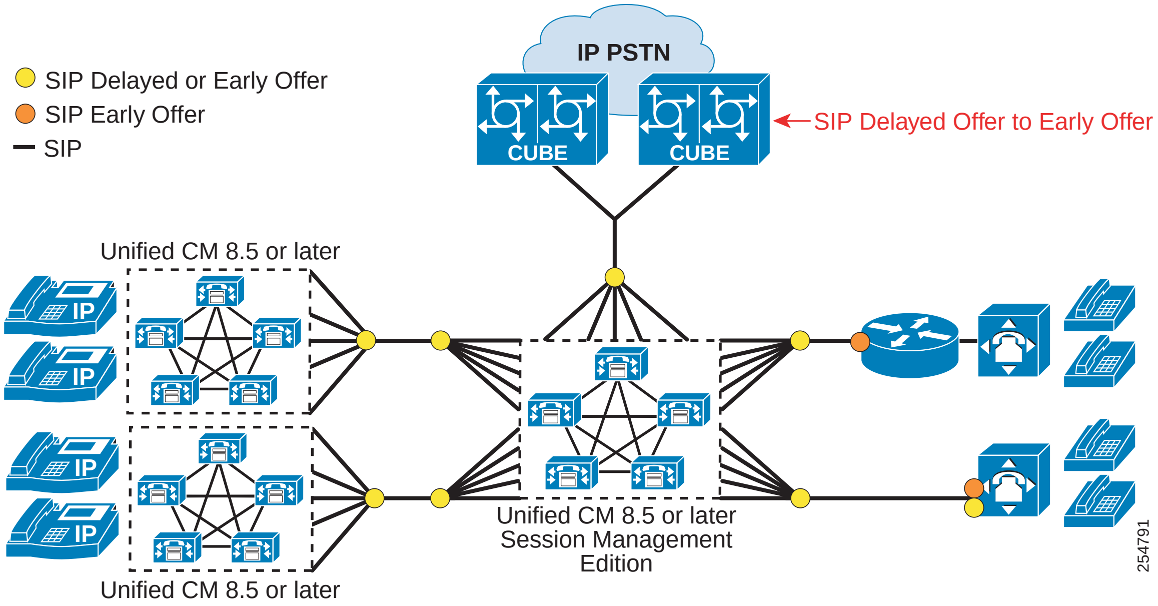
For SIP trunks to an IP PSTN, SIP Early Offer is typically required by the service provider, and most providers support voice calls only. However, if required, video calls and encrypted media are also supported. Where SIP Early Offer is required by the Service Provider, the Cisco Unified Border Element (SIP Delayed Offer to Early Offer feature) can be used as an alternative to configuring Early Offer on the Unified CM SIP trunk. For inbound calls to Unified CM SIP trunks, Early Offer or Delayed Offer (or a mixture of both Early Offer and Delayed Offer) can be used. When connecting to a service provider's IP PSTN network, Cisco strongly recommends the use of the Cisco Unified Border Element as an enterprise edge Session Border Controller to provide a controlled demarcation and security point between your enterprise and the service provider's network.

SIP trunks to third-party unified communications systems may support voice, video, and encrypted media. Check the capabilities of the end system to determine the SIP trunk features and media capabilities that it supports. For outbound calls from Unified CM SIP trunks, Delayed Offer or Early Offer can be used. For inbound calls to Unified CM SIP trunks, Early Offer or Delayed Offer (or a mixture of both Early Offer and Delayed Offer) can be used.

Note SIP trunks on Cisco IOS gateways always send Early Offer.

For SIP trunk connections to the IP PSTN and third-party unified communications systems, normalization and transparency scripts can be used to address SIP interoperability issues.

Figure 14-13 Multi-Cluster Deployments with Unified CM 8.5 and Later Releases



Multiple Clusters Running Unified CM 8.5 and Prior Releases

When the leaf clusters are running Unified CM 8.5 or a later release in combination with leaf clusters running prior releases of Unified CM, the following trunk types and features should be used (see Figure 14-14):

When the leaf cluster is running an earlier version (pre-8.5) of Unified CM and voice, video, and encryption are required, use H.323 Slow Start intercluster trunks and Annex M1 (QSIG) if desired. Deploy one or more H.323 Slow Start intercluster trunks using standard Unified CM Groups and up to three destination IP addresses. If multiple trunks are used with route lists, to avoid the Route Local rule (described earlier) ensure that the primary server in the route list's Unified CM Group does not reside on the same node as an associated outbound H.323 trunk.

For leaf clusters running Unified CM 8.5 and later releases, use a SIP Delayed Offer or Early Offer (for voice and video (insert MTP if needed)) intercluster trunk, enable **Run on All Active Unified CM Nodes**, and use multiple destination IP addresses and SIP OPTIONS Ping for high availability and even call distribution. If multiple trunks are used with route lists, enable the **Run on All Active Unified CM Nodes** feature on the route lists.

Using SIP Delayed Offer or Early Offer (for voice and video (insert MTP if needed)) intercluster trunks on Unified CM 8.5 (or later release) leaf clusters and H.323 Slow Start intercluster trunks on leaf clusters using earlier versions of Unified CM, allows voice, video, and encrypted calls to be made between clusters and reduces the number MTPs required.

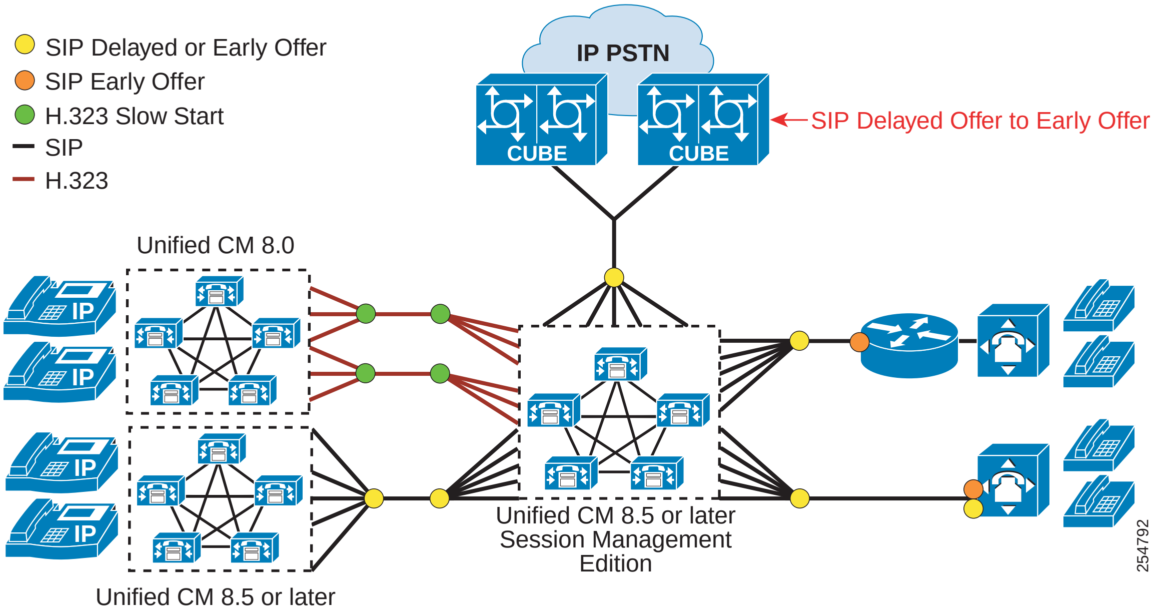
For SIP trunks to an IP PSTN, SIP Early Offer is typically required by the service provider, and most providers support voice calls only. However, if required, video calls and encrypted media are also supported. Where SIP Early Offer is required by the service provider, the Cisco Unified Border Element (SIP Delayed Offer to Early Offer feature) can be used as an alternative to configuring Early Offer on the Unified CM SIP trunk. For inbound calls to Unified CM SIP trunks, Early Offer or Delayed Offer (or a mixture of both Early Offer and Delayed Offer) can be used. When connecting to a service provider's IP PSTN network, Cisco strongly recommends the use of the Cisco Unified Border Element as an enterprise edge Session Border Controller to provide a controlled demarcation and security point between your enterprise and the service provider's network.

SIP trunks to third-party unified communications systems may support voice, video, and encrypted media. Check the capabilities of the end system to determine the SIP trunk features and media capabilities that it supports. For outbound calls from Unified CM SIP trunks, Delayed Offer or Early Offer can be used. For inbound calls to Unified CM SIP trunks, Early Offer or Delayed Offer (or a mixture of both Early Offer and Delayed Offer) can be used.

Note SIP trunks on Cisco IOS gateways always send Early Offer.

For Unified CM SIP trunk connections to the IP PSTN and third-party unified communications systems, normalization and transparency scripts can be used to address SIP interoperability issues.

Figure 14-14 Multi-Cluster Deployments with Unified CM 8.5 and Prior Releases



Trunk Design Considerations for Clustering over the WAN

When deploying clustering over the WAN for spatial resilience and redundancy, SIP trunk features such as OPTIONS Ping and QSIG can be used as required and appropriate. SIP and H.323 trunk features such as **Run on all Unified CM Nodes** and multiple destination addresses should be used with consideration, primarily because of the mechanism that trunks use to identify and accept inbound calls. (A trunk will accept a call if the incoming source IP address matches one of the addresses defined as its destination IP address.)

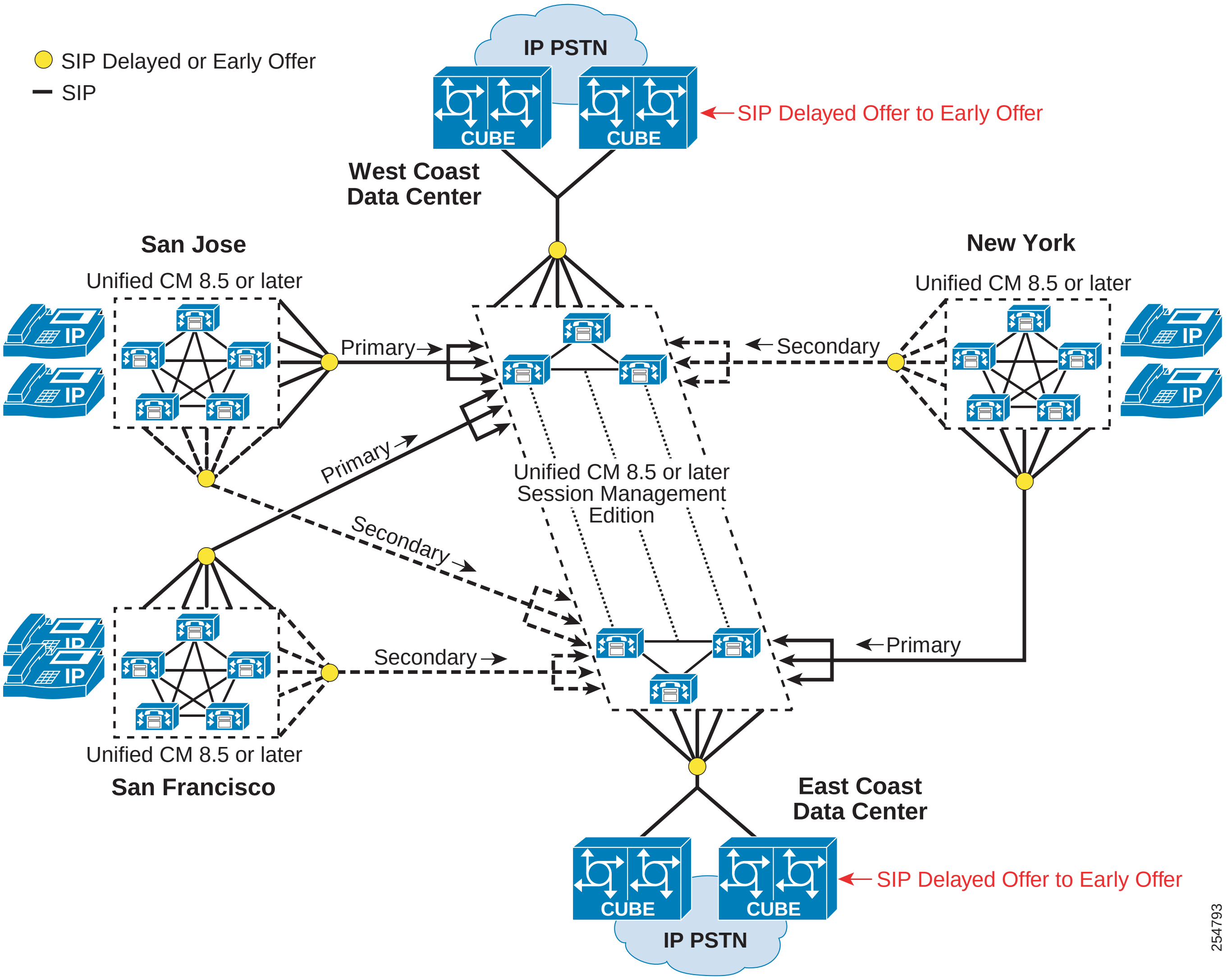
For clustering over the WAN deployments where calls need to be routed to different groups of Unified CM nodes based on their geographic location, consideration should be given to the trunk configuration for both inbound and outbound calls. This is described in the following section, using a Unified CM Session Management Edition cluster that is clustered over the WAN as an example.

Design Guidance for Clustering over the WAN with Leaf Cluster Trunks

Create and prioritize multiple SIP trunks in route lists in each leaf cluster to distribute calls to each group of Unified CM Session Management Edition nodes in each data center, and run route lists on all nodes. (See Figure 14-15.)

Enable **Run on all Nodes** on each leaf cluster SIP trunk (each SIP trunk must use a unique incoming port number). Define destination IP addresses per trunk for geographic call distribution.

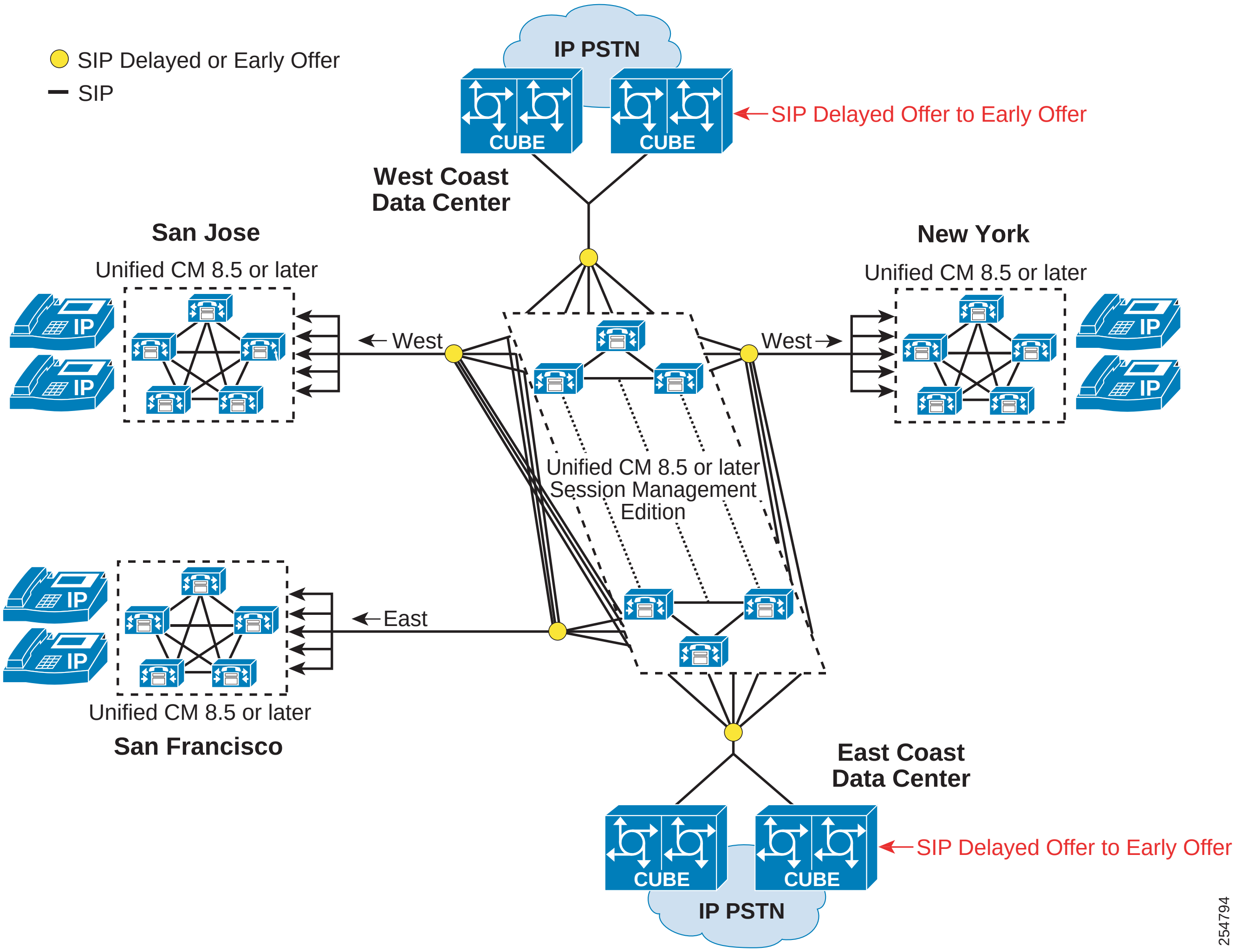
Figure 14-15 Calls from Leaf Clusters to Unified CM Session Management Edition



Design Guidance for Clustering over the WAN with Unified CM Session Management Edition Cluster Trunks

For each leaf cluster, create a single SIP trunk on the Unified CM Session Management Edition cluster. Enable **Run on all Unified CM Nodes** on this SME trunk and configure destination IP addresses for each call processing node in the leaf cluster. (See Figure 14-16.)

Figure 14-16 Calls from Unified CM Session Management Edition to Leaf Clusters



Other SIP Trunk Deployment Considerations

Voice clipping, if observed, can be minimized or eliminated by enabling PRACK on the trunk. The PRACK feature can be enabled via the **SIP Rel1XX Options** of the Trunk Specific Configuration section in the SIP trunk's SIP Profile Configuration. Note that PRACK must be supported by both SIP Unified Communications systems.

Other operating parameters for security settings and the types of messages accepted over a SIP trunk can be enabled in the SIP Trunk Security Profile. Here you can set parameters not only for TLS and Digest Authentication, but also for whether or not the trunk will accept Presence Subscription, an out-of-dialog REFER message, Replaces header, or an Unsolicited Notify message.

SIP trunks support topology-aware RSVP call admission control using SIP Preconditions and locations-based call admission control which is unaware of the underlying WAN topology.

For connection to service provider networks, Cisco recommends the use of the Cisco Unified Border Element. In addition to providing a demarcation point between the enterprise and service provider networks, the Cisco Unified Border Element can also be used for address hiding and enhancing SIP signaling interoperability between the two networks.

For more information on the Cisco Unified Border Element, refer to the documentation available at

http://www.cisco.com/en/US/products/sw/voicesw/ps5640/index.html

H.323 Trunks Overview

H.323 trunks provide connectivity to other H.323 devices such as gateways, Unified CM Session Management Edition, gatekeepers, Unified Communications applications, and other Unified CM clusters. Cisco Unified CM provides the following call routing enhancement for all H.323 trunk types:

- Run route lists on all Unified CM nodes

In addition to this, H.323 non-gatekeeper controlled intercluster trunks also support the following features:

- Run on all Unified CM nodes
- Up to 16 destination IP addresses per trunk

These two features improves outbound call distribution from Unified CM clusters and reduce the number of H.323 non-gatekeeper controlled intercluster trunks required between clusters.

These features and their operation are discussed in detail later in this section.

For the complete list of new enhancements for H.323 trunks, refer to the latest Cisco Unified Communications Manager product release notes available at

http://www.cisco.com/en/US/products/sw/voicesw/ps556/prod_release_notes_list.html

General H.323 Intercluster Trunk Deployment Considerations

Prior to Unified CM 8.5, H.323 Annex M1 trunks were the preferred choice for connections between Unified CM clusters. Unified CM SIP trunks now offer a greater set of features in comparison with H.323 intercluster trunks, thus making SIP the protocol of choice for intercluster trunk connections. However, the majority of Unified CM clusters using earlier software versions are likely to be deployed with H.323 Annex M1 intercluster trunks, and this may determine the intercluster trunk type that you use to these clusters.

Basic Operation of H.323 Trunks

H.323 trunks provide connectivity to other Unified CM clusters and other H.323 devices such as gateways. H.323 trunks support most of the audio and video codecs that Unified CM supports for intra-cluster communications, with the exception of wideband audio and wideband video.

H.323 trunks use the Empty Capabilities Set (ECS) to provide supplementary call services such as hold/resume and transfer. This method is a standard H.245 mechanism to stop or close a media stream (or channel) and start or open it to the same or a different endpoint address. This method allows Unified CM to keep a call active while still being able to control the source and destination of the media streams on the fly.

For example, consider a call between two clusters (A and B) using the H.323 trunk. When a user in cluster A places a user in cluster B on hold, the media streams between the two users are closed and the user in cluster B is connected to a music on hold (MoH) server in cluster A. The MoH server is instructed to send media (the music file) to the user. When the user in cluster A resumes the call, the MoH stream is closed and the two-way media streams are reopened between the two users. (Unified CM does not support H.450 for supplementary call services.) In this case, MoH is an example of an ECS operation. H.323 trunks support multicast MoH, therefore the media resource group list (MRGL) for the H.323 trunks can contain both unicast and multicast MoH sources. (For details, see Music on Hold, page 17-21.)

The bandwidth used for calls on H.323 trunks can be controlled by the use of regions configured in Unified CM and assigned to each trunk. A region limits the amount of bandwidth allocated for calls by specifying the inter-region Max Audio Bit Rate for audio and the inter-region Max Video Call Bit Rate setting for video (that includes audio). Calls between one region and another region must be within the specified bandwidth limit. If the device making the call over the H.323 trunk is in a more restrictive region or does not support a particular codec such as video, then it is a subset of codecs that are allowed for that call.

H.323 Trunk Types

The following major types of H.323 trunks can be configured in a Unified CM:

- Intercluster Trunk (Non-Gatekeeper Controlled), page 14-39
- Intercluster Trunk (Gatekeeper Controlled), page 14-46
- H.225 Trunk (Gatekeeper Controlled), page 14-47

Each of these H.323 trunk types and their specific design considerations are discussed in the following sections.

Intercluster Trunk (Non-Gatekeeper Controlled)

This trunk is the simplest H.323 trunk type and is used for connecting to other Unified CM clusters in either a multi-cluster single campus or a distributed call processing deployment. This trunk does not use a gatekeeper for call admission control, although it may use locations configured in Unified CM if bandwidth control is required.

Cisco Unified CM supports the following trunk features and call routing enhancements for H.323 non-gatekeeper controlled intercluster trunks:

- Run on all active Unified CM nodes
- Up to 16 destination IP addresses per trunk
- Run route lists on all Unified CM nodes

These features are discussed in the following sections.

H.323 Non-Gatekeeper Intercluster Trunks Running on all Active Unified CM Nodes

When the **Run on all Active Unified CM Nodes** option is checked on a H.323 non-gatekeeper intercluster trunk, Unified CM creates an instance of the H.323 trunk daemon on every call processing Unified CM Groups.) This allows H.323 non-gatekeeper intercluster trunk calls to be made or received on any call processing subscriber. With **Run on all Active Unified CM Nodes** enabled, outbound H.323 non-gatekeeper intercluster trunk calls originate from the same server on which the inbound call (for example, from a phone or trunk) is received. As with all Unified CM H.323 non-gatekeeper intercluster trunks, the H.323 daemons associated with the trunk will accept only inbound calls from end systems with IP addresses that are defined in the trunk's destination address fields. Running the H.323 non-gatekeeper intercluster trunk on all nodes is recommended where the H.323 a non-gatekeeper intercluster trunk is required to process a large number of calls, so that outbound and inbound call distribution can be evenly spread across all call processing subscribers within a cluster. Bear in mind that (unlike SIP trunks) H.323 non-gatekeeper intercluster trunks use a fixed destination port and an ephemeral source, and therefore H. 323 non-gatekeeper intercluster trunks cannot be differentiated using port numbers. When configuring H.323 non-gatekeeper intercluster trunks, make sure that each trunk uses different destination IP addresses when **Run on all Active Unified CM Nodes** is enabled.

Up to 16 Destination IP Addresses per H.323 Non-Gatekeeper Intercluster Trunk

An H.323 non-gatekeeper intercluster trunk can be configured with up to 16 destination IP addresses. Support for additional destination IP addresses reduces the need to create multiple trunks associated with route lists and route groups for call distribution between two Unified Communications systems, thus simplifying Unified CM trunk design. This feature can be used in conjunction with the **Run on all Active Unified CM Nodes** feature. Bear in mind, however, that the H.323 daemons associated with a Unified CM H.323 non-gatekeeper intercluster trunk will accept only inbound calls from end systems with IP addresses that are defined in the trunk's destination address fields.

Route Lists Running on All Active Unified CM Nodes

Although this is not specifically an H.323 non-gatekeeper intercluster trunk feature, running route lists on all nodes provides benefits for trunks in route lists and route groups. Running route lists on all nodes improves outbound call distribution by using the Route Local rule to avoid unnecessary intra-cluster traffic.

For route lists, the Route Local rule operates as follows:

> For outbound calls that use route lists and associated route groups and trunks, when a call from a registered phone or inbound trunk arrives at the node with the route list instance associated with the trunk selected for the outbound call, Unified CM checks to see if an instance of the selected outbound trunk exists on the same node as the route list. If so, Unified CM will use this node to establish the outbound trunk call.

If both the route list and the selected outbound trunk have **Run on all Active Unified CM Nodes** enabled, outbound call distribution will be determined by the node on which the inbound call arrives. When the selected outbound trunk uses Unified Groups instead of running on all nodes, Unified CM will apply the Route Local rule if an instance of the selected outbound trunk exists on the same node on which the inbound call arrived. If an instance of the trunk does not exist on this node, then Unified CM will forward the call (within the cluster) to a node where the trunk is active.

If the route list does not have **Run on all Active Unified CM Nodes** enabled, the route list will be active on one node within the cluster (the primary node in the route list's Unified Group) and the Route Local rule will be applied on this node.

As a general recommendation, **Run on all Active Unified CM Nodes** should be enabled for all route lists.

Design Considerations for H.323 Non-Gatekeeper Intercluster Trunks

For intercluster trunk connections, the H.323 non-gatekeeper intercluster trunk configured in each cluster may be using standard Unified CM Groups or the **Run on all Active Unified CM Nodes** feature. The reasons for using each type of feature will typically be determined by the Unified CM version used by a cluster, or if clustering over the WAN has been deployed and geographically based call distribution is required.

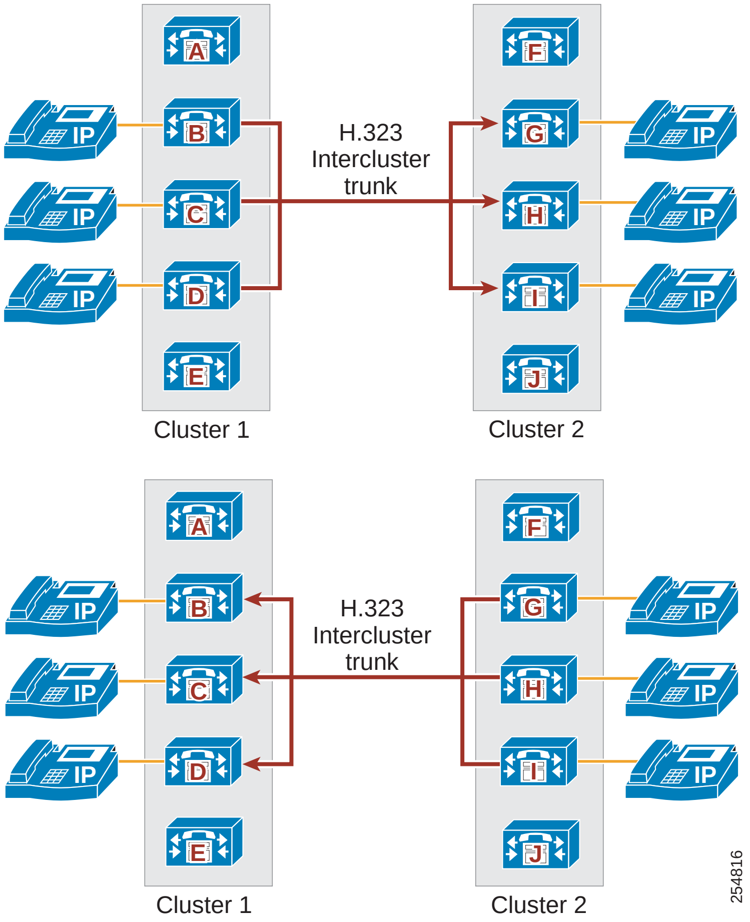
Using Standard Unified CM Groups with H.323 Non-Gatekeeper Intercluster Trunks

In this type of deployment, standard Unified CM Groups are used by H.323 non-gatekeeper intercluster trunks in each cluster. When defining this type of trunk with standard Unified CM Groups, you should define a maximum of three remote Unified CM servers in the destination cluster. The trunk will automatically load-balance calls across all servers defined as remote destination addresses. In the remote cluster, it is important to configure a corresponding intercluster trunk (non-gatekeeper controlled) that has the same Unified CM nodes in its Unified CM Group as those defined as remote destination Unified CM servers in the first cluster.

For example, if Cluster 1 has a trunk to Cluster 2, and Cluster 2 has a trunk to Cluster 1, the following configurations would be needed (see Figure 14-17):

- Cluster 1
  - Servers B, C, and D are configured as members of the Unified CM Group defined in the device pool associated with the non-gatekeeper controlled trunk to Cluster 2.
  - The non-gatekeeper controlled trunk has Cluster 2's remote servers G, H, and I configured as destinations.
- Cluster 2
  - Servers G, H, and I are configured as members of the Unified CM Group defined in the device pool associated with the non-gatekeeper controlled trunk to Cluster 1.
  - The non-gatekeeper controlled trunk has Cluster 1's remote servers B, C, and D configured as destinations.

Figure 14-17 H.323 Non-Gatekeeper Intercluster Trunks Using Standard Unified CM Groups



Using Run on All Active Unified Nodes with H.323 Non-Gatekeeper Intercluster Trunks

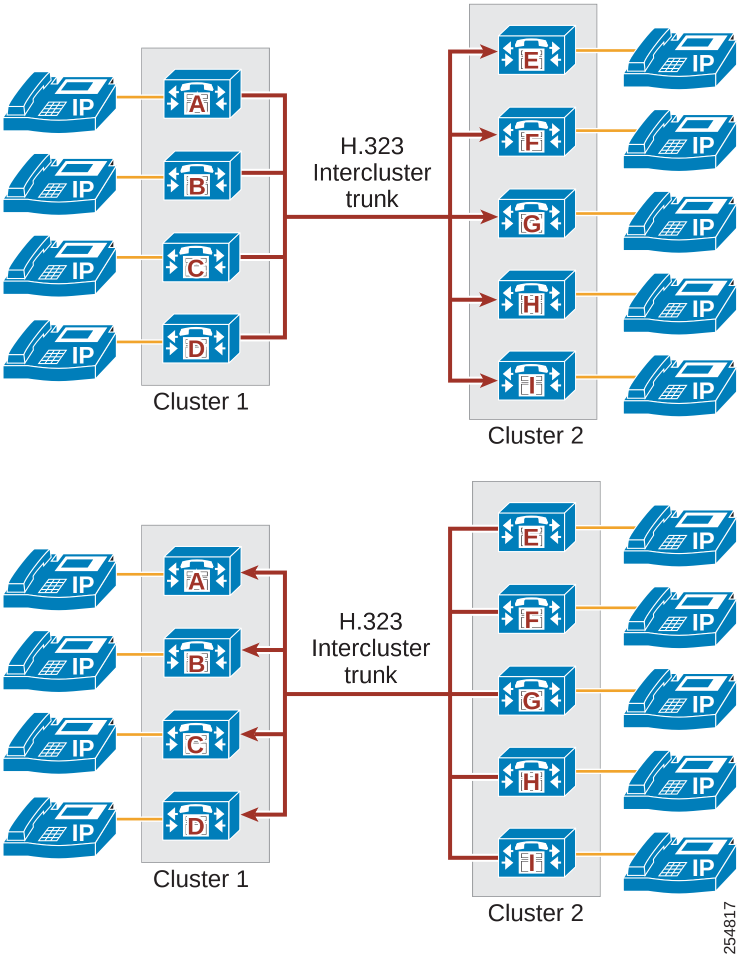
In this type of deployment, **Run on all Active Unified CM Nodes** is used by the H.323 non-gatekeeper intercluster trunks in each cluster. When defining this type of trunk, you may define up to 16 remote Unified CM servers in the destination cluster. (The number of remote servers that you need will depend on the number of active Unified CM nodes in the destination cluster.) The trunk will automatically load-balance calls across all defined remote destination Unified CM servers. In the remote cluster, it is important to configure a corresponding intercluster trunk (non-gatekeeper controlled) that has **Run on all Active Unified CM Nodes** configured, where these nodes are defined as the remote destination Unified CM servers in the first cluster.

For example, if Cluster 1 (four nodes) has a trunk to Cluster 2, and Cluster 2 (five nodes) has a trunk to Cluster 1, the following configurations would be needed (see Figure 14-18):

- Cluster 1 has four active Unified CM nodes (A, B, C, and D).
  - Enabling **Run on all active Unified CM Nodes** causes servers A, B, C, and D to have active H.323 trunk daemons associated with the non-gatekeeper controlled trunk to Cluster 2.
  - The non-gatekeeper controlled trunk has Cluster 2's remote servers E, F, G, H, and I configured as destinations.

- Cluster 2 has five active Unified CM nodes (E, F, G, H, and I).
  - Enabling Run on all active Unified CM Nodes causes servers E, F, G, H, and I to have active H.323 trunk daemons associated with the non-gatekeeper controlled trunk to Cluster 2.
  - The non-gatekeeper controlled trunk has Cluster 1's remote servers A, B, C, and D configured.

Figure 14-18 H.323 Non-Gatekeeper Intercluster Trunks Using Run on All Active Unified Nodes



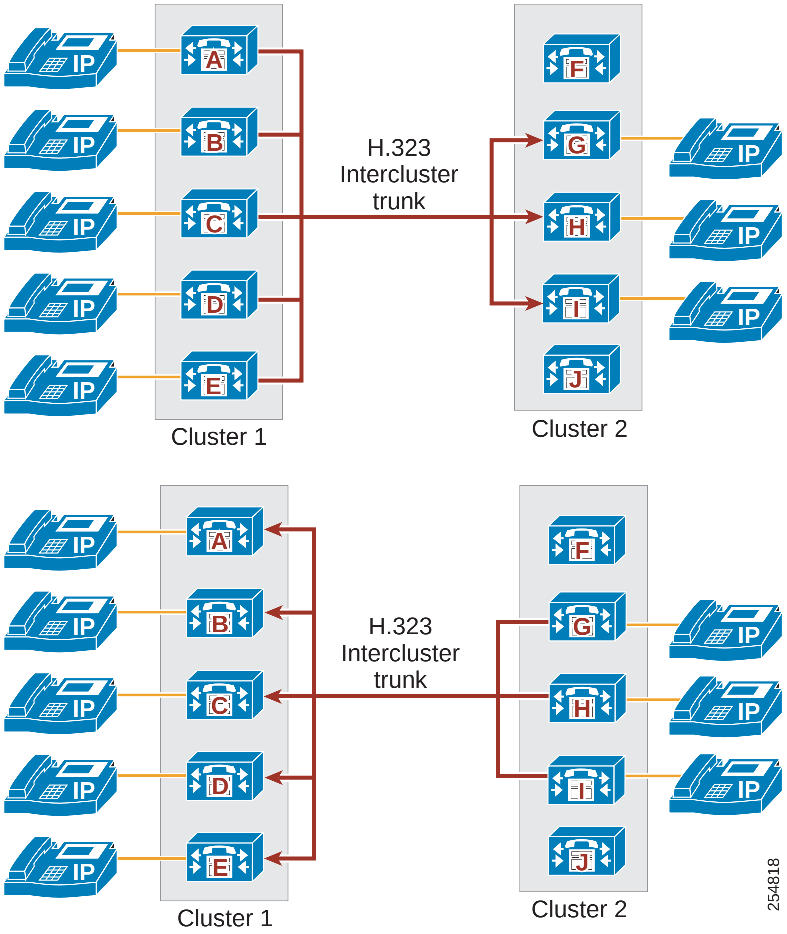
Using Standard Unified CM Groups and Run on All Active Unified CM Nodes with H.323 Non-Gatekeeper Intercluster Trunks

In this type of deployment, **Run on all Active Unified CM Nodes** is used by the H.323 non-gatekeeper intercluster trunk in one cluster, and standard Unified CM Groups are used by the H.323 non-gatekeeper intercluster trunk in the other cluster. When configuring these trunks, the number of remote Unified CM server destinations that you define should match the number of active Unified CM nodes for the corresponding trunk in the destination cluster. The trunk will automatically load-balance calls across all defined remote destination Unified CM servers. In the remote cluster, it is important to configure a corresponding intercluster trunk (non-gatekeeper controlled) that has Unified CM nodes with active H.323 daemons where these nodes are defined as remote destination Unified CM servers in the first cluster.

For example, if Cluster 1 has a trunk to Cluster 2, and Cluster 2 has a trunk to Cluster 1, the following configurations would be needed (see Figure 14-19):

- Cluster 1 has five active Unified CM nodes (A, B, C, D. and E).
  - Enabling **Run on all Active Unified CM Nodes** causes servers A, B, C, D, and E to have active H.323 trunk daemons associated with the non-gatekeeper controlled trunk to Cluster 2.
  - The non-gatekeeper controlled trunk has Cluster 2's remote servers G, H, and I configured as destinations.
- Cluster 2 has five active Unified CM nodes and uses an intercluster trunk with a Unified CM Group containing nodes G, H, and I.
  - Servers G, H, and I are configured as members of the Unified CM Group defined in the device pool associated with the non-gatekeeper controlled trunk to Cluster 1.
  - The non-gatekeeper controlled trunk has Cluster 1's remote servers A, B, C, D, and E configured as destinations.

Figure 14-19 H.323 Non-Gatekeeper Intercluster Trunks Using Standard Unified CM Groups and Run on All Active Unified CM Nodes



High Availability for Non-Gatekeeper Controlled Intercluster Trunks

High availability and redundancy for H.323 non-gatekeeper intercluster trunks can be provided by using multiple source Unified CM servers for originating calls and multiple destination IP addresses per trunk.

Multiple Source Unified CM Servers for Originating H.323 Non-Gatekeeper Intercluster Trunk Calls

- Using standard Unified CM Groups

  The nodes defined in the Unified CM Group associated with an individual trunk make up the set of servers that can place or receive calls over that trunk. Up to three nodes can be defined in a Unified CM Group, thus ensuring high availability of the trunk itself.

- Using **Run on all Active Unified CM Nodes**

  The **Run on all Active Unified CM Nodes** feature creates and enables an H.323 trunk instance on each call processing subscriber within the cluster, thus allowing these nodes to place or receive calls over that trunk.

- The Unified CM Route Local feature and its effect of subscriber selection for outbound H.323 non-gatekeeper intercluster trunk calls

  The Route Local feature in Unified CM is designed to reduce intra-cluster traffic. The feature operates as follows: When a device such as a phone is making an outbound call over H.323 intercluster trunk ICT 1, if an instance of H.323 ICT 1 is active on the same node as the one to which the phone is registered, then always use this co-located H.323 ICT 1 instance rather than internally route the call to another H.323 ICT 1 instance on another node within the cluster.

  The effect of the Route Local feature on node selection depends on whether Unified CM Groups or **Run on all Active Unified CM Nodes** is configured on the trunk. For trunks with **Run on all Active Unified CM Nodes** configured, the node to which the calling device is registered is used to make the outbound H.323 intercluster trunk call. When Unified CM Groups are used on the trunk, if the calling device is registered to one of the nodes in the trunk's Unified CM Group, then the Route Local rule applies. If the calling device is not registered to one of the nodes in the trunk's Unified CM Group, then Unified CM will randomly distribute the call over the nodes in the trunk's Unified CM Group.

In general, using **Run on all Active Unified CM Nodes** is recommended for H.323 intercluster trunks because call distribution across nodes is determined by the calling device and intra-cluster traffic is minimized.

Multiple Destination IP Addresses per H.323 Non-Gatekeeper Intercluster Trunks

A single H.323 non-gatekeeper intercluster trunk can be configured with up to 16 destination IP addresses. Unified CM uses round-robin distribution to the configured destination IP addresses when placing calls over an H.323 non-gatekeeper intercluster trunk.

Design Considerations When Using Run on All Active Unified CM Nodes

When using **Run on All Active Unified CM Nodes** in conjunction with multiple destination addresses, be aware that to accept inbound calls, the inbound source IP address received on the H.323 trunk must match with a configured destination IP address on the inbound trunk. Where clustering over the WAN designs are deployed and geographic call distribution and failover are required, use standard Unified CM Groups on multiple intercluster trunks (each with up to three destination IP addresses) in conjunction with route lists and route groups.

Load Balancing for H.323 Non-Gatekeeper Intercluster Trunks

When designing load balancing for H.323 non-gatekeeper intercluster trunks, consider both the node that sources the call and its destination. With H.323 non-gatekeeper intercluster trunks, the node that originates the call is determined by the Route Local rule, the number of nodes on which the outbound trunk is active, and whether a route list is used in conjunction with multiple outbound trunks. These considerations are discussed below.

Outbound Calls over a Single H.323 Non-Gatekeeper Intercluster Trunk

A single H.323 non-gatekeeper intercluster trunk can run on up to three Unified CM nodes in a Unified CM Group or on all active Unified CM nodes in the cluster. To select the source node for outbound calls, Unified CM applies the following decision processes:

> Where an instance of the trunk runs on all nodes, the Route Local rule applies and the node used for each outbound call is determined by the node on which the call arrives (for example, the node to which the calling phone is registered or the node on which the inbound trunk call arrives). Where Unified CM Groups are used, the route local rule still applies for those calling devices that are registered to the same node as the nodes in the trunk's Unified CM Group. For calling devices that are registered to other servers within the cluster, Unified CM will randomly distribute calls across the nodes in the trunk's Unified CM Group. Unified CM uses round-robin call distribution across the trunk's configured destination addresses. H.323 non-gatekeeper intercluster trunks may be configured with up to 16 destination IP addresses.

Outbound Calls over Multiple H.323 Non-Gatekeeper Intercluster Trunks

Because H.323 non-gatekeeper intercluster trunks can run on all active Unified CM nodes and have up to 16 destination addresses, you typically do not have to use multiple H.323 non-gatekeeper intercluster trunks to provide even call distribution between two Unified Communications clusters. Where multiple trunks are used with route lists and route groups, route lists should be enabled to run on all active Unified CM nodes. Multiple H.323 trunks are often used in conjunction with route lists to provide failover to the PSTN or to a group of Unified CM servers in a different site as part of a clustering over the WAN deployment. The selection of the Unified CM node used to initiate an outbound trunk call and the distribution of calls over the trunk's configured destination IP addresses is determined in the same way as described for single trunks. Where clustering over the WAN designs are deployed and geographic call distribution and failover are required, use multiple intercluster trunks (each with up to three destination IP addresses) with standard Unified CM Groups in conjunction with route lists and route groups.

Intercluster Trunk (Gatekeeper Controlled)

The intercluster gatekeeper controlled trunk can be used instead of the non-gatekeeper controlled trunk to interconnect a large number of Unified CM clusters. The advantages of using the gatekeeper controlled trunk are mainly the overall administration of the cluster and failover times. With non-gatekeeper controlled trunks, if a subscriber server in a cluster becomes unreachable, there will be a 5-second (default) timeout while the call is attempted. If an entire cluster is unreachable, the number of attempts before either call failure or rerouting over the PSTN will depend on the number of remote servers defined for the trunk and on the number of trunks in the route list or route group (if any). If there are many remote servers and many non-gatekeeper controlled trunks, the call delay can become excessive.

With a H.323 gatekeeper controlled trunk, you configure only one trunk that can then communicate by means of the gatekeeper with all other clusters registered to the gatekeeper. If a cluster or subscriber becomes unreachable, the gatekeeper automatically directs the call to another subscriber in the cluster or rejects the call if no other possibilities exist, thus allowing the call to be rerouted over the PSTN (if required) with little incurred delay. With a single Cisco gatekeeper, it is possible to have 100 clusters all registering a single trunk each, with all clusters being able to call each other. The gatekeeper controlled intercluster trunk should be used for communicating only with other Unified CMs because the use of this trunk with other H.323 devices might cause problems with supplementary services.

Note Gatekeeper controlled trunks do not support the **Run on All Active Unified CM Nodes** feature, and only standard Unified CM Groups are supported. Destination addresses are returned to Unified CM by the gatekeeper. Where gatekeeper controlled trunks are used in route lists, Cisco recommends enabling the **Run on All Active Unified CM Nodes** feature on the route list.

H.225 Trunk (Gatekeeper Controlled)

The H.225 gatekeeper controlled trunk is essentially the same as the intercluster gatekeeper controlled trunk except that it has the capability of working with Unified CM clusters as well as other H.323 devices such as gateways, conferencing systems, and clients. This capability is achieved through a discovery mechanism on a call-by-call basis. (See H.323 Operation in Unified CM, page 14-53, for details of this discovery process.)

Note Gatekeeper controlled trunks do not support the **Run on All Active Unified CM Nodes** feature, and only standard Unified CM Groups are supported. Destination addresses are returned to Unified CM by the gatekeeper. Where gatekeeper controlled trunks are used in route lists, Cisco recommends enabling the **Run on All Active Unified CM Nodes** feature on the route list.

High Availability for Gatekeeper Controlled Trunks

Redundancy can be achieved in several ways, depending on the requirements of the design. The simplest method is to configure a gatekeeper controlled trunk and assign up to three subscribers in the Unified CM Group associated with the device pool assigned to that trunk. This configuration will cause all servers to register with the same gatekeeper in the same zone with the same technology prefix. However, the H.323 trunk name that is used for the h323_id will have a suffix of "_*n*" where *n* is the node number in the cluster. This ID is automatically generated and cannot be changed. You configure a single trunk, but the gatekeeper registers multiple trunks, one from each subscriber in the Unified CM Group.

If you have additional redundancy requirements, it is possible to configure another gatekeeper controlled trunk with a different name and different subscribers in the Unified CM Group, but with all the other parameters identical to the first trunk. This second trunk will cause additional subscribers to register with the gatekeeper.

Cisco recommends assigning device pools that contain a Unified CM Group consisting of the two servers that make up the standard subscriber pair. (See Call Processing Subscriber Redundancy, page 8-18, for more information on subscriber redundancy.) For complete redundancy in each full cluster, four trunks would be needed, using four different device pools and resulting in eight subscribers registering with the gatekeeper. (The same result could be achieved with three trunks and larger Unified CM Groups.)

During registration, several parameters are passed between Unified CM and the gatekeeper. Unified CM uses an ephemeral User Datagram Protocol (UDP) port for gatekeeper Registration Admission Status (RAS) messages. This port would normally be UDP 1719. However, Unified CM must be able to identify precisely which H.323 daemon is the originator of a RAS message from a particular server; therefore it uses a range of UDP ports and assigns them dynamically.

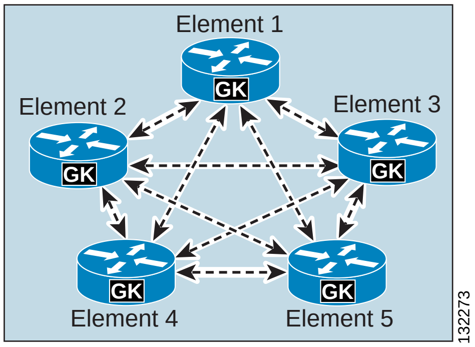
During the registration process, a trunk registers the following information for the other subscribers in its Unified CM Group:

- H.225 call signaling port
- h323_id
- CanMapAlias support
- Technology prefix
- H.225 call signaling address

If the recommended clustered gatekeepers are used, the gatekeeper will return a list of alternate gatekeeper addresses that may be used if the primary gatekeeper fails or does not have sufficient available resources.

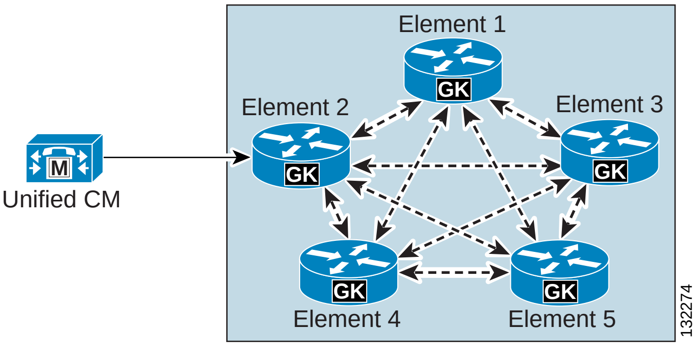
Figure 14-20 shows a cluster of gatekeepers that use Gatekeeper Update Protocol (GUP) to communicate. (See the chapter on Call Processing, page 8-1, for more information on gatekeepers.)

Figure 14-20 Gatekeeper Cluster



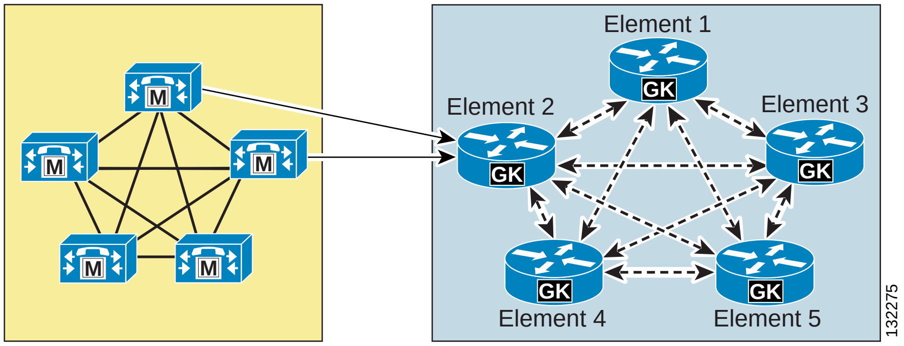
If an H.323 trunk has only a single subscriber in its Unified CM Group, there will be only one connection between the configured gatekeeper in Unified CM and the gatekeeper cluster, as illustrated in Figure 14-21.

Figure 14-21 H.323 Trunk with a Single Unified CM Subscriber



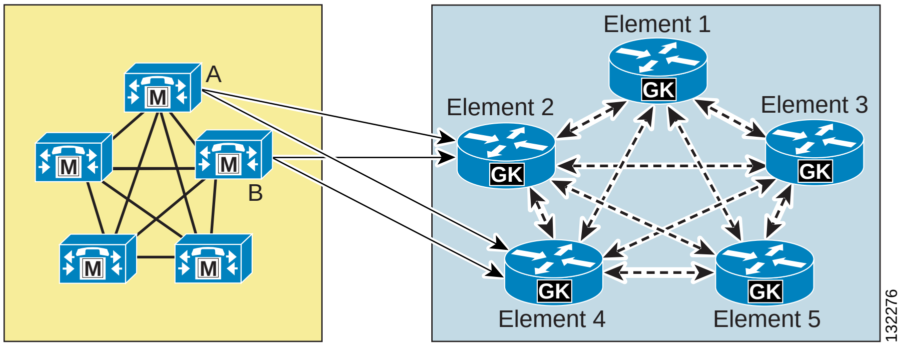
If there are multiple subscribers in the Unified CM Group associated with the trunk, additional connections will be established between the Unified CM cluster and the gatekeeper cluster, as illustrated in Figure 14-22.

Figure 14-22 H.323 Trunk with Multiple Unified CM Subscribers



This approach provides redundancy for subscriber failures as well as gatekeeper failures after registration because the alternate gatekeeper is communicated when the trunk registers. This approach does not, however, provide redundancy if the configured gatekeeper is unavailable at initial registration or following a reset because the list of alternate gatekeepers is dynamic and not stored in the database. To provide an additional level of redundancy as well as load balancing, an additional gatekeeper from the gatekeeper cluster is configured in Unified CM. For example, if the original trunk is registered with Element 2, the additional gatekeeper could be configured as Element 4, as illustrated in Figure 14-23.

Figure 14-23 Additional Gatekeeper Configured for Load Balancing and Additional Redundancy



The Unified CM configuration for the example in Figure 14-23 would contain the following components:

- Two gatekeepers for Element 2 and Element 4
- Two H.323 trunks defined with a Unified CM Group containing subscriber servers A and B

Using this approach, the Unified CM cluster will still be able to register when either Element 2 or Element 4 is not reachable during initial registration (that is, during power-up or trunk reset).

Load balancing of calls inbound to the Unified CM cluster is done automatic by default because the gatekeeper randomly selects one of the subscribers registered within the zone. If this is not the desired behavior, you can use the **gw-priority** configuration command in the gatekeeper to modify this default behavior, as illustrated in Example 14-3.

Example 14-3 Using the gw-priority Command to Direct Calls to a Particular Trunk

```
gatekeeper
 zone local SJC cisco.com 10.0.1.10
 zone prefix SJC 1408....... gw-priority 10 sjc-trunk_2
 zone prefix SJC 1408....... gw-priority 9 sjc-trunk_3
 zone prefix SJC 1408....... gw-default-priority 0
 gw-type-prefix 1#* default-technology
 arq reject-unknown-prefix
 no shutdown
 endpoint ttl 60
```

In Example 14-3, the H.323 trunk was configured as sjc-trunk in Unified CM, and the “_2” and “_3” suffixes are appended automatically by the Unified CM subscribers to indicate which node number they are in the cluster. Therefore, this example uses node 2 as the first choice, which should be the highest-priority Unified CM in the Unified CM Group for this trunk. Node 3 is the second choice in this case.

The use of **gw-default-priority 0** is optional. It was used in this example to disable the use of any other trunk that might accidentally be configured to register in this zone.

Load Balancing Outbound Calls over H.323 Gatekeeper Controlled Trunks

With Unified CM H.323 gatekeeper controlled trunks, the node from which the call originates is determined by the Route Local rule, the number of nodes on which the outbound trunk is active, and whether a route list is used in conjunction with multiple outbound trunks. These considerations are discussed below.

Outbound Call Load Balancing When Deploying a Single H.323 Gatekeeper Controlled Trunk

For the initiation of outbound calls over a single H.323 gatekeeper controlled trunk, the route local rule applies and the following factors within a Unified CM cluster determine which server is selected:

- Which Unified CM servers have an active H.323 daemon for the selected trunk
- Whether the phone originating the call is registered to a Unified CM server with an active H.323 daemon for the selected trunk

For a single H.323 gatekeeper controlled trunk, the Route Local process of server selection for outgoing calls operates as follows:

- If there is an active H.323 daemon for the selected trunk on the Unified CM server to which the phone or device originating the call is registered (that is, if the server is one of those listed in the trunk's Unified CM Group), then use this Unified CM server to originate the H.323 call.
- If there is no active H.323 daemon for the selected trunk on the Unified CM server to which the phone or device originating the call is registered, then select a server on a round-robin basis from the Unified CM Group of the selected trunk.

Outbound Call Load Balancing When Deploying Route Lists in Conjunction with H.323 Gatekeeper Controlled Trunks

In configurations where route lists are employed to select a trunk for outbound calls, enable **Run on all Active Unified CM Nodes** for all route lists. Running route lists on all nodes improves outbound call distribution by using the Route Local rule to avoid unnecessary intra-cluster traffic. For route lists, the Route Local rule operates as follows:

> For outbound calls that use route lists (and associated route groups and trunks), when a call (from a registered phone or inbound trunk) arrives at the node with the route list instance associated with the outbound trunk call, Unified CM checks to see if an instance of the selected outbound trunk call exists on the same node as the route list. If so, Unified CM will use this node to establish the outbound trunk call.

If the route list has **Run on all Active Unified CM Nodes** enabled: For gatekeeper controlled trunks using Unified CM Groups, Unified CM will apply the route local rule if an instance of the selected outbound trunk exists on the same node on which the inbound call arrived. If an instance of the trunk does not exist on this node, then Unified CM will forward the call (within the cluster) to a node where the trunk is active.

If the route list does not have **Run on all Active Unified CM Nodes** enabled, the route list will be active on one node within the cluster (the primary node in the route list's Unified CM Group) and the Route Local rule will be applied on this node.

H.323 Outbound Fast Start Call Connections

Calls that are placed from IP phones over large WAN topologies with long delays can experience voice clipping when the called party goes off-hook to answer the call. When H.323 trunks or gateways are separated from the Unified CM server, significant delays can occur because of the many H.245 messages that are exchanged when a call is set up.

With the Fast Start feature, information that is required to complete a media connection between two parties gets exchanged during the H.225 portion of call setup, and this exchange eliminates the need for H.245 messages. The connection experiences one round-trip WAN delay during call setup, and the calling party does not experience voice clipping when the called party answers the call.

Unified CM uses media termination points (MTPs) for making an H.323 outbound Fast Start call. Unified CM starts an outbound Fast Start call by allocating an MTP and opening the receive channel. Next, the H.323 Fast Connect procedure sends the SETUP message with a Fast Start element to the called endpoint. The Fast Start element includes information about the receiving channel for the MTP.

By default, H.323 Fast Start is disabled. To enable H.323 Fast Start, check the **MTP Required** and **Enable Outbound FastStart** checkboxes on the H.323 trunk, and select the desired Codec For Outbound Fast Start. Also note that Inbound Fast Start is enabled separately with check box **Enable Inbound FastStart**. (Inbound Fast Start does not require an MTP or a codec selection.)

Note When H.323 Fast Start is enabled, an MTP is assigned for every outbound H.323 trunk call. MTPs used for H.323 Fast Start support a single voice codec only, and therefore video and encrypted calls are not supported. H.323 Fast Start is disabled by default on H.323 trunks, and MTPs are not required for outbound or inbound calls. As a general rule, this default H.323 (Slow Start) trunk configuration is preferred so that voice, video, and encrypted calls are supported over H.323 trunk connections.

H.323 Trunks with Media Termination Points

Media termination points (MTPs) are generally not required for normal operation of the H.323 trunk. They are, however, required for communication with devices that are H.323 Version 1, that do not support the Empty Capabilities Set (ECS) for supplementary services, or that require H.323 Fast Start.

To test whether or not an MTP is required, use the following simple procedure:

1. Place a call from a phone through the H.323 trunk to the other device. This call should work normally.
2. Place the call on hold, then resume it. If the call drops, then it is highly likely that an MTP is required to ensure interoperability between Unified CM and the other device.

DTMF Transport

The H.323 trunk supports DTMF signaling for both out-of-band DTMF using H.245 and in-band DTMF using RTP Named Telephone Events (RFC 2833). There are no configuration options. An MTP may be allocated dynamically to convert between out-of-band DTMF relay to in-band DTMF relay, if required. For an explanation of when the H.323 trunk uses which method and when MTPs are required, refer to the chapter on Media Resources, page 17-1.

H.323 Trunk Transport Protocols

H.323 trunks use TCP for H.225 call control and H.245 media control signaling, and UDP for gatekeeper H.225 Registration Admission Status (RAS) signaling.

Secure H.323 Trunks

Securing H.323 trunks involves two processes: configuring the trunk to encrypt media and configuring the trunk to encrypt signaling.

Media Encryption

Media encryption can be configured on H.323 trunks by checking the trunk's **SRTP allowed** check box. It is important to understand that checking the **SRTP allowed** checkbox will cause the media for calls to be encrypted but the trunk signaling will not be encrypted, therefore the session keys used to establish the secure media stream will be sent in the clear. It is therefore important to ensure that signaling between Unified CM and its destination H.323 trunk device is also encrypted so that keys and other security-related information do not get exposed during call negotiations.

Signaling Encryption

H.323 trunks use IPSec for signaling encryption. You may configure IPSec in the network infrastructure, or you may configure IPSec between Cisco Unified Communications Manager (Unified CM) and the remote gateway or trunk. If you implement one method to set up IPSec, you do not need to implement the other method. Using IPSec on Unified CM servers can incur a significant impact on server performance, therefore Cisco recommends that you provision IPSec in the network infrastructure rather than in Unified CM itself.

If the system can establish a secure media or signaling path and if the end devices support SRTP, the system uses an SRTP connection. If the system cannot establish a secure media or signaling path or if at least one device does not support SRTP, the system uses an RTP connection. SRTP-to-RTP fallback (and vice versa) may occur for transfers from a secure device to a non-secure device, conferencing, transcoding, music on hold, and so on.

For SRTP-configured devices, Unified CM classifies a call as encrypted if the **SRTP Allowed** check box is checked for the device and if the SRTP capabilities for the devices are successfully negotiated for the call. If the preceding criteria are not met, Unified CM classifies the call as non-secure. If the device is connected to a phone that can display security icons, the phone displays the lock icon when the call is encrypted.

MTPs that are statically assigned to an H.323 trunk using the **MTP Required** checkbox do not support SRTP because they do not support the pass-through codec. To ensure that SRTP is supported for all calls, do not configure the H.323 trunk for H.323 Outbound Fast Start (that is, do not select **MTP Required**). SRTP is supported for Inbound Fast Start. (Inbound Fast Start does not require an MTP or a codec selection.)

H.323 Operation in Unified CM

This section provides information on how the H.323 protocol is used and implemented in Unified CM, and it explains how and why certain features work the way they do.

The most important point to understand is which subscribers run the call signaling daemons. These daemons are pieces of code that make and receive H.323 calls. They are usually referred to as H.323 or H.225 daemons (H.323Ds or H.225Ds). H.225 is part of the H.323 protocol and is mainly responsible for call control. H.245 is the other major component of H.323 that is responsible for the media control of a call.

For the majority of H.323 devices, the subscribers listed in the Unified CM Group for a particular H.323 device determine which subscribers run the daemons and when. For H.323 non-gatekeeper controlled intercluster trunks, standard Unified CM Groups can be used or the **Run on All Active Unified CM Nodes** can be enabled, in which case the daemon will run on all active nodes.

For devices using Unified CM Groups, it is important to know which nodes will run H.225 daemons because calls sent to an incorrect subscriber might be rejected. For example, this situation would occur if a Cisco IOS H.323 gateway is configured with dial peers that send calls to subscriber C in a Unified CM cluster but the Unified CM Group for that gateway has only subscribers A and B in its list. In such a case, the call will fail or be handled by an H.323 trunk daemon if one happens to be configured on the subscriber.

The following scenarios describe where and when H.225Ds are created on subscribers:

- H.323 client

  The H.225D is active on only the highest-priority subscriber available in the Unified CM Group associated with the H.323 client.

  If the H.323 client is gatekeeper controlled, the RasAggregator device registers from only the highest-priority subscriber available in the Unified CM Group associated with the gatekeeper controlled H.323 client.

  The RasAggregator is a special device that registers in gatekeeper zones for the purpose of providing two specific features:

  – If H.323 clients use DHCP, they cannot be used with a Unified CM using DNS unless they support Dynamic DNS. With the RasAggregator, Unified CM can obtain the IP address of a specific H.323 client that is registered with the gatekeeper whenever a call is placed. The gatekeeper registration is done using standard RAS ARQ messages that contain the E.164 address of the H.323 client. The gatekeeper resolves the E.164 address and provides the IP address back to Unified CM in an ACF message.

  - – The RasAggregator also ensures that all calls by the H.323 clients are made through Unified CM and not directly between the clients themselves, thus ensuring that dialing rules and codec restrictions are enforced.

- H.323 gateway

  The H.225D is active on all subscribers in the Unified CM Group associated with the H.323 gateway.

- H.323 gatekeeper controlled trunks

  The H.225D is active on all subscribers in the Unified CM Group associated with the H.323 trunk. A RAS daemon registers the trunk with the gatekeeper from all subscribers in the associated Unified CM Group.

- H.323 non-gatekeeper controlled trunks using Unified CM Groups

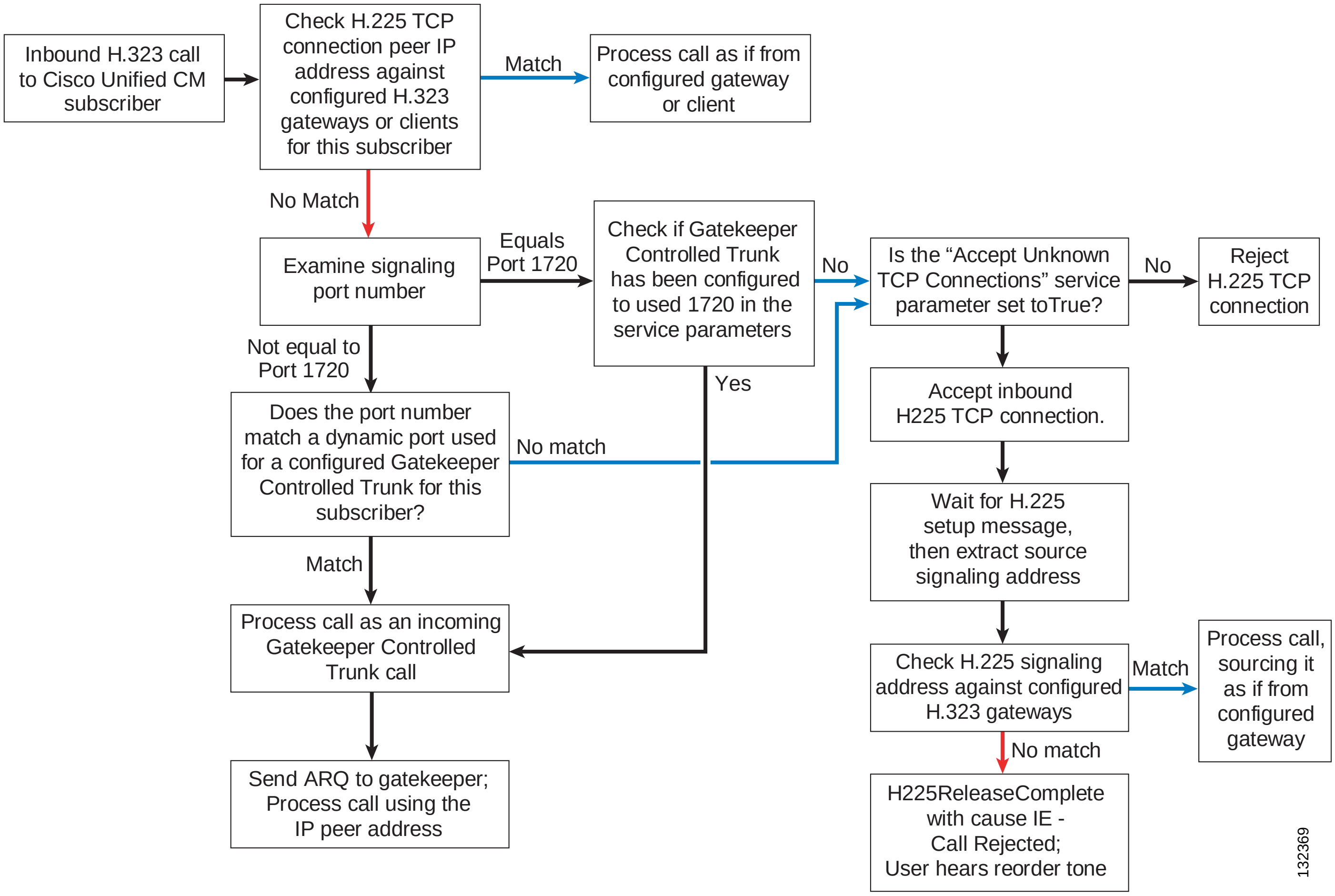
  The H.225D is active on all subscribers in the Unified CM Group associated with the H.323 trunk.

- H.323 non-gatekeeper controlled trunks using **Run on all Active Unified CM Nodes**

  The H.225D is active on all active Unified CM subscribers in the cluster.

When an incoming H.323 call is made to a subscriber in a Unified CM cluster, various decisions are made to determine if the call is accepted or rejected and which H.225D will receive the call if it is accepted. Figure 14-24 shows how this process works.

Figure 14-24 Process for Determining if an H.323 Call is Accepted or Rejected


Inbound H.323 call to Cisco Unified CM subscriber
Check H.225 TCP connection peer IP address against configured H.323 gateways or clients for this subscriber
Match
Process call as if from configured gateway or client
No Match
Examine signaling port number
Equals Port 1720
Check if Gatekeeper Controlled Trunk has been configured to used 1720 in the service parameters
No
Is the "Accept Unknown TCP Connections" service parameter set toTrue?
No
Reject H.225 TCP connection
Not equal to Port 1720
Yes
Does the port number match a dynamic port used for a configured Gatekeeper Controlled Trunk for this subscriber?
No match
Accept inbound H225 TCP connection.
Wait for H.225 setup message, then extract source signaling address
Match
Process call as an incoming Gatekeeper Controlled Trunk call
Check H.225 signaling address against configured H.323 gateways
Match
Process call, sourcing it as if from configured gateway
No match
Send ARQ to gatekeeper; Process call using the IP peer address
H225ReleaseComplete with cause IE - Call Rejected; User hears reorder tone
132369

Unified CM H.323 protocol includes the following additional features:

- Protocol Auto Detect

  This feature provides the ability to determine, on a call-by-call basis, if the calling device is from Cisco Unified CM. Whenever a call is received, Unified CM looks for an H.225 User-to-User Information Element (UUIE) that indicates if the other end is another Unified CM. If it is, it will always use the Intercluster Trunk Protocol. If no UUIE is found, it will use the configured protocol for that device. This feature enables an H.225 gatekeeper controlled trunk to switch between Intercluster Trunk Protocol and H.225 on a call-by-call basis, allowing a mixture of Unified CM clusters and other H.323 devices to use the gatekeeper. Intercluster Trunk Protocol is the same as H.225 except for several differences that enable specific features to work correctly between Unified CM clusters.

- Tunneled QSIG or H.323 Annex M1 (ISO and ECMA variants supported per trunk)

  This feature can be enabled on all H.323 trunks. It allows specific H.323 Annex M1 features to be implemented between Unified CM clusters and other verified systems that also support H.323 Annex M1. These features include:

  - Path replacement
  - Message waiting indication (MWI)
  - Callback

- Alternate Endpoints

  When registering with a gatekeeper that supports this feature, such as a Cisco Multimedia Conference Manager (MCM) Gatekeeper, Unified CM can inform the gatekeeper of alternate destinations for calls to the H.323 trunk. These alternate endpoints or destinations are sent to the calling device by the gatekeeper when this H.323 trunk is called. They are the other subscribers listed in the Unified CM Group associated with the H.323 trunk that registers with the gatekeeper.

- Alternate Gatekeeper

  When an H.323 trunk registers with a gatekeeper that supports this feature (for example, a Cisco gatekeeper cluster), Unified CM is dynamically informed about other gatekeepers that can process registrations, call admission requests, and other RAS functions in the event that this gatekeeper fails or exhausts its own resources.

- CanMapAlias

  When an H.323 trunk sends an admission request (ARQ) to the gatekeeper, it might receive a different E.164 number in the admission confirmation message (ACF), indicating that the original called number should be replaced with this new one. This feature requires a route server using Gatekeeper Transaction Message Protocol (GKTMP) to communicate with Cisco gatekeepers.

  Note CanMapAlias is supported for the called number only.

- Bandwidth Requests

  H.323 trunks can update the gatekeeper with bandwidth information to indicate a change in the requested bandwidth allocated to a specific call. This feature is disabled by default and is controlled by setting the Unified CM service parameter **BRQ Enabled** to **True**, under the H.323 section. This feature is especially important when video is used on an H.323 trunk because the original bandwidth request is for the maximum amount allowed. Enabling this feature ensures that call admission control uses the actual bandwidth negotiated during call setup.

Other Design Considerations for H.323 Trunks

Unified CM SIP trunks now offer a greater set of features in comparison with H.323 intercluster trunk, making SIP the protocol of choice for intercluster trunk connections, although H.323 Annex M1 may still be preferred for intercluster trunk connections to Unified CM clusters using earlier software versions. For more information on deploying intercluster trunks in multi-cluster and clustering over the WAN environments, see Design Considerations for SIP Trunks, page 14-28.

General SIP and H.323 Trunk Design Considerations

This section covers the following general design considerations:

- Deterministic Outbound Call Load Balancing over Unified CM Trunks, page 14-57
- Codec Selection Over IP Trunks, page 14-58
- Other MTP Uses, page 14-62

Deterministic Outbound Call Load Balancing over Unified CM Trunks

In the majority of cases, using **Run on all Active Unified CM Nodes** or assigning Unified CM Groups to devices is sufficient to handle the call distribution of outbound calls over trunks from call processing subscribers. Due to the Route Local rule, trunk calls might appear to originate randomly from call processing subscribers, but the trade-off for this random call origination is reduced call processing and reduced Intra-Cluster Communication Signaling (ICCS) traffic within the cluster.

Deterministic load balancing of outbound IP trunk calls across call processing servers is possible but can be counter-productive because the advantages gained from predictable call origination within the cluster can be outweighed by the increase in ICCS traffic created by calls from phones registered to one subscriber extending their communication to another server within the cluster to originate the outgoing IP trunk call.

Predictable and deterministic subscriber-based load balancing of outgoing IP trunk calls can be achieved as follows:

- To deterministically load-balance outbound trunk calls across a subset of the call processing servers in the cluster, define multiple trunks and assign only a single subscriber to the Unified CM Group of each trunk. Place these trunks into a route group and use circular call distribution.

  For example, to spread outbound trunk calls across four subscribers in the cluster, perform the following tasks:

  – Configure four H.323 trunks or four SIP trunks with individual Unified CM Groups, all contained within a route group with circular call distribution.

  – Define Unified CM Groups as follows:

    Group A: Subscriber A

    Group B: Subscriber B

    Group C: Subscriber C

    Group D: Subscriber D

  With no backup subscribers defined, if the primary subscriber for the specified trunk fails, Unified CM will re-route outgoing calls to the next trunk in the route group.

- To spread outbound trunk calls across all eight subscribers in a cluster, perform the following tasks:
  - Configure eight H.323 trunks or eight SIP trunks with individual Unified CM Groups, each containing only one subscriber and all contained within a circular route group.
  - Define Unified CM Groups as follows:

    Group A: Subscriber A

    Group B: Subscriber B

    Group C: Subscriber C

    Group D: Subscriber D

    Group E: Subscriber E

    Group F: Subscriber F

    Group G: Subscriber G

    Group H: Subscriber H

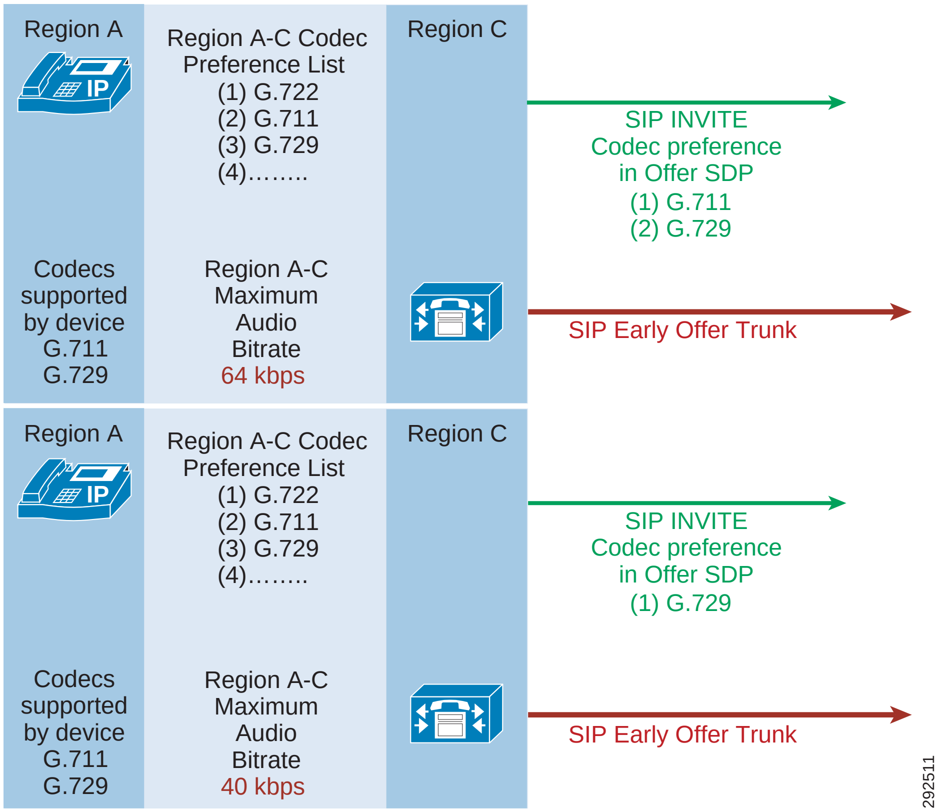
Codec Selection Over IP Trunks

Before media can be established between communicating entities, both the entities must agree on the codec(s) that they want to use. This codec (or codecs if both audio and video are involved) is derived from the intersection of codecs supported by communicating entities involved and the configured policy in Unified CM. Policy in Unified CM is configured by region settings.

The region settings in Unified CM 9.*x* provide for configurable audio codec preference lists. In addition to the default Lossy and Low Loss audio codec preference lists that can be selected via a region's Link Loss Type, multiple custom audio codec preference lists can also be configured. Audio codec preference lists can be used for codec selection for calls within a region and between regions. The **Maximum Audio Bit rate** is still applied for calls within a region and between regions; but rather that using the highest audio quality codec (as in earlier Unified CM releases) based on the maximum bit rate setting, the codec selection is made based on the codec order in the audio codec preference list and the codecs that the endpoints support. (See Figure 14-25.)

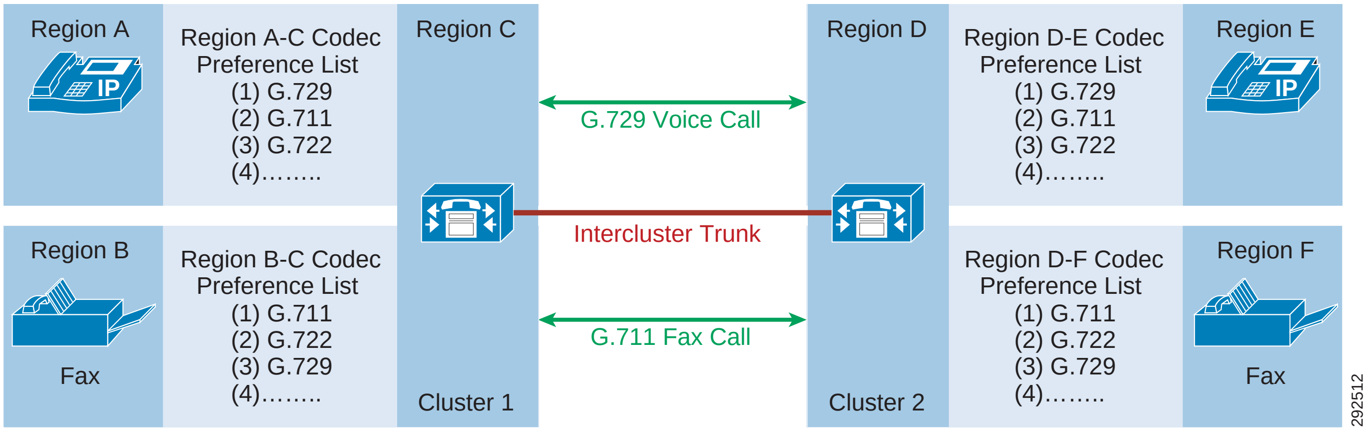
The Audio Codec Preference List is a list of all the codec types supported by Unified CM. The preference order of this list of codecs can be modified and saved as a custom preference list. (Note that codecs cannot be removed from the Audio Codec Preference List). The list of codecs used for codec negotiation during call setup is the subset of codecs supported by the device and those in the codec preference list, limited by the maximum audio bit rate for the region or region pair.

Figure 14-25 Examples of How Codecs Are Selected for Codec Negotiation During Call Setup



For calls between two Unified CM clusters via SIP or H.323 intercluster trunks, audio codec preference lists allow the codec to be selected for a call based upon the codec preferences of the calling and called devices. By grouping devices in each cluster into regions based on their codec preferences, a single intercluster trunk can be used to support multiple calls, each using a specific codec. (See Figure 14-26.)

Figure 14-26 Audio Codec Preference Lists for Voice and Fax Calls Between Two Unified CM Clusters



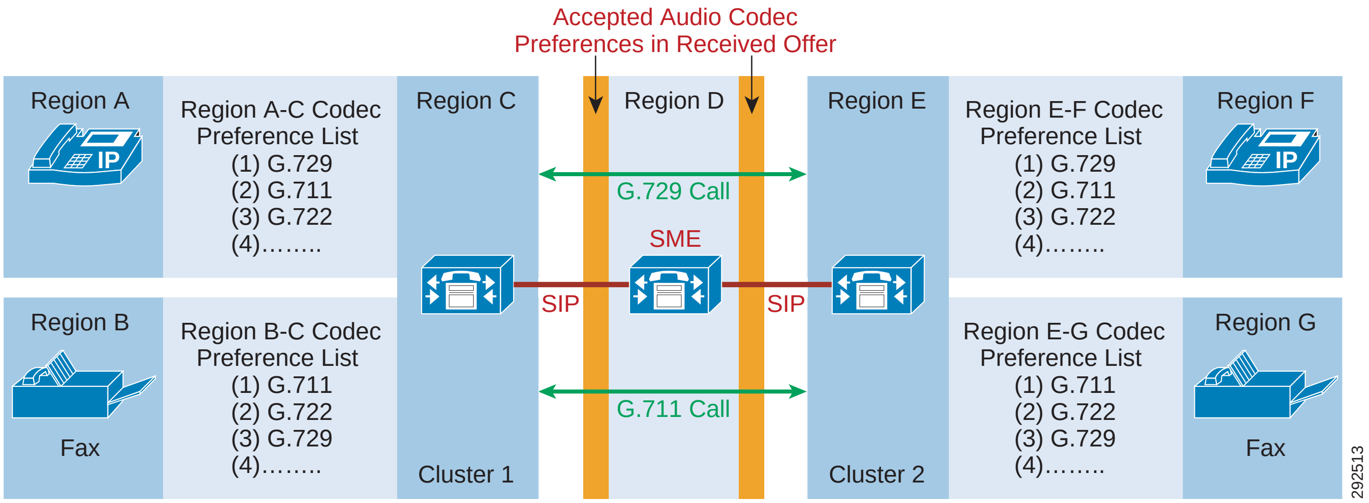
Note Equivalent audio codec preference lists for each device type region should be configured in each cluster to ensure that a common codec is selected for each device type, irrespective of call direction or trunk configuration. If the audio codec preference lists in each cluster are not equivalent, the codecs used per call may vary based on call direction and trunk configuration. (Ordinarily, the codec preference order is not honored by the cluster receiving the codec preference list.)

Note Avoid using SIP trunks configured for Early Offer using **MTP Required** or H.323 trunks with Fast Start enabled. These trunk configurations insert an MTP for outbound calls, which limits calls to a single audio codec only, thereby overriding codec preference and selection. In these cases, the maximum audio bit rate setting must also be configured appropriately to allow this codec to be used.

Accept Audio Codec Preferences in Received Offer

In deployments where calls can pass through more than one Unified CM cluster (for example, SME deployments), the inter-region audio codec preference list of the intermediary Unified CM cluster can override the preferred codec selection between the calling and called device. To ensure that the endpoints' codec preferences are honored as calls pass through SME, enable the SIP Profile feature **Accept Audio Codec Preferences in Received Offer** on all SME SIP trunks. (See Figure 14-27.)

Figure 14-27 SME Deployment Using "Accept Audio Codec Preferences in Received Offer" on SIP Trunks



Note The **Accept Audio Codec Preferences in Received Offe**r feature is available only on SIP trunks (a SIP Profile feature). This feature does not offer consistent results if used in an SME deployment where the SME cluster uses a combination of SIP, H.323, and/or MGCP trunks. Therefore, the **Accept Audio Codec Preferences in Received Offe**r feature should be used when the SME cluster is deployed using only SIP trunks.

Note For Unified CM 8.5 and later releases, Cisco recommends the use of SIP trunks instead of H.323 and MGCP trunks because SIP offers the richest feature set. (See Table 14-2 for a comparison of SIP and H.323 intercluster trunks. For a comparison of features for SIP, H.323, and MGCP trunks, see the *Cisco Unified Communications Manager Session Management Edition Deployment Guide*, available at http://www.cisco.com/en/US/products/ps10661/products_implementation_design_guides_list.html.)

Cisco Unified CM and Cisco Unified Border Element SIP Trunk Codec Preference

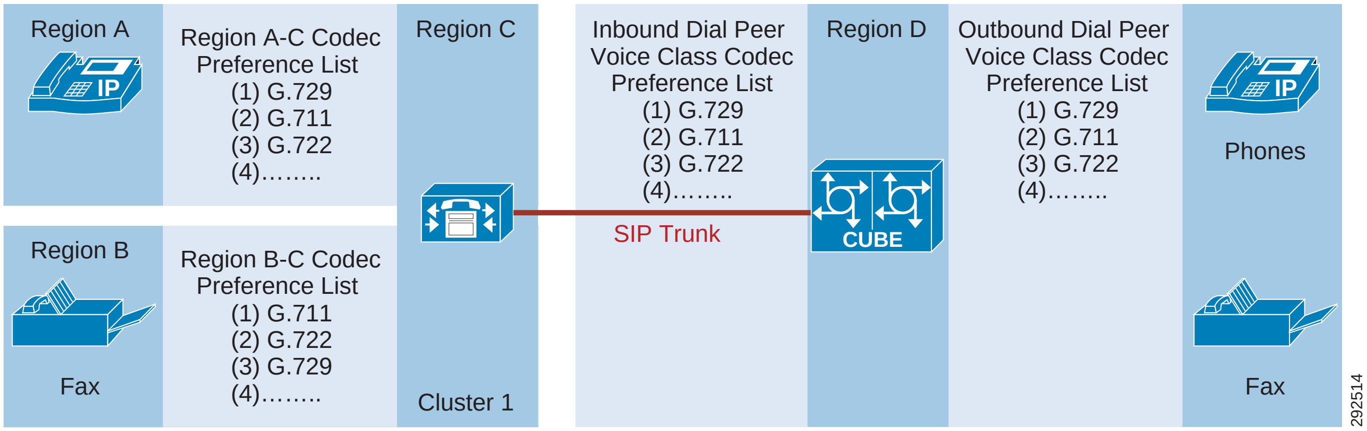
Unified CM audio codec preference lists can be used in Unified Communications deployments with Cisco Unified Border Element (CUBE) to simplify Unified CM-to-CUBE SIP trunk configuration. For example, instead of using dedicated SIP trunks to CUBE for voice and fax calls, a single Unified CM SIP trunk can be used where the codec preference for each device type is honored as calls pass through CUBE.

In Figure 14-28, the Voice Class Codec Preference lists defined on CUBE's inbound and outbound dial peers do not change the preference of the listed codecs in the received Offer. CUBE does codec filtering on the received Offer, both on the inbound and outbound dial-peer, and passes across the common codecs in the same preference order as received in the inbound Offer to the peer leg.

If codecs, in addition to those received in an Offer, are defined in the voice class codec list, then these codecs will be appended to those received in the ordered list and sent out in the outbound Offer.

Thus, a single inbound and outbound dial-peer can be configured on CUBE for all device types. Cisco recommends using the same voice class codec preference list for both the inbound and outbound dial-peer, with that list containing the codecs that you want to negotiate with the service provider. As mentioned above, the order of the codecs will be dictated first by the order received in the inbound Offer and then by the order defined in the voice class codec preference list.

Figure 14-28 Cisco Unified CM and Cisco Unified Border Element SIP Trunk Codec Preference



Other MTP Uses

MTPs are very useful for terminating media streams from other devices that make calls over trunks and for re-originating the media streams with the same voice payload; however, in such cases the IP address is changed to that of the MTP. With this fact in mind, you can utilize MTPs in the following scenarios:

- If the phones, gateways, and other devices within your enterprise all use RFC 1918 private addresses, you might still want to connect to other systems on a public network without using Network Address Translation (NAT) for all your voice and video devices. If the Unified CM subscriber that communicates to the public network is using a public IP address, the signaling will be routed. If all MTPs are also using public addresses, the media from the devices with RFC 1918 addresses will be terminated on the MTP and then originated again, but this time with a public address that is routable on the public network. This approach allows tens of thousands of devices with RFC 1918 addresses to communicate with the public network. This same method can be used to conceal the real IP addresses of devices in an enterprise network when communicating with other enterprises or service providers.
- Trust boundaries can be established to traverse firewalls or to allow access through an access control list (ACL). Normally, for media to traverse a firewall, you could either use an Application Layer Gateway (ALG) or fix-up to provide access dynamically for the media streams or you could allocate a wide range of addresses and ports for use by all voice devices that need to communicate across the firewall. All calls that use a trunk and traverse a firewall or ACL will have media that is sourced from the MTP(s), which may use either a single IP address or a small range of IP addresses.

With both of these methods, if the **MTP Required** box is checked, the default behavior is to allow calls on SIP and H.323 trunks even if MTP resources are unavailable or exhausted. This default behavior might result in no voice path for the call, but the behavior can be changed by setting the Unified CM service parameter **Fail Call if MTP allocation fails** under the SIP and H.323 sections to **True**.

Cisco Unified CM Trunks and Emergency Services

IP trunks might be unable to deliver emergency 911 calls or, like centralized PSTN trunks, might be unable to deliver such calls to the appropriate Public Safety Answering Point (PSAP) for the caller's location. Customers must investigate carefully the capabilities of the IP trunk service provider to deliver emergency 911 calls and caller locations to the appropriate PSAP. Cisco Emergency Responder may be used to provide the location-specific calling party number to the IP trunk service provider for emergency 911 calls.

Centralized IP or PSTN trunks might also temporarily become unavailable for emergency 911 calls from remote locations due to WAN congestion or failure. For this reason, remote locations should always have local gateways to the PSTN that are capable of delivering emergency 911 calls. For more information, see Emergency Services, page 10-1.

Capacity Planning for Unified CM IP Trunks

Cisco 7800 Series Media Convergence Servers support the following trunk capacities:

- An MCS-7845 cluster or Cisco Unified Computing System (UCS) equivalent cluster can support up to 2100 trunks.
- An MCS-7835 cluster can support up to 1100 trunks.
- An MCS-7825 cluster can support up to 1100 trunks.
- An MCS-7816 cluster can support up to 200 trunks.

While the above values represent the nominal maximum capacities, actual trunk scalability and performance ultimately depend on several factors including all other applications and tasks that the individual subscribers are processing, the busy hour call attempts (BHCA) across the trunks, and so forth. To determine the overall system capacity, use the Cisco Unified Communications Sizing Tool (Unified CST), which is available to Cisco employees and partners with proper login authentication at

http://tools.cisco.com/cucst

To obtain the most trunk throughput from a cluster, ensure that the trunk load for both incoming and outgoing calls is distributed uniformly over all of the subscribers in the cluster as much as possible.

IP PSTN and IP Trunks to Service Provider Networks

Service providers are increasing their offerings of non-TDM PSTN connections to enterprise customers. Apart from the key benefit of the cost savings from deploying non-TDM interfaces, these IP-based PSTN connections might also offer additional voice features compared to traditional PSTN interfaces.

SIP-based services dominate the available offerings, and although earlier H.323 services were available in select geographies, they are being phased out. This is mainly due to the increasing popularity of SIP within the enterprise and the promise of additional capabilities such as Presence and support for many rich media applications (such as instant messaging). SIP will probably become the most widely deployed Unified Communications protocol in the long term.

When connecting to a service provider's IP PSTN network, Cisco strongly recommends the use of the Cisco Unified Border Element as an enterprise edge Session Border Controller to provide a controlled demarcation and security point between your enterprise and the service provider's network.

Cisco Unified Border Element

The Cisco Unified Border Element provides a wide range of signaling and media functionality between the enterprise and service-provider Cisco Unified Communications networks. Cisco Unified Border Element provides a Session Border Controller network-to-network interface point for:

- Address and port translations (privacy and Level 7 topology hiding)
- SIP Delayed Offer to Early Offer conversion
- Protocol interworking (H.323 and SIP) and normalization
- Media interworking (DTMF, fax, codec transcoding, and transrating volume and gain control)
- Call admission control (based on total calls, CPU, memory, call arrival spike detection, or maximum calls per destination)

- Security (including SIP malformed packet detection, non-dialog RTP packet drops, SIP listening port configuration, digest authentication, simultaneous call limits, call rate limits, toll fraud protection, and a number of signaling and media encryption options)
- PPI/PAI/Privacy and RPID Interworking with service providers
- QoS and bandwidth management (QoS marking using ToS, DSCP, and bandwidth enforcement using RSVP and codec filtering)
- Simultaneous connectivity to SIP trunks from multiple service providers

High availability with in-box or box-to-box failover options (depending on platform and release)

- Billing statistics and CDR collection

The Cisco Unified Border Element is a licensed Cisco IOS application available on the Cisco Integrated Service Routers Generation 2 (ISR G2), the Cisco AS5000XM Media Gateways, and the Cisco 1000 Series Aggregation Services Routers (ASR). Depending on your choice of hardware platform, the Cisco Unified Border Element can provide session scalability for up to 16,000 concurrent voice calls with in-box or box-to-box failover options.

For more information on the Cisco Unified Border Element, refer to the documentation at

http://www.cisco.com/go/cube

Trunk Aggregation Platforms

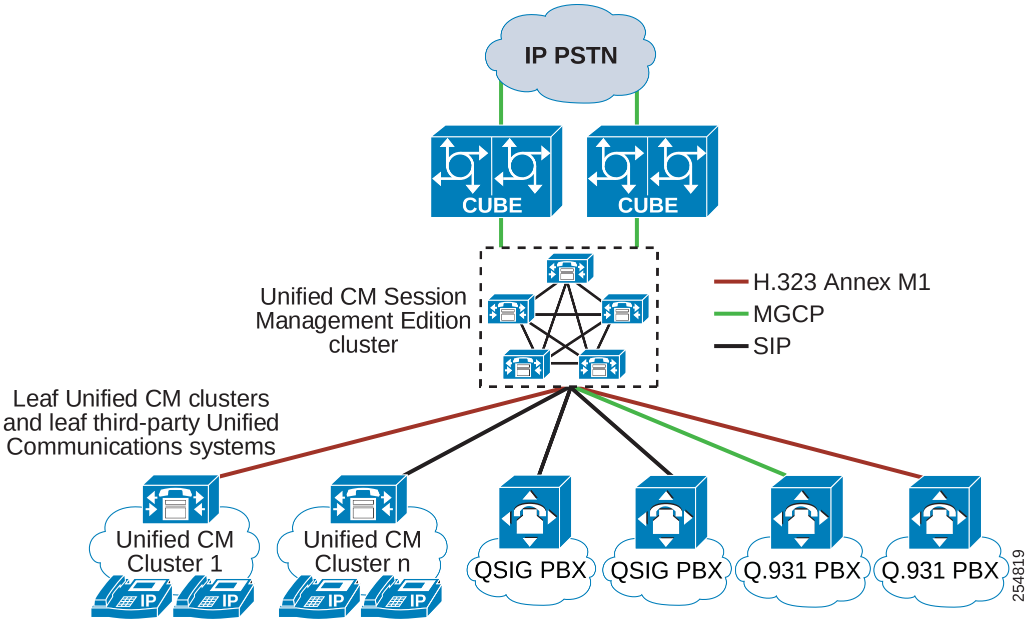
Large-scale IP PSTN deployments often involve the requirement to aggregate trunks from many Unified Communications systems prior to connecting them by means of the Cisco Unified Border Element to the service provider's IP PSTN. In most cases, the choice of aggregation platform depends on the protocol(s) the end systems are using. Cisco Unified CM Session Management Edition and Cisco Unified SIP Proxy are two commonly used aggregation platforms and are discussed in this section. Cisco's H.323 gatekeeper is also an option but tends to be less widely used today since SIP has become the preferred choice of protocol for IP PSTN connections. (Cisco H.323 gatekeeper options are discussed in the section on H.323 Trunk Types, page 14-39.)

Unified CM Session Management Edition

Unified Communications deployments using Unified CM Session Management Edition are a variation on the multisite distributed call processing deployment model and are typically employed to interconnect large numbers of Unified Communications systems and also to provide connections to an IP PSTN.

Cisco Unified CM Session Management Edition is essentially a Unified CM cluster with trunk interfaces only and no IP endpoints. It enables aggregation of multiple unified communications systems, referred to as leaf systems. (See Figure 14-29.)

Figure 14-29 Cisco Unified CM Session Management Edition


IP PSTN
CUBE
CUBE
Unified CM Session Management Edition cluster
H.323 Annex M1
MGCP
SIP
Leaf Unified CM clusters and leaf third-party Unified Communications systems
Unified CM Cluster 1
Unified CM Cluster n
IP
QSIG PBX
QSIG PBX
Q.931 PBX
Q.931 PBX
254819

Unified CM Session Management Edition deployments can be used to migrate a deployment of multiple PBXs and associated phones to a Unified CM cluster with IP phones and relatively few trunks. The Unified CM Session Management Edition cluster may start with a large number of trunks interconnecting third-party PBXs and migrate over time to a Unified CM cluster deployment with thousands of IP phones. Unified CM Session Management Edition may also be used to connect to third-party unified communications systems such as IP PSTN connections and centralized unified communications applications.

Unified CM Session Management Edition supports the following features:

- SIP trunks
- H.323 trunks
- MGCP trunks
- Voice calls
- Video calls
- Encrypted calls
- Fax calls

Because Unified CM Session Management Edition uses the same software as Unified CM, all Unified CM features such as scalability, availability, load balancing, SIP message normalization, call routing, and number modification are available to the Unified CM Session Management Edition cluster.

For more information on Cisco Unified CM Session Management Edition, see the chapter on Unified Communications Deployment Models, page 5-1.

Cisco Unified SIP Proxy

The Cisco Unified SIP Proxy provides SIP proxy functionality on a Cisco NME-522 network module that can be plugged into a network module slot on the Cisco 3800 Series Integrated Services Routers (ISRs). This ISR does not have to be dedicated to hosting the network module and running the proxy alone, but can be used simultaneous for other network functions such as to run the Cisco Unified Border Element described above.

The Cisco Unified SIP Proxy brings the following benefits to a network using Unified CM SIP trunks:

- Aggregation and routing

  The Unified SIP Proxy is capable of connecting several SIP servers to each other without each of the servers connecting to every other one in a full-mesh configuration

- Scalability

  The Unified SIP Proxy can be used to terminate calls to and from the enterprise and IP PSTN service providers. The proxy, in turn, distributes the calls across a pool of Unified Border Elements. More Unified Border Elements may be added to increase capacity.

- Availability and load balancing

  The Unified SIP Proxy distributes calls over the pool of available Unified Border Elements and monitors the status of each Unified Border Element to ensure reliable call completion.

- Message normalization

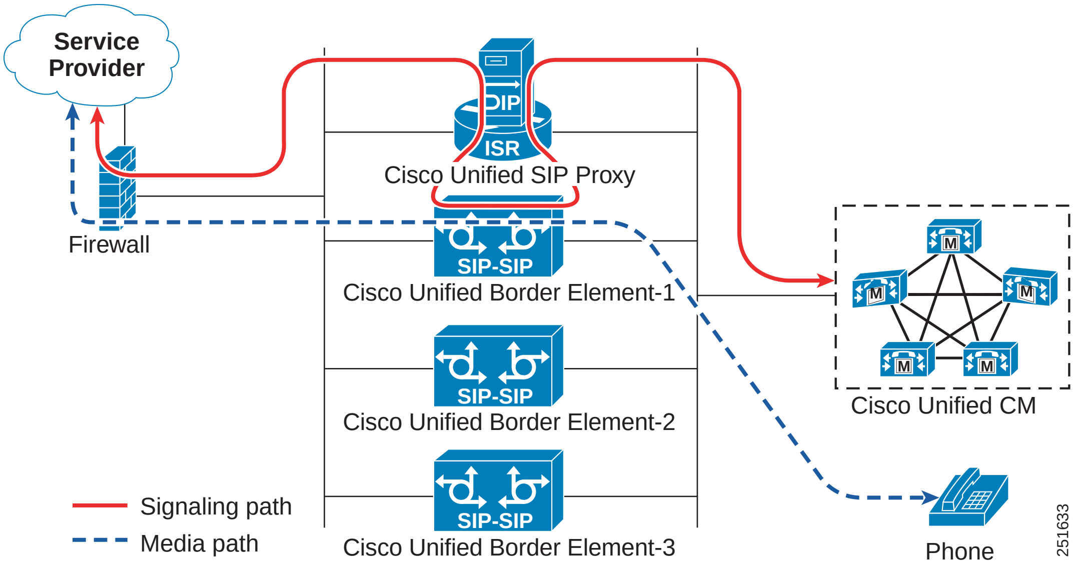
  The Unified SIP Proxy serves to hide differences in SIP protocol messaging by providing the means to manipulate headers and contents of the messages as they pass through the Unified SIP Proxy.

The following design considerations should be taken into account when deploying the Cisco Unified SIP Proxy with the Cisco Unified Border Element:

- Configure a load balancing scheme on the Unified SIP Proxy so that none of the attached Unified Border Elements is overloaded.
- Set up trunk monitoring in the Unified SIP Proxy so that it can detect and act on any failure of the Unified CM or Unified Border Element.

Figure 14-30 shows the call flow using the Cisco Unified SIP Proxy with Cisco Unified Border Element.

Figure 14-30 Call Flow for Cisco Unified SIP Proxy with Cisco Unified Border Element



To originate a call to the service provider, Unified CM sends the call to the Unified SIP Proxy. The Unified SIP Proxy determines that the request came from Unified CM and forwards the call to a Unified Border Element. The Unified Border Element terminates and re-originates the call, and sends it back to the Unified SIP Proxy. The Unified SIP Proxy determines that the call came from a Unified Border Element, and this time forwards it to the service provider. Media is then established directly from the originating phone to the service provider through the Unified Border Element.

Large-Scale Session Border Controller

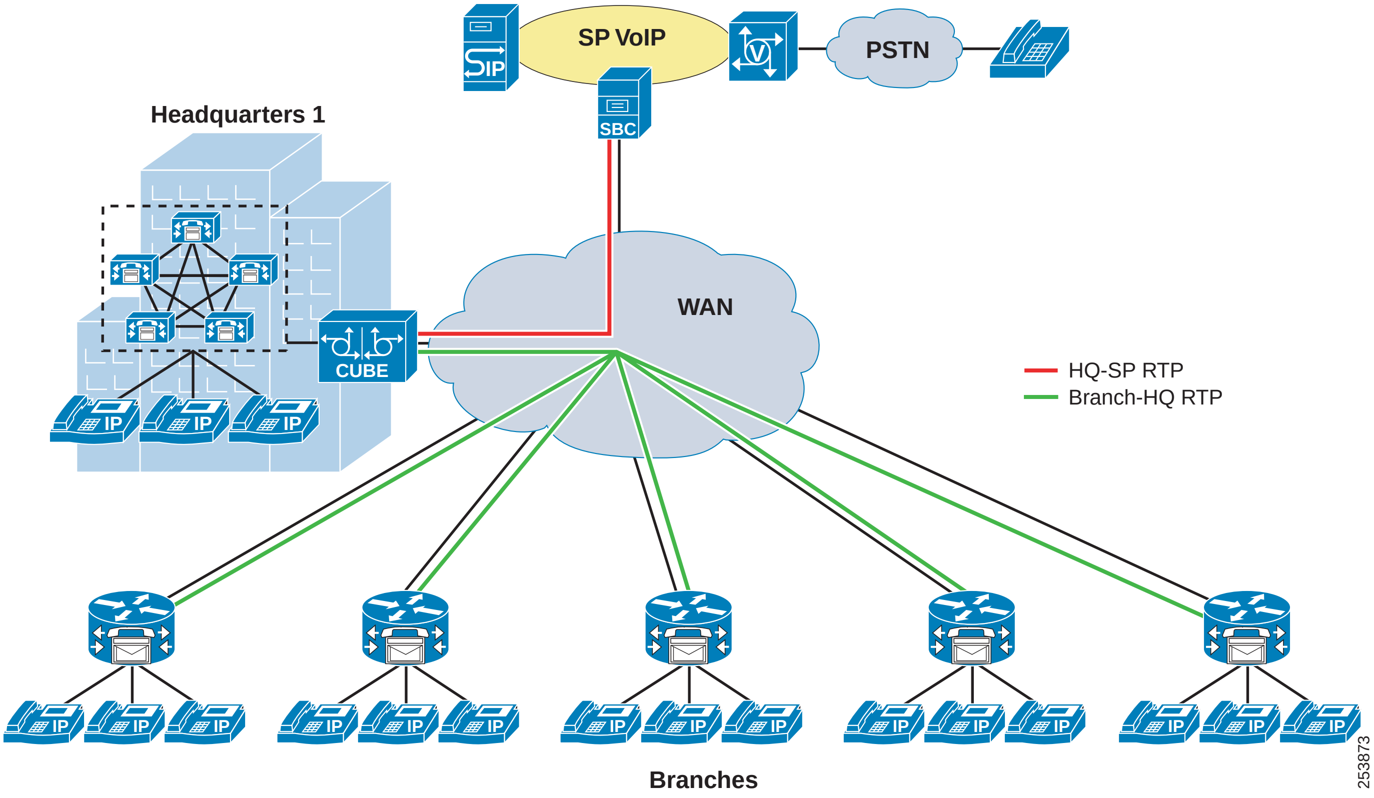
Depending on hardware platform, Cisco Unified Border Element on a single hardware chassis can aggregate up to 16,000 simultaneous calls on SIP trunks from one or more providers, and if desired, also with stateful failover between redundant chassis.

Trunk IP-PSTN Connection Models

Trunks may be connected to IP PSTN service providers in several different ways, depending on the desired architecture. The two most common architectures for this connectivity are centralized trunks and distributed trunks.

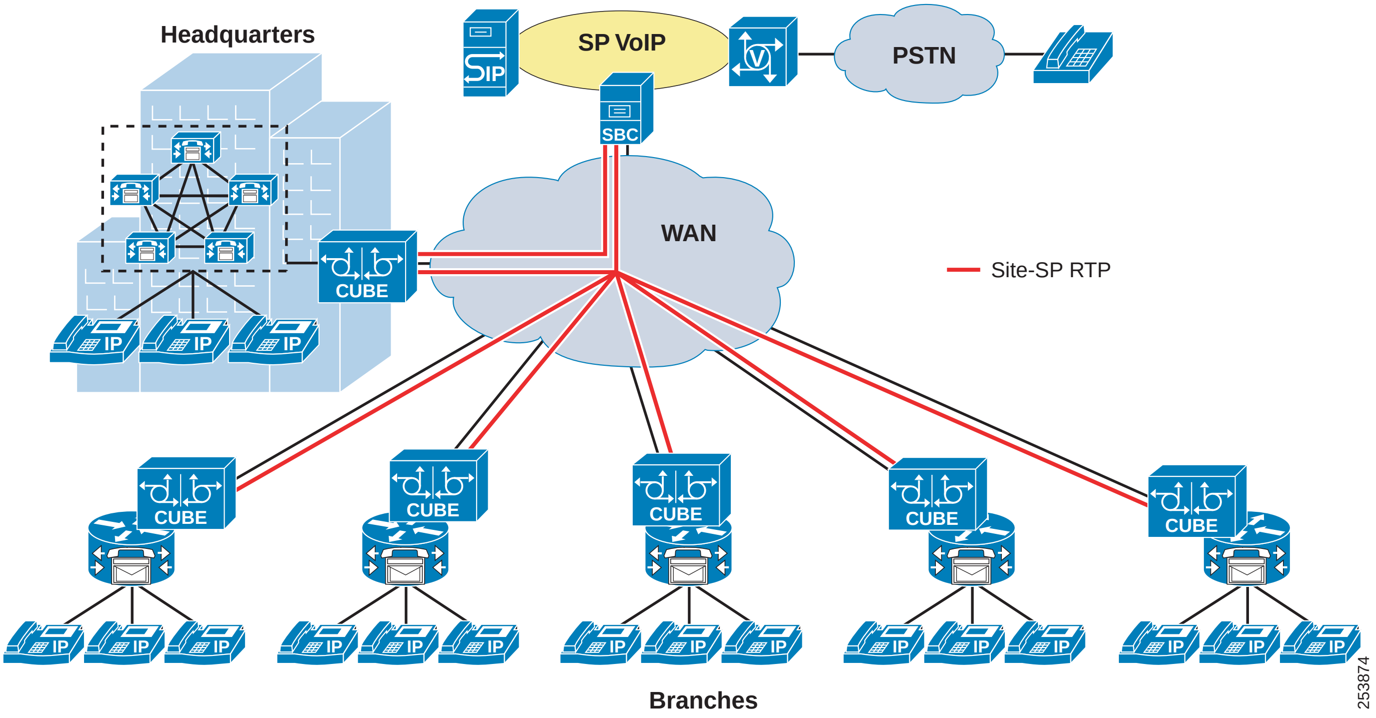
Centralized trunks connect to the service provider (SP) through one logical connection (although there may be more than one physical connection for redundancy) with Session Border Controllers (SBCs) such as the Cisco Unified Border Element. (See Figure 14-31.) All calls to and from the enterprise use this set of trunks. If the enterprise hosts a single central Unified CM cluster at its headquarters, with remote branches connected to the headquarters through a WAN, then media and signaling for PSTN calls to and from each of the sites traverse the WAN.

Figure 14-31 Centralized or Aggregated SIP Trunk Model



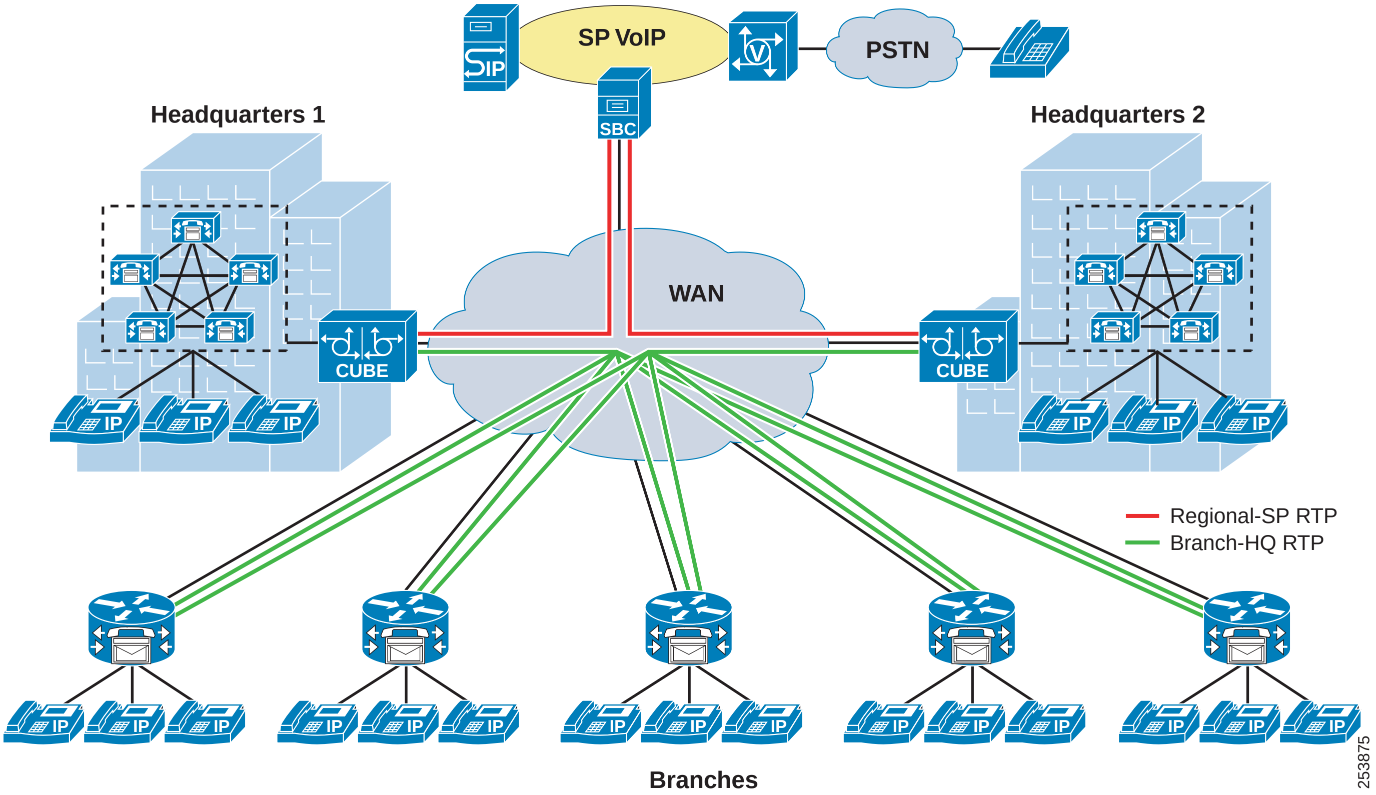
Distributed trunks connect to the service provider through several logical connections. (See Figure 14-32.) Each branch of an enterprise may have its own local trunk to the service provider. Media from branches no longer needs to traverse the WAN but flows to the service provider interface through a local SBC.

Figure 14-32 Distributed SIP Trunk Model



Each connectivity model has its own advantages and disadvantages. Centralized trunks are generally easier to deploy in terms of both physical equipment and configuration complexity. Distributed trunks have the advantage of local hand-off of media and better number portability from local providers. As illustrated in Figure 14-33, a hybrid connectivity model that groups some of the branches together for connectivity, or that provides trunks from each Unified CM cluster of a multi-cluster deployment, captures the advantages of both forms of deployment.

Figure 14-33 Hybrid SIP Trunk Model with Regional Aggregation



PART 3

Unified Communications Call Control

CHAPTER 15

Overview of Cisco Unified Communications Call Control

Revised: June 28, 2012; OL-27282-05

After the network infrastructure and call routing have been properly designed and deployed for your Cisco Unified Communications System, the next phase involves deploying a group of core call control components. These call control components allow users to initiate calls more easily, enhance user capabilities, and enhance the experience of remote callers as well. The following aspects are essential to Unified Communications call control components:

- Integration with central Lightweight Directory Access Protocol (LDAP) directories
- Access to media resources such as audio conferencing or codec transcoding
- Capabilities for music on hold for callers into the Unified Communications System
- Capabilities and feature sets for Unified Communications endpoints
- Applications embedded in the call routing, such as click-to-call dialing, manager-assistant applications, and the ability for users to log in to any phone

This part of the SRND provides coverage for all the various call control components mentioned above. Each chapter provides an introduction to the call control components, followed by discussions surrounding architecture, high availability, capacity planning, and design considerations. The content of each chapter focuses on design-related information rather than product-specific support and configuration information.

This part of the SRND includes the following chapters:

- LDAP Directory Integration, page 16-1

  This chapter covers aspects of Unified Communications integration with the LDAP directories, including the Cisco Unified Communications Manager directory architecture itself as well as design considerations for LDAP synchronization and authentication. Directory access from Unified Communications endpoints and security considerations are also explored.

- Media Resources, page 17-1

  This chapter examines all components classified as Unified Communications media resources. Digital signal processors (DSPs) and their deployment for voice termination, conferencing and transcoding capabilities, and music on hold (MoH) are all discussed. Media termination points (MTPs), how they function, and design considerations with SIP and H.323 trunks are also covered. In addition, design considerations surrounding Trusted Relay Points, RSVP Agents, annunciator, MoH, and secure conferencing are included in the chapter.

- Unified Communications Endpoints, page 18-1

  This chapter discusses all the Unified Communications endpoints available in the Cisco portfolio. Endpoints covered include software-based endpoints, wireless and hard-wired desk phones, video endpoints, and analog gateways and interface modules that provide foreign exchange station (FXS) ports for analog connectivity.

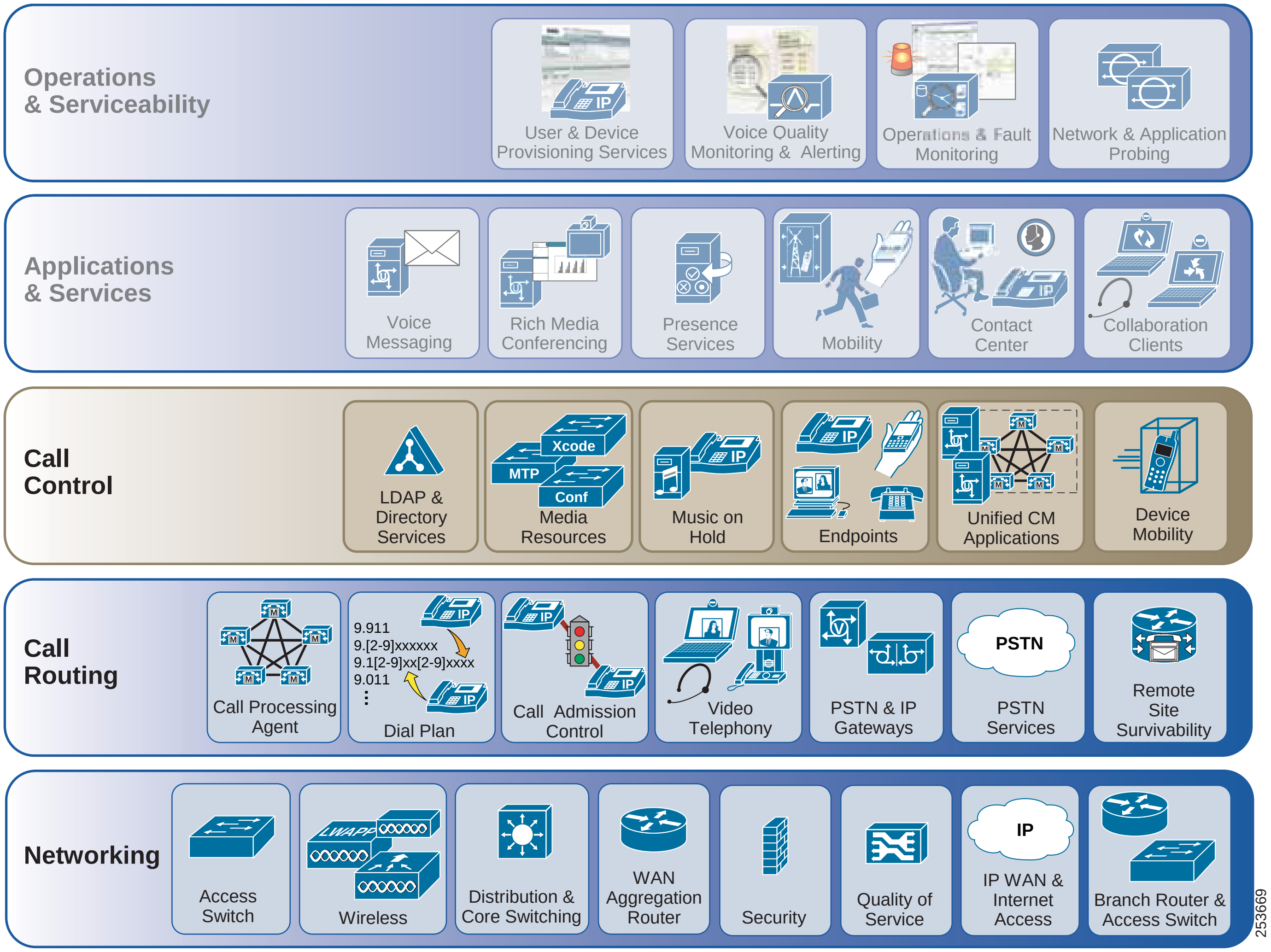
- Cisco Unified CM Applications, page 19-1

  This chapter covers the inherent applications built into Cisco Unified Communications Manager (Unified CM): IP Phone Services, WebDialer, Unified CM Assistant, and Extension Mobility (EM). In addition, this chapter covers Attendant Console applications and their integration through CTI to Unified CM. The chapter first explains the architecture behind the applications, followed by a discussion of design considerations. The chapter also explores variations in the applications such as Extension Mobility Cross Cluster (EMCC) and Unified CM Assistant proxy-line versus shared-line mode.

Architecture

As with other network and application technology systems, unified communications call control components build upon the underlying network and system infrastructures. Figure 15-1 shows the logical location of unified communications call control components in the overall Cisco Unified Communications System architecture.

Figure 15-1 Cisco Unified Communications Call Control Architecture



Unified communications call control components such as conferencing resources, music on hold, directory integration, and endpoints require the unified communications networking infrastructure and unified communications call routing architecture to be well designed and already deployed. These call control components build on the unified communications system and provide enhanced (and usually required) user features.

High Availability

As with the network and call routing, call control infrastructure should be made highly available to ensure that required features and functionality remain available during outages in the network or call processing entities. It is important to understand the various types of failures that can occur and the design considerations around these failures. In some cases, the failure of a single server or feature can impact multiple services because many unified communications components are dependent on others. For example, while the various service components of Cisco Unified Communications Manager (Unified CM) applications may be functioning properly, the loss of the Unified CM call processing service will effectively render the Unified CM applications unusable because the deployment is dependent upon Unified CM to place or receive calls. In many cases, all or part of the functionality can be handled by a redundant resource, thus giving end users the ability to continue to leverage services in the event of certain failures.

For media resources and music on hold, high availability considerations include temporary loss of functionality due to network outages and server or DSP platform failures. This could lead to a poor user experience (for example, initiating a conference only to see a "Resources Not Available" or similar message on the phone) as well as a poor experience for callers into the system (for example, possibly hearing silence instead of a specific advertising message while on hold). Design details around configuration best practices and deployment of redundant resources are discussed in the respective chapters.

Unified CM applications are deployed by enabling specific services on Unified CM nodes in the cluster. If there is a service outage, it will result in a degraded or completely non-functional user experience. Users logging in to phones or accessing IP phone services will experience long delays and typically re-initiate the connection several times, further exacerbating the problem. The chapters in this part of the SRND explain the architecture of the applications and provide design considerations related to which nodes and/or how many nodes in the cluster to enable for application-specific features.

Similarly, for LDAP directory integrations, an LDAP directory server can go offline or the connection between LDAP and the call processing entity can become unavailable. There must be design considerations to allow for LDAP authentication or user directory lookups to continue to function with an alternate server or through an alternate path.

Capacity Planning

Network, call routing, and call control infrastructures must be designed and deployed with an understanding of the capacity and scalability of the individual components and the overall system. When deploying various unified communications call control components, it is important to consider not only the scalability of the components themselves, but also the underlying infrastructures. Certainly the network infrastructure must have available bandwidth and be capable of handling the additional traffic load these components create. Likewise, the call routing infrastructure must be capable of handling user and device configuration and registration as well as additional load surrounding protocols and connections associated with call control elements.

There are capacity planning considerations across all of the chapters of this part of the SRND. For LDAP directory integrations, the most common consideration is the number of users that can be synchronized in the unified communications database, along with polling updates and how they can affect system performance. DSP media resources have a finite number of conferencing or transcoding sessions they each can handle, so proper sizing and allocation of DSPs is critical to a good design. Each Unified CM application has its own set of upper limits, whether it is the supported Extension Mobility login rate or the number of IP Manager Assistants that can be configured in the system. Each call control chapter contains a capacity planning section that offers capacity design guidance and assists in architecting sound Unified Communications designs.

For a complete discussion of system sizing, capacity planning, and deployment considerations related to sizing, refer to the chapter on Unified Communications Design and Deployment Sizing Considerations, page 29-1.

CHAPTER 16

LDAP Directory Integration

Revised: April 30, 2013; OL-27282-05

Directories are specialized databases that are optimized for a high number of reads and searches, and occasional writes and updates. Directories typically store data that does not change often, such as employee information, user policies, user privileges, and group membership on the corporate network.

Directories are extensible, meaning that the type of information stored can be modified and extended. The term *directory schema* defines the type of information stored, its container (or attribute), and its relationship to users and resources.

The Lightweight Directory Access Protocol (LDAP) provides applications with a standard method for accessing and potentially modifying the information stored in the directory. This capability enables companies to centralize all user information in a single repository available to several applications, with a remarkable reduction in maintenance costs through the ease of adds, moves, and changes.

This chapter covers the main design principles for integrating a Cisco Unified Communications system based on Cisco Unified Communications Manager (Unified CM) with a corporate LDAP directory. The main topics include:

- What is Directory Integration?, page 16-2

  This section analyzes the various requirements for integration with a corporate LDAP directory in a typical enterprise IT organization.

- Directory Access for Unified Communications Endpoints, page 16-3

  This section describes the technical solution to enable directory access for Cisco Unified Communications endpoints and provides design best-practices around it.

- Directory Integration with Unified CM, page 16-5

  This section describes the technical solutions and provides design considerations for directory integration with Cisco Unified CM, including the LDAP synchronization and LDAP authentication functions.

The considerations presented in this chapter apply to Cisco Unified CM as well as the following applications bundled with it: Cisco Extension Mobility, Cisco Unified Communications Manager Assistant, WebDialer, Bulk Administration Tool, and Real-Time Monitoring Tool.

For Cisco Unity, refer to the *Cisco Unity Design Guide* and to the following white papers: *Cisco Unity Data and the Directory*, *Active Directory Capacity Planning*, and *Cisco Unity Data Architecture and How Cisco Unity Works*, also available at

http://www.cisco.com

What's New in This Chapter

Table 16-1 lists the topics that are new in this chapter or that have changed significantly from previous releases of this document.

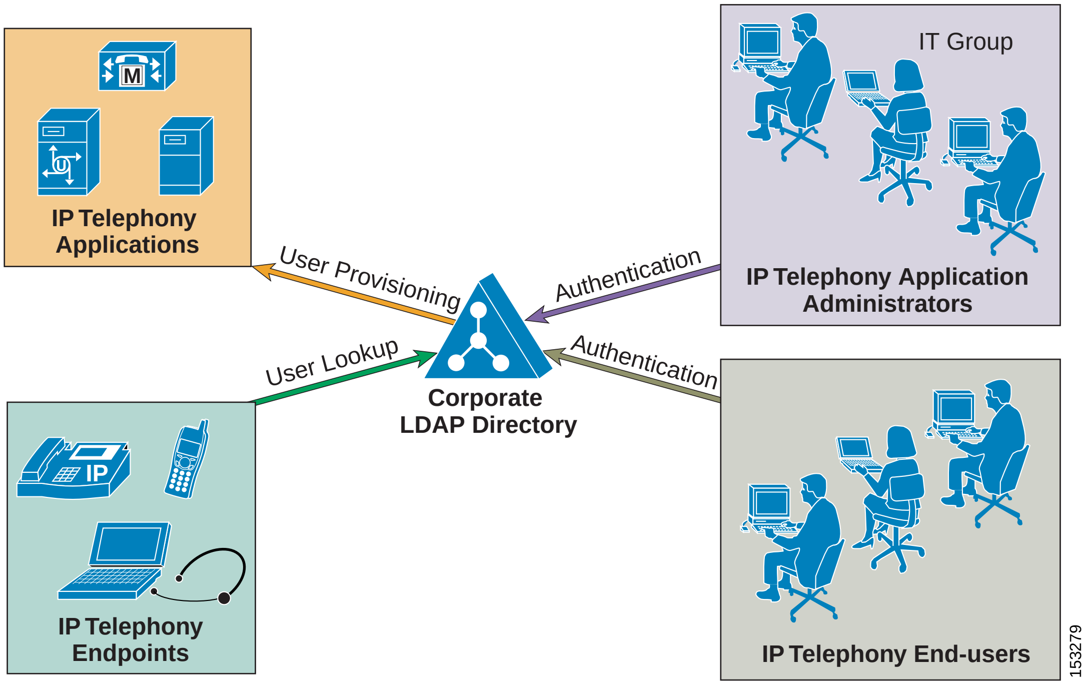
Table 16-1 New or Changed Information Since the Previous Release of This Document

New or Revised Topic	Described in	Revision Date
Minor corrections and changes	Various sections	April 30, 2013
Local and LDAP synchronized users supported simultaneously	LDAP Synchronization, page 16-9	June 28, 2012
Custom LDAP filter strings	Extending the Default Filter, page 16-26	June 28, 2012
Minor updates for Cisco Unified Communications System Release 9.0	Various sections throughout this chapter	June 28, 2012

What is Directory Integration?

Integrating voice applications with a corporate LDAP directory is a common task for many enterprise IT organizations. However, the exact scope of the integration varies from company to company, and can translate into one or more specific and independent requirements, as shown in Figure 16-1.

Figure 16-1 Various Requirements for Directory Integration



One common requirement is to enable user lookups (sometimes called the "white pages" service) from IP phones or other voice and/or video endpoints, so that users can dial contacts quickly after looking up their numbers in the directory.

Another requirement is to provision users automatically from the corporate directory into the user database for applications. This method avoids having to add, remove, or modify core user information manually each time a change occurs in the corporate directory.

Authentication of end users and administrators of the voice and/or video applications using their corporate directory credentials is also a common requirement. Enabling directory authentication allows the IT department to deliver single log-on functionality while reducing the number of passwords each user needs to maintain across different corporate applications.

As shown in Table 16-2, within the context of a Cisco Unified Communications system, the term *directory access* refers to mechanisms and solutions that satisfy the requirement of user lookups for Cisco Unified Communications endpoints, while the term *directory integration* refers to mechanisms and solutions that satisfy the requirements of user provisioning and authentication (for both end users and administrators).

Table 16-2 Directory Requirements and Cisco Solutions

Requirement	Cisco Solution	Cisco Unified CM Feature
User lookup for endpoints	Directory access	Cisco Unified IP Phone Services SDK
User provisioning	Directory integration	LDAP Synchronization
Authentication for Unified Communications end users	Directory integration	LDAP Authentication
Authentication for Unified Communications application administrators	Directory integration	LDAP Authentication

The remainder of this chapter describes how to address these requirements in a Cisco Unified Communications system based on Cisco Unified CM.

Note Another interpretation of the term *directory integration* revolves around the ability to add application servers to a Microsoft Active Directory domain in order to centralize management and security policies. Cisco Unified CM is an appliance that runs on a customized embedded operating system, and it cannot be added to a Microsoft Active Directory domain. Server management for Unified CM is provided through the Cisco Real Time Monitoring Tool (RTMT). Strong security policies tailored to the application are already implemented within the embedded operating system.

Directory Access for Unified Communications Endpoints

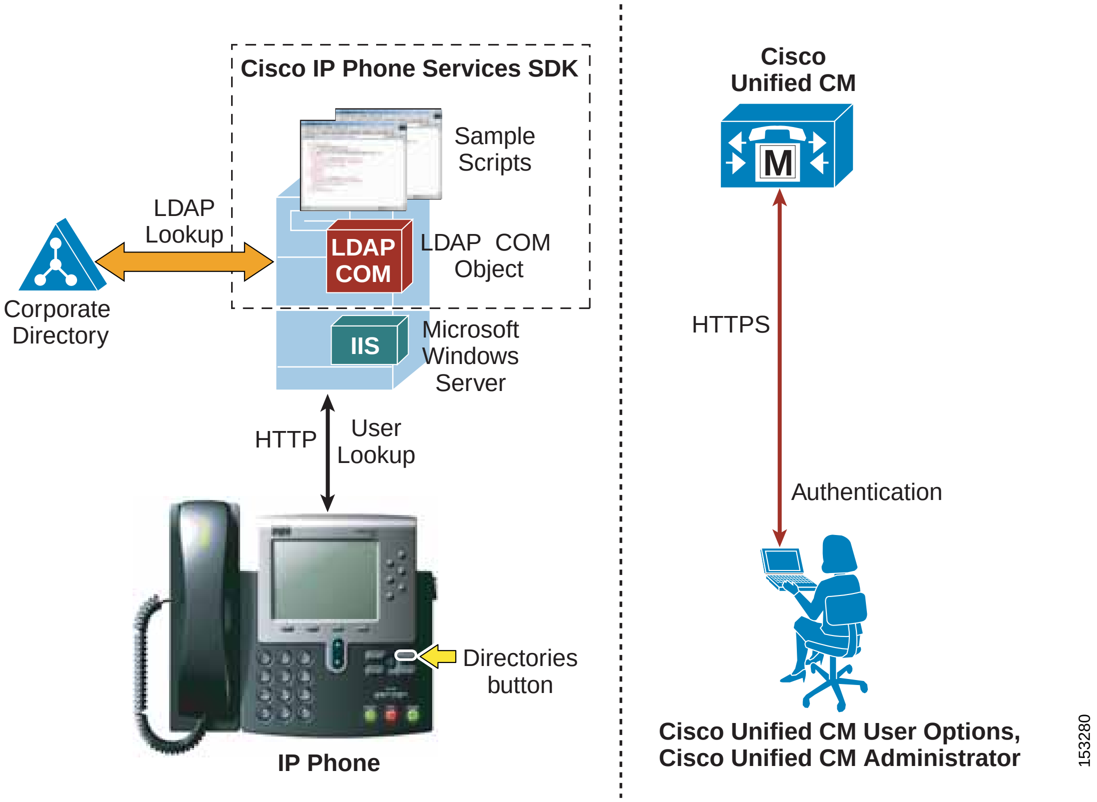
This section describes how to configure corporate directory access to any LDAP-compliant directory server to perform user lookups from Cisco Unified Communications endpoints (such as Cisco Unified IP Phones). The guidelines contained in this section apply regardless of whether Unified CM or other Unified Communications applications have been integrated with a corporate directory for user provisioning and authentication.

Cisco Unified IP Phones equipped with a display screen can search a user directory when a user presses the Directories button on the phone. The IP Phones use Hyper-Text Transfer Protocol (HTTP) to send requests to a web server. The responses from the web server contain specific Extensible Markup Language (XML) objects that the phone interprets and displays.

By default, Cisco Unified IP Phones are configured to perform user lookups against Unified CM's embedded database. However, it is possible to change this configuration so that the lookup is performed on a corporate LDAP directory. In this case, the phones send an HTTP request to an external web server that operates as a proxy by translating the request into an LDAP query which is then processed by the corporate directory. The web server encapsulates the LDAP response into an XML object that is sent back to the phone using HTTP, to be rendered to the end user.

Figure 16-2 illustrates this mechanism in a deployment where Unified CM has not been integrated with the corporate directory. Note that, in this scenario, Unified CM is not involved in the message exchange. The authentication mechanism to Unified CM web pages, shown on the right half of Figure 16-2, is independent of how directory lookup is configured.

Figure 16-2 Directory Access for Cisco Unified IP Phones Using the Cisco Unified IP Phone Services SDK



In the example shown in Figure 16-2, the web server proxy function is provided by the Cisco LDAP Search Component Object Model (COM) server, which is included in the Cisco Unified IP Phone Services Software Development Kit (SDK). You can download the latest Cisco Unified IP Phone Services SDK from the Cisco Developer Community at

http://developer.cisco.com/web/ipps/home

The IP Phone Services SDK can be installed on a Microsoft Windows web server running IIS 4.0 or later, but it cannot be installed on a Unified CM server. The SDK includes some sample scripts to provide simple directory lookup functionality.

To set up a corporate directory lookup service using the IP Phone Services SDK, perform the following steps:

Step 1 Modify one of the sample scripts to point to your corporate LDAP directory, or write your own script using the LDAP Search COM Programming Guide provided with the SDK.

Step 2 In Unified CM, configure the URL Directories parameter (under **System** > **Enterprise Parameters**) to point to the URL of the script on the external web server.

Step 3 Reset the phones to make the changes take effect.

Note If you want to offer the service only to a subset of users, configure the URL Directories parameter directly within the Phone Configuration page instead of the Enterprise Parameters page.

In conclusion, the following design considerations apply to directory access with the Cisco Unified IP Phone Services SDK:

- User lookups are supported against any LDAP-compliant corporate directory.
- When querying Microsoft Active Directory, you can perform lookups against the Global Catalog by pointing the script to a Global Catalog server and specifying port 3268 in the script configuration. This method typically results in faster lookups. Note that a Global Catalog does not contain a complete set of attributes for users. Refer to Microsoft Active Directory documentation for details.
- There is no impact on Unified CM when this functionality is enabled, and only minimal impact on the LDAP directory server.
- The sample scripts provided with the SDK allow only a minimal amount of customization (for example, you can prefix a digit string to all returned numbers). For a higher degree of manipulation, you will have to develop custom scripts, and a programming guide is included with the SDK to aid in writing the scripts.
- This functionality does not entail provisioning or authentication of Unified CM users with the corporate directory.

Directory Integration with Unified CM

This section describes the mechanisms and best practices for directory integration with Cisco Unified CM to allow for user provisioning and authentication with a corporate LDAP directory. This section covers the following topics:

- Cisco Unified Communications Directory Architecture, page 16-6

  This section provides an overview of the user-related architecture in Unified CM.

- LDAP Synchronization, page 16-9

  This section describes the functionality of LDAP synchronization and provides design guidelines for its deployment, with additional considerations for Microsoft Active Directory.

- LDAP Authentication, page 16-18

  This section describes the functionality of LDAP authentication and provides design guidelines for its deployment, with additional considerations for Microsoft Active Directory.

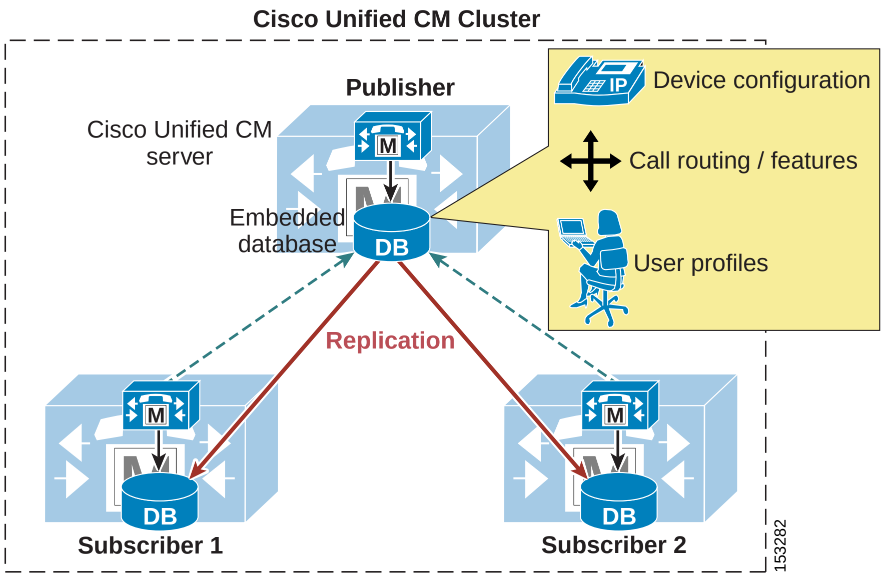
For a list of supported LDAP directories, refer to the latest version of the *Cisco Unified Communications Manager System Guide*, available at

http://www.cisco.com/en/US/products/sw/voicesw/ps556/prod_maintenance_guides_list.html

Cisco Unified Communications Directory Architecture

Figure 16-3 shows the basic architecture of a Unified CM cluster. The embedded database stores all configuration information, including device-related data, call routing, feature provisioning, and user profiles. The database is present on all servers within a Unified CM cluster and is replicated automatically from the publisher server to all subscriber servers.

Figure 16-3 Cisco Unified CM Architecture



By default, all users are provisioned manually in the publisher database through the Unified CM Administration web interface. Cisco Unified CM has two types of users:

- End users — All users associated with a physical person and an interactive login. This category includes all Unified Communications users as well as Unified CM administrators when using the User Groups and Roles configuration (equivalent to the Cisco Multilevel Administration feature in prior Unified CM versions).
- Application users — All users associated with other Cisco Unified Communications features or applications, such as Cisco Attendant Console, Cisco Unified Contact Center Express, or Cisco Unified Communications Manager Assistant. These applications need to authenticate with Unified CM, but these internal "users" do not have an interactive login and serve purely for internal communications between applications.

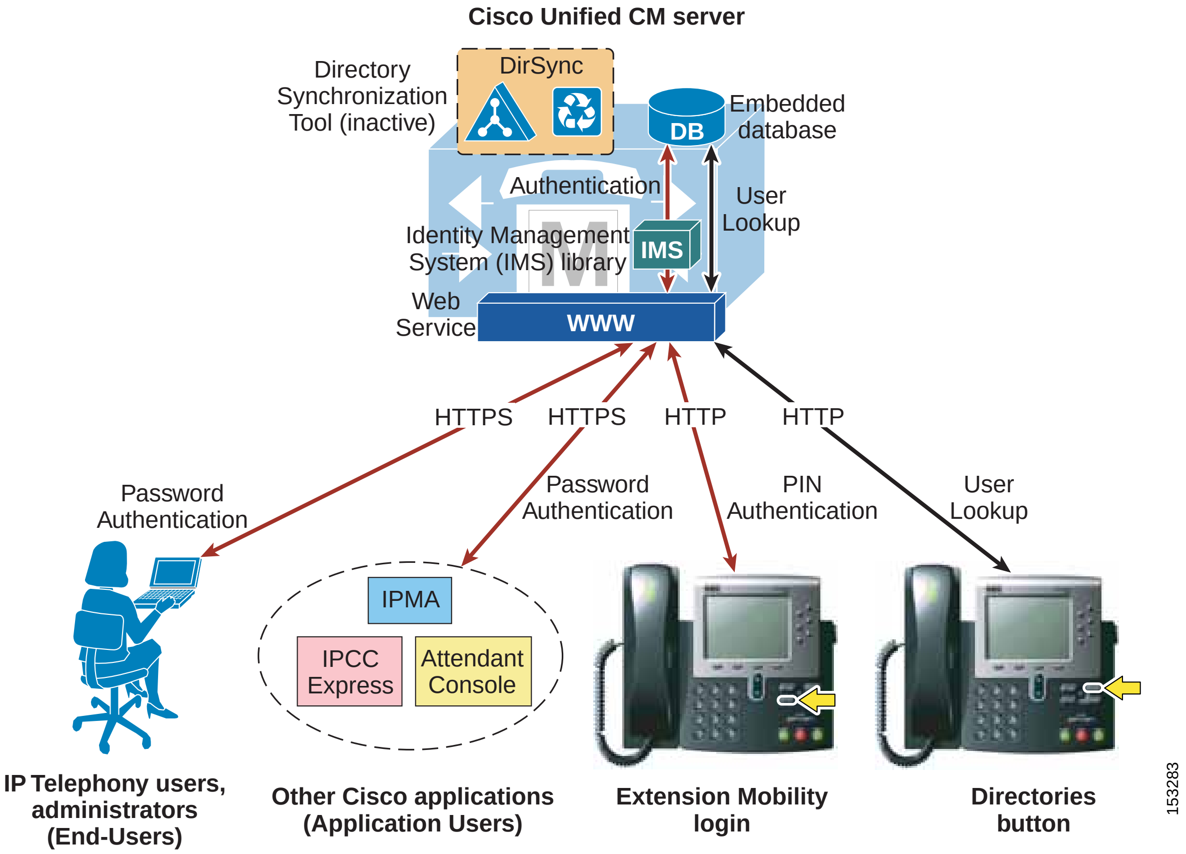
Table 16-3 lists the application users created by default in the Unified CM database, together with the feature or application that uses them. Additional application users can be created manually when integrating other Cisco Unified Communications applications (for example, the **ac** application user for Cisco Attendant Console, the **jtapi** application user for Cisco Unified Contact Center Express, and so forth).

Table 16-3 Default Application Users for Unified CM

Application User	Used by:
CCMAdministrator	Unified CM Administration (default "super user")
CCMQRTSecureSysUser	Cisco Quality Reporting Tool
CCMQRTSysUser	
CCMSysUser	Cisco Extension Mobility
IPMASecureSysUser	Cisco Unified Communications Manager Assistant
IPMASysUser	
WDSecureSysUser	Cisco WebDialer
WDSysUser	

Based on these considerations, Figure 16-4 illustrates the default behavior in Unified CM for user-related operations such as lookups, provisioning, and authentication.

Figure 16-4 Default Behavior for User-Related Operations for Unified CM



End users access the Unified CM User Options page via HTTPS and authenticate with a user name and password. If they have been configured as administrators by means of User Groups and Roles, they can also access the Unified CM Administration pages with the same credentials.

Similarly, other Cisco features and applications authenticate to Unified CM via HTTPS with the user name and password associated with their respective application users.

The authentication challenge carried by the HTTPS messages are relayed by the web service on Unified CM to an internal library called Identity Management System (IMS). In its default configuration, the IMS library authenticates both end users and application users against the embedded database. In this way, both "physical" users of the Unified Communications system and internal application accounts are authenticated using the credentials configured in Unified CM.

End users may also authenticate with their user name and a numeric password (or PIN) when logging into the Extension Mobility service from an IP phone. In this case, the authentication challenge is carried via HTTP to Unified CM but is still relayed by the web service to the IMS library, which authenticates the credentials against the embedded database.

In addition, user lookups performed by Unified Communications endpoints via the Directories button communicate with the web service on Unified CM via HTTP and access data on the embedded database.

The importance of the distinction between End Users and Application Users becomes apparent when integration with a corporate directory is required. As mentioned in the previous section, this integration is accomplished by means of the following two separate processes:

- LDAP synchronization

  This process uses an internal tool called Cisco Directory Synchronization (DirSync) on Unified CM to synchronize a number of user attributes (either manually or periodically) from a corporate LDAP directory. When this feature is enabled, users are automatically provisioned from the corporate directory in addition to local user provisioning through the Unified CM administration GUI. This feature applies only to End Users, while Application Users are kept separate and are still provisioned via the Unified CM Administration interface. In summary, End Users are defined in the corporate directory and synchronized into the Unified CM database, while Application Users are stored only in the Unified CM database and do not need to be defined in the corporate directory.

- LDAP authentication

  This process enables the IMS library to authenticate user credentials of LDAP synchronized End Users against a corporate LDAP directory using the LDAP standard Simple_Bind operation. When this feature is enabled, End User passwords of LDAP synchronized End Users are authenticated against the corporate directory, while Application User passwords and passwords of local End Users are still authenticated locally against the Unified CM database. Cisco Extension Mobility PINs are also still authenticated locally.

Maintaining and authenticating the Application Users internally to the Unified CM database provides resilience for all the applications and features that use these accounts to communicate with Unified CM, independently of the availability of the corporate LDAP directory.

Cisco Extension Mobility PINs are also kept within the Unified CM database because they are an integral part of a real-time application, which should not have dependencies on the responsiveness of the corporate directory.

The next two sections describe in more detail LDAP synchronization and LDAP authentication, and they provide design best-practices for both functions.

Note As illustrated in the section on Directory Access for Unified Communications Endpoints, page 16-3, user lookups from endpoints can also be performed against a corporate directory by configuring the Cisco Unified IP Phone Services SDK on an external web server.

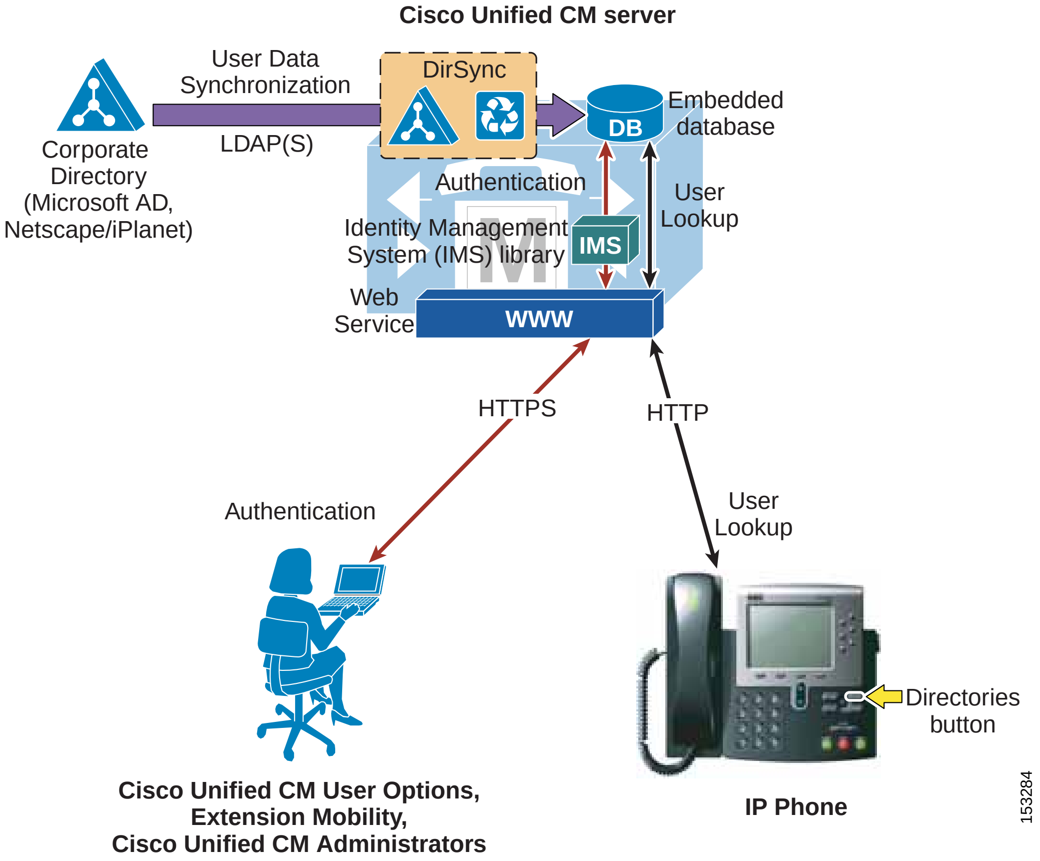
LDAP Synchronization

Synchronization of Unified CM with a corporate LDAP directory allows the administrator to provision users easily by mapping Unified CM data fields to directory attributes. Critical user data maintained in the LDAP store is copied into the appropriate corresponding fields in the Unified CM database on a scheduled or on-demand basis. The corporate LDAP directory retains its status as the central repository. Unified CM has an integrated database for storing user data and a web interface within Unified CM Administration for creating and managing user accounts and data. When LDAP synchronization is enabled, the local database is still used, and additional local end-user accounts can be created. Management of end-user accounts is then accomplished through the interface of the LDAP directory and the Unified CM administration GUI. (See Figure 16-5.). Accounts for application users can be created and managed only through the Unified CM Administration web interface.

The user account information is imported from the LDAP directory into the database located on the Unified CM publisher server. Information that is imported from the LDAP directory may not be changed by Unified CM. Additional user information specific to Cisco Unified Communications is managed by Unified CM and stored only within its local database. For example, device-to-user associations, speed dials, call forward settings, and user PINs are all examples of data that is managed by Unified CM and does not exist in the corporate LDAP directory. The user data is then propagated from the Unified CM publisher server to the subscriber servers through the built-in database synchronization mechanism.

User information synchronized from the LDAP directory can be converted to local user information so that the user information then can be edited locally on Unified CM. After converting an LDAP synchronized user to a local user, the information of that user will not be synchronized from LDAP any longer. Local end-users can be added manually using the Unified CM administration GUI.

Figure 16-5 Enabling Synchronization of User Data



When LDAP synchronization is activated, only one type of LDAP directory may be chosen globally for the cluster at any one time. Also, one attribute of the LDAP directory user is chosen to map into the Unified CM User ID field. Unified CM uses standard LDAPv3 for accessing the data.

Cisco Unified CM imports data from standard attributes. Extending the directory schema is not required. Table 16-4 lists the attributes that are available for mapping to Unified CM fields. The data of the directory attribute that is mapped to the Unified CM User ID must be unique within all entries for that cluster. The attribute mapped to the Cisco UserID field must be populated in the directory and the **sn** attribute must be populated with data, otherwise those records are skipped during this import action. If the primary attribute used during import of end-user accounts matches any application user in the Unified CM database or any local end user, that end user is not imported from the LDAP directory.

Table 16-4 lists the attributes that are imported from the LDAP directory into corresponding Unified CM user fields, and it describes the mapping between those fields. Some Unified CM user fields might be mapped from one of several LDAP attributes.

Table 16-4 Synchronized LDAP Attributes and Corresponding Unified CM Field Names

Unified CM User Field	Microsoft Active Directory	Active Directory Application Mode (ADAM) or Active Directory Lightweight Directory Service (AD LDS)	Netscape, iPlanet, or Sun ONE	OpenLDAP
User ID	*One of:* sAMAccountName mail employeeNumber telephoneNumber userPrincipalName	*One of:* uid mail employeeNumber telephoneNumber userPrincipalName	*One of:* uid mail employeeNumber telephonePhone	*One of:* uid mail employeeNumber telephonePhone
First Name	givenName	givenName	givenname	givenname
Middle Name	*One of:* middleName initials	*One of:* middleName initials	initials	initials
Last Name	sn	sn	sn	sn
Manager ID	manager	manager	manager	manager
Department	department	department	departmentnumber	departmentnumber
Phone Number	One of: telephoneNumber ipPhone	One of: telephoneNumber ipPhone	telephonenumber	telephonenumber
Mail ID	*One of:* mail sAMAccountName	*One of:* mail uid	*One of:* mail uid	*One of:* mail uid

Table 16-5 contains a list of additional attributes that are imported by the Dirsynch process and copied into the Unified CM database but are not displayed in the administrator user configuration web pages. The attribute msRTCSIP-PrimaryUserAddress is populated in AD when Microsoft OCS is used. This table is included for completeness.

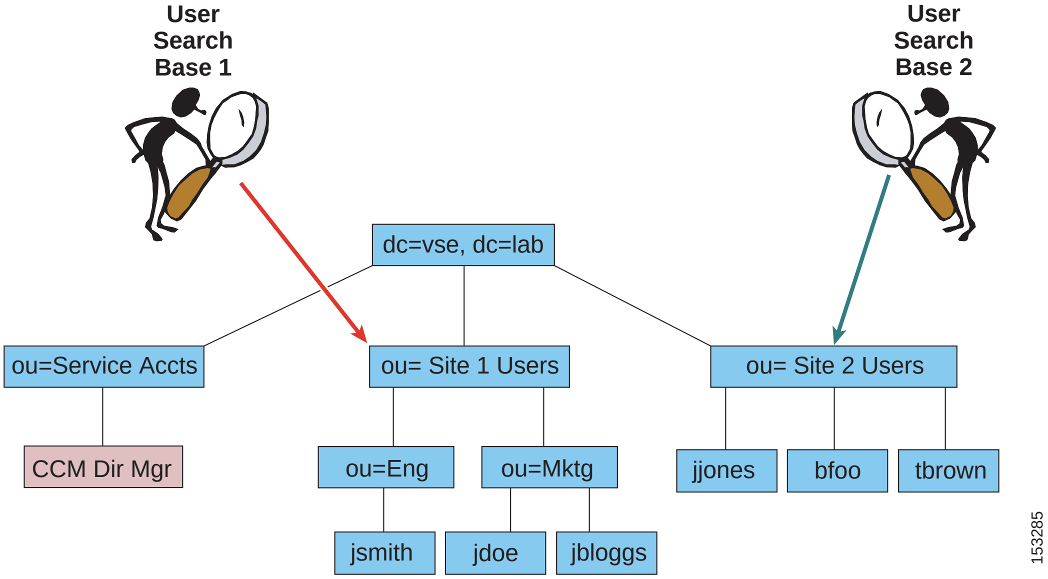
Table 16-5 Synchronized LDAP Attributes that Are Not Displayed

Unified CM User Field	Microsoft Active Directory	Netscape, iPlanet, or Sun ONE	OpenLDAP
objectGUID	objectGUID	Not applicable	Not applicable
OCSPrimaryUserAddress	msRTCSIP-PrimaryUserAddress	Not applicable	Not applicable
Title	title	Title	title
Home Phone Number	homePhone	Homephone	hometelephonenumber
Mobile Phone Number	mobile	Mobile	Mobiletelephonenumber
Pager Number	pager	Pager	pagertelephonenumber

The synchronization is performed by a process called Cisco DirSync, which is enabled through the Serviceability web page. When enabled, it allows one to five synchronization agreements to be configured in the system. An agreement specifies a search base that is a position in the LDAP tree where Unified CM will begin its search for user accounts to import. Unified CM can import only users that exist in the domain specified by the search base for a particular synchronization agreement.

In Figure 16-6, two synchronization agreements are represented. One synchronization agreement specifies User Search Base 1 and imports users jsmith, jdoe and jbloggs. The other synchronization agreement specifies User Search Base 2 and imports users jjones, bfoo, and tbrown. The CCMDirMgr account is not imported because it does not reside below the point specified by a user search base. When users are organized in a structure in the LDAP directory, you can use that structure to control which user groups are imported. In this example, a single synchronization agreement could have been used to specify the root of the domain, but that search base would also have imported the Service Accts. The search base does not have to specify the domain root; it may specify any point in the tree.

Figure 16-6 User Search Bases



To import the data into the Unified CM database, the system performs a bind to the LDAP directory using the account specified in the configuration as the LDAP Manager Distinguished Name, and reading of the database is done with this account. The account must be available in the LDAP directory for Unified CM to log in, and Cisco recommends that you create a specific account with permissions to allow it to read all user objects within the sub-tree that was specified by the user search base. The sync agreement specifies the full Distinguished Name of that account so that the account may reside anywhere within that domain. In the example in Figure 16-6, CCMDirMgr is the account used for the synchronization.

It is possible to control the import of accounts through use of permissions of the LDAP Manager Distinguished Name account. In this example, if that account is restricted to have read access to ou=Eng but not to ou=Mktg, then only the accounts located under Eng will be imported.

Synchronization agreements have the ability to specify multiple directory servers to provide redundancy. You can specify an ordered list of up to three directory servers in the configuration that will be used when attempting to synchronize. The servers are tried in order until the list is exhausted. If none of the directory servers responds, then the synchronization fails, but it will be attempted again according to the configured synchronization schedule.

Synchronization Mechanism

The synchronization agreement specifies a time for synchronizing to begin and a period for re-synchronizing that can be specified in hours, days, weeks, or months (with a minimum value of 6 hours). A synchronization agreement can also be set up to run only once at a specific time.

When synchronization is enabled for the first time on a Unified CM publisher server, user accounts that exist in the corporate directory are imported into the Unified CM database. Then either existing Unified CM end-user accounts are activated and data is updated, or a new end-user account is created according to the following process:

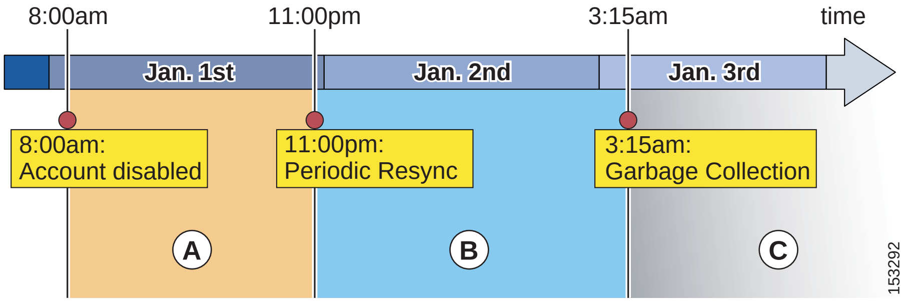
1. If end-user accounts already exist in the Unified CM database and a synchronization agreement is configured, all pre-existing accounts that have been synchronized from LDAP previously are marked inactive in Unified CM. The configuration of the synchronization agreement specifies a mapping of an LDAP database attribute to the Unified CM UserID. During the synchronization, accounts from the LDAP database that match an existing Unified CM account cause that Unified CM account to be marked active again. If accounts from LDAP match an existing Unified CM account that is not marked as an LDAP synchronized account, then these accounts are ignored.
2. After the synchronization is completed, any LDAP synchronized accounts that were not set to active are permanently deleted from Unified CM when the garbage collection process runs. Garbage collection is a process that runs automatically at the fixed time of 3:15 AM, and it is not configurable.
3. Subsequently when changes are made in the corporate directory, the synchronization from Microsoft Active Directory occurs as a full re-synchronization at the next scheduled synchronization period. On the other hand, the iPlanet and Sun ONE directory products perform an incremental synchronization triggered by a change in the directory. The following sections present examples of each of these two scenarios.

Note Once users are synchronized from LDAP into the Unified CM database, deletion of a synchronization configuration will cause users that were imported by that configuration to be marked inactive in the database. Garbage collection will subsequently remove those users.

Account Synchronization with Active Directory

Figure 16-7 shows an example timeline of events for a Unified CM deployment where LDAP Synchronization and LDAP Authentication have both been enabled. The re-synchronization is set for 11:00 PM daily.

Figure 16-7 Change Propagation with Active Directory



After the initial synchronization, the creation, deletion, or disablement of an account will propagate to Unified CM according to the timeline shown in Figure 16-7 and as described in the following steps:

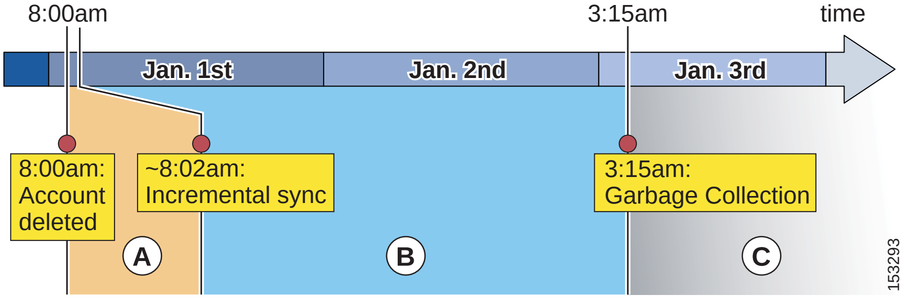
1. At 8:00 AM on January 1, an account is disabled or deleted in AD. From this time and during the whole period A, password authentication (for example, Unified CM User Options page) will fail for this user because Unified CM redirects authentication to AD. However, PIN authentication (for example, Extension Mobility login) will still succeed because the PIN is stored in the Unified CM database.
2. The periodic re-synchronization is scheduled for 11:00 PM on January 1. During that process, Unified CM will verify all accounts. Any accounts that have been disabled or deleted from AD will at that time be tagged in the Unified CM database as inactive. After 11:00 PM on January 1, when the account is marked inactive, both the PIN and password authentication by Unified CM will fail.
3. Garbage collection of accounts occurs daily at the fixed time of 3:15 AM. This process permanently deletes user information from the Unified CM database for any record that has been marked inactive for over 24 hours. In this example, the garbage collection that runs at 3:15 AM on January 2 does not delete the account because it has not been inactive for 24 hours yet, so the account is deleted at 3:15 AM on January 3. At that point, the user data is permanently deleted from Unified CM.

If an account has been created in AD at the beginning of period A, it will be imported to Unified CM at the periodic re-synchronization that occurs at the beginning of period B and will immediately be active on Unified CM.

Account Synchronization with iPlanet or Sun ONE

The iPlanet and Sun ONE products support incremental synchronization agreements and use a different synchronization timeline than Microsoft Active Directory. The synchronization makes use of the Persistent Search mechanism defined by an IETF Draft and supported by many LDAP implementations. Figure 16-8 shows an example of this synchronization timeline for a Unified CM deployment with LDAP Synchronization and LDAP Authentication both enabled.

Figure 16-8 Change Propagation with iPlanet and Sun ONE



The example in Figure 16-8 involves the following steps:

1. An account is deleted from the corporate directory at 8:00 AM on January 1, which causes an incremental update to be sent from the LDAP server to Unified CM. Unified CM sets its corresponding copy of the data to inactive. Because LDAP authentication is configured, the user will be unable to log in via password as soon as the LDAP server has deleted the record. Also, the PIN may not be used for login at the moment the Unified CM record is marked inactive.
2. During period B, the user's record is still present in Unified CM, albeit inactive.

3. When the garbage collection runs at 3:15 AM on January 2, the record has not yet been inactive for 24 hours. The data remains in the Unified CM database until the beginning of period C on January 3, when the garbage collection process runs again at 3:15 AM and determines that the record has been inactive for 24 hours or more. The record is then permanently deleted from the database.

Accounts that are newly created in the directory are synchronized to Unified CM via incremental updates as well, and they may be used as soon as the incremental update is received.

Security Considerations

During the import of accounts, no passwords or PINs are copied from the LDAP directory to the Unified CM database. If LDAP authentication is not enabled in Unified CM, the password for the end user is managed by using Unified CM Administration. The password and PIN are stored in an encrypted format in the Unified CM database. The PIN is always managed on Unified CM. If you want to use the LDAP directory password to authenticate an end user, see the section on LDAP Authentication, page 16-18.

The connection between the Unified CM publisher server and the directory server can be secured by enabling Secure LDAP (SLDAP) on Unified CM and the LDAP server. Secure LDAP enables LDAP to be sent over a Secure Socket Layer (SSL) connection and can be enabled by uploading the SSL certificate from within the Unified CM Platform Administration. For detailed procedure steps, refer to the Unified CM product documentation available at http://www.cisco.com. Refer to the documentation of the LDAP directory vendor to determine how to enable SLDAP.

Design Considerations for LDAP Synchronization

Observe the following design and implementation best practices when deploying LDAP synchronization with Cisco Unified CM:

- Use a specific account within the corporate directory to allow the Unified CM synchronization agreement to connect and authenticate to it. Cisco recommends that you use an account dedicated to Unified CM, with minimum permissions set to "read" all user objects within the desired search base and with a password set never to expire. The password for this account in the directory must be kept in synchronization with the password configuration of the account in Unified CM. If the service account password changes in the directory, be sure to update the account configuration in Unified CM.
- All synchronization agreements on a given cluster must integrate with the same family of LDAP servers.
- Stagger the scheduling of synchronization agreements so that multiple agreements are not querying the same LDAP servers simultaneously. Choose synchronization times that occur during quiet periods (off-peak hours).
- If security of user data is required, enable Secure LDAP (SLDAP) by checking the **Use SSL** field on the LDAP Directory configuration page in Unified CM Administration.
- Ensure that the LDAP directory attribute chosen to map into the Unified CM UserID field is unique within all synchronization agreements for that cluster.
- The attribute chosen as UserID must not be the same as that for any of the Application Users defined in Unified CM.
- The LDAP attribute sn(lastname) is a mandatory attribute for LDAP Synchronization of users.
- An existing account in the Unified CM database before synchronization is maintained only if an account imported from the LDAP directory has a matching attribute. The attribute that is matched to the Unified CM UserID is determined by the synchronization agreement.

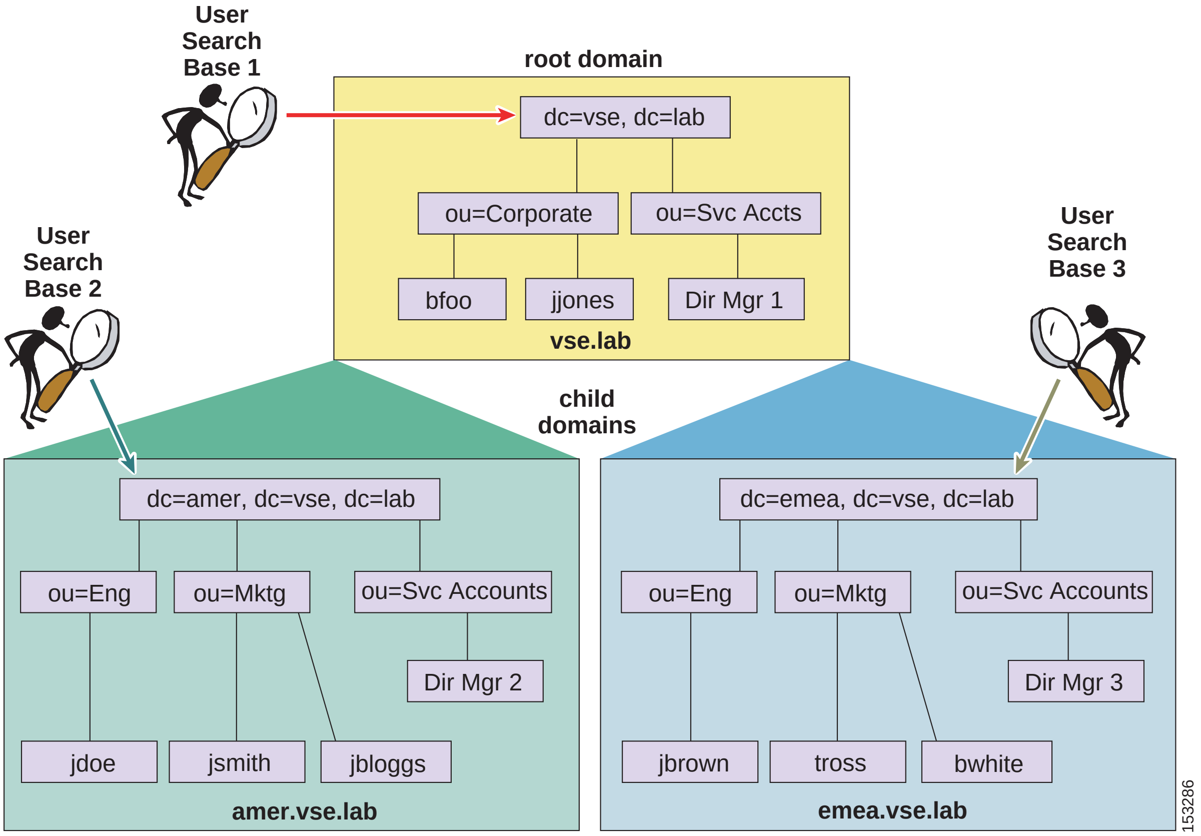
- Administer end-user accounts through the LDAP directory's management tools, and manage the Cisco-specific data for those accounts through the Unified CM Administration web page.
- LDAP Synchronization is supported only with Microsoft NT LAN Manager (NTLM). Kerberos and NTLMv2 are not supported.
- For AD deployments, the ObjectGUID is used internally in Unified CM as the key attribute of a user. The attribute in AD that corresponds to the Unified CM User ID may be changed in AD. For example, if sAMAccountname is being used, a user may change their sAMAccountname in AD, and the corresponding user record in Unified CM would be updated.

  With all other LDAP platforms, the attribute that is mapped to User ID is the key for that account in Unified CM. Changing that attribute in LDAP will result in a new user being created in Unified CM, and the original user will be marked inactive.

Additional Considerations for Microsoft Active Directory

A synchronization agreement for a domain will not synchronize users outside of that domain nor within a child domain because Unified CM does not follow AD referrals during the synchronization process. The example in Figure 16-9 requires three synchronization agreements to import all of the users. Although Search Base 1 specifies the root of the tree, it will not import users that exist in either of the child domains. Its scope is only VSE.LAB, and separate agreements are configured for the other two domains to import those users.

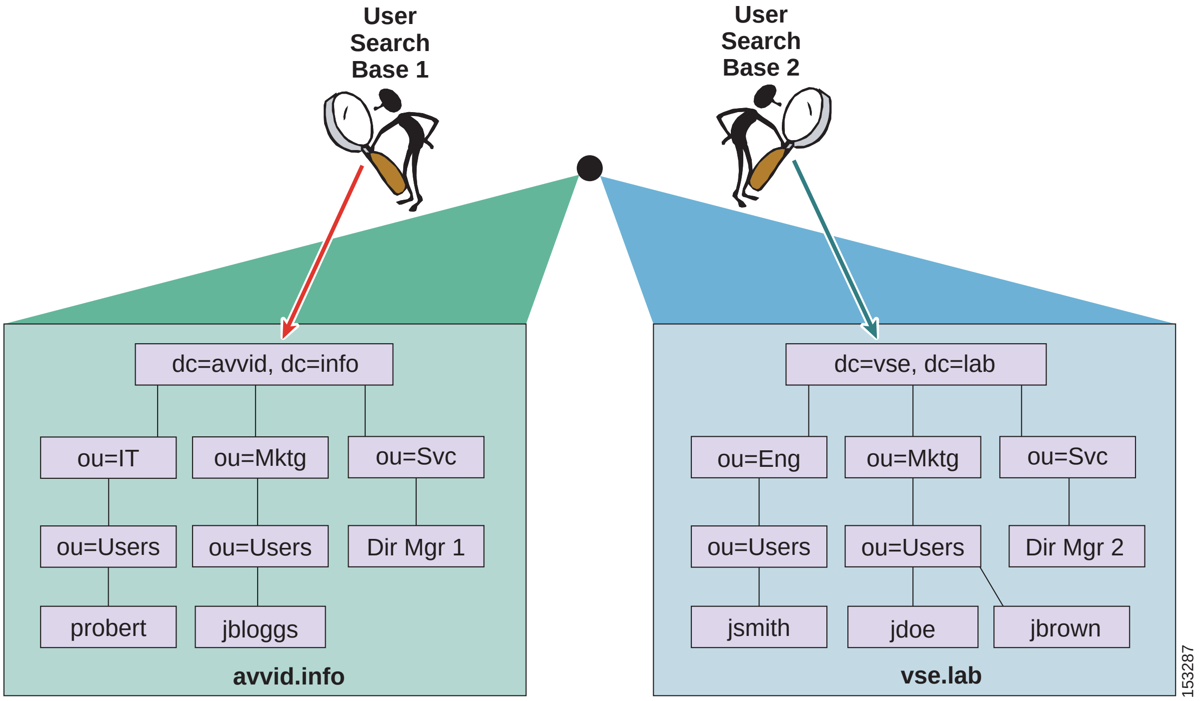
Figure 16-9 Synchronization with Multiple Active Directory Domains



In Figure 16-9, each of the domains and sub-domains contains at least one domain controller (DC) associated to them, and the three synchronization agreements each specify the appropriate domain controller. The DCs have information only on users within the domain where they reside, therefore three synchronization agreements are required to import all of the users.

When synchronization is enabled with an AD forest containing multiple trees, as shown in Figure 16-10, multiple synchronization agreements are still needed for the same reasons listed above. Additionally, the UserPrincipalName (UPN) attribute is guaranteed by Active Directory to be unique across the forest and must be chosen as the attribute that is mapped to the Unified CM UserID. For additional considerations on the use of the UPN attribute in a multi-tree AD scenario, see the section on Additional Considerations for Microsoft Active Directory, page 16-21.

Figure 16-10 Synchronization with Multiple AD Trees (Discontiguous Namespaces)



Unified CM sends a default LDAP search filter string to AD when performing the synchronization of accounts. One of the clauses is to not return accounts that have been marked as disabled in AD. An account marked disabled by AD, such as when failed login attempts are exceeded, will be marked inactive if synchronization runs while the account is disabled.

Unified CM Multi-Forest LDAP Synchronization

A Unified CM deployment using a multi-forest LDAP infrastructure can be supported by using Active Directory Lightweight Directory Services (AD LDS) as a single forest view integrating with the multiple disparate forests. The integration also requires the use of LDAP filtering (see User Filtering for

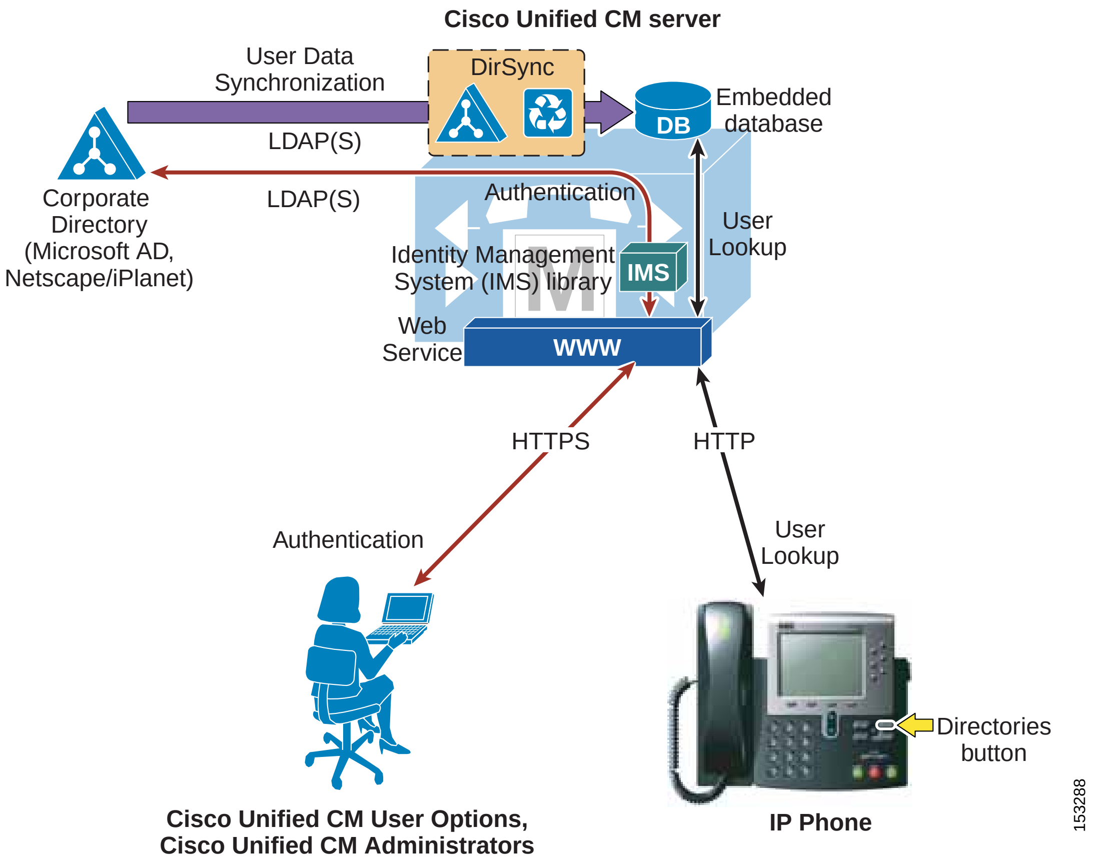
Directory Synchronization and Authentication, page 16-23). For full details, refer to the document on *How to Configure Unified Communication Manager Directory Integration in a Multi-Forest Environment*, available at

http://www.cisco.com/en/US/products/sw/voicesw/ps556/products_configuration_example09186a0080b2b103.shtml

LDAP Authentication

The LDAP authentication feature enables Unified CM to authenticate LDAP synchronized users against the corporate LDAP directory. Application users and locally configured users are always authenticated against the local database. Also PINs of all end users are always checked against the local database only. This authentication is accomplished with an LDAPv3 connection established between the IMS module within Unified CM and a corporate directory server, as shown in Figure 16-11.

Figure 16-11 Enabling LDAP Authentication



To enable authentication, a single authentication agreement may be defined for the entire cluster. The authentication agreement supports configuration of up to three LDAP servers for redundancy and also supports secure connections LDAP over SSL (SLDAP) if desired. Authentication can be enabled only when LDAP synchronization is properly configured and used.

The following statements describe Unified CM's behavior when authentication is enabled:

- End user passwords of users imported from LDAP are authenticated against the corporate directory by a simple bind operation.
- Application user passwords and passwords of locally configured end users are authenticated against the Unified CM database.
- End user PINs are authenticated against the Unified CM database.

This behavior is in line with the guiding principle of providing single logon functionality for end users while making the operation of the real-time Unified Communications system independent of the availability of the corporate directory, and is shown graphically in Figure 16-12.

Figure 16-12 Authenticating LDAP Synchronized End User Passwords, Application User Passwords, and End User PINs

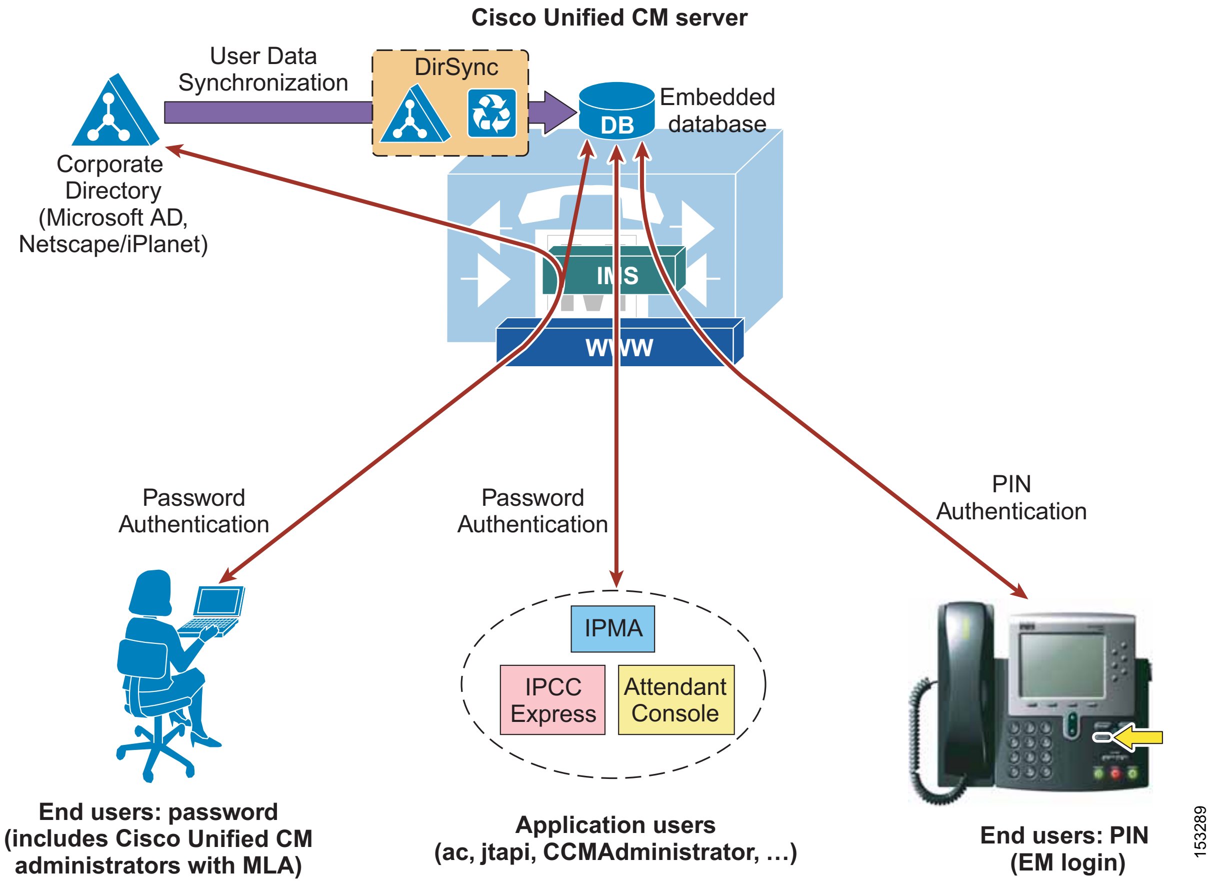
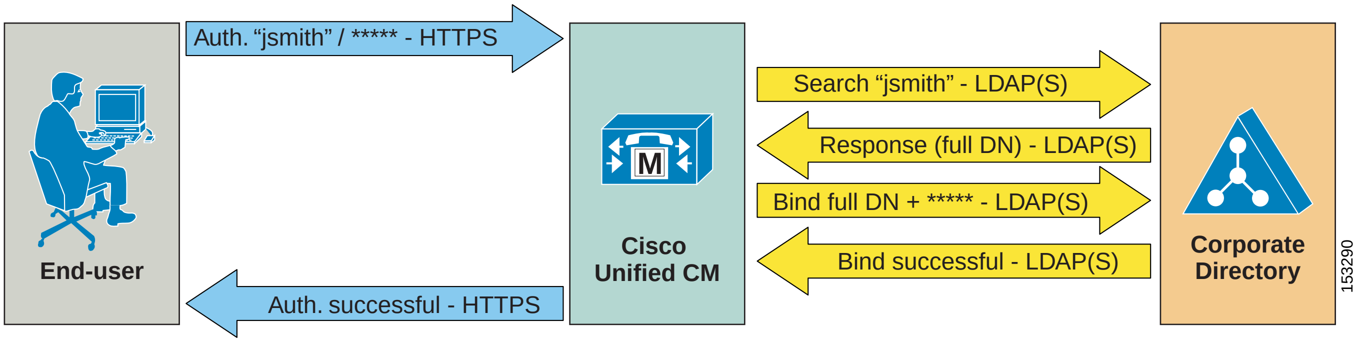

Cisco Unified CM server
User Data Synchronization
DirSync
Embedded database
DB
Corporate Directory (Microsoft AD, Netscape/iPlanet)
IMS
WWW
Password Authentication
Password Authentication
PIN Authentication
IPMA
IPCC Express
Attendant Console
End users: password (includes Cisco Unified CM administrators with MLA)
Application users (ac, jtapi, CCMAdministrator, …)
End users: PIN (EM login)
153289

Figure 16-13 illustrates the following process, adopted by Unified CM to authenticate an end user synchronized from LDAP against a corporate LDAP directory:

1. A user connects to the Unified CM User Options page via HTTPS and attempts to authenticate with a user name and password. In this example, the user name is jsmith.
2. Unified CM issues an LDAP query for the user name jsmith, using the value specified in the LDAP Search Base on the LDAP Authentication configuration page as the scope for this query. If SLDAP is enabled, this query travels over an SSL connection.

3. The corporate directory server replies via LDAP with the full Distinguished Name (DN) of user jsmith (for example, "cn=jsmith, ou=Users, dc=vse, dc=lab").
4. Unified CM then attempts to validate the user's credentials by using an LDAP bind operation to pass the full DN and password provided by the user.
5. If the LDAP bind is successful, Unified CM allows the user to proceed to the configuration page requested.

Figure 16-13 Authentication Process


Auth. “jsmith” / ***** - HTTPS
End-user
Cisco Unified CM
Search “jsmith” - LDAP(S)
Response (full DN) - LDAP(S)
Bind full DN + ***** - LDAP(S)
Bind successful - LDAP(S)
Corporate Directory
153290
Auth. successful - HTTPS

Design Considerations for LDAP Authentication

Observe the following design and implementation best-practices when deploying LDAP authentication with Cisco Unified CM:

- Create a specific account within the corporate directory to allow Unified CM to connect and authenticate to it. Cisco recommends that you use an account dedicated to Unified CM, with minimum permissions set to "read" all user objects within the desired search base and with a password set to never expire. The password for this account in the directory must be kept in synchronization with the password configuration of the account in Unified CM. If the account password changes in the directory, be sure to update the account configuration in Unified CM. If LDAP synchronization is also enabled, you can use the same account for both functions.
- Enable LDAP authentication on Unified CM by specifying the credentials of the aforementioned account under LDAP Manager Distinguished Name and LDAP Password, and by specifying the directory subtree where all the users reside under LDAP User Search Base.
- This method provides single logon functionality to all end users synchronized from LDAP. When they log in to the Unified CM User Options page, they can use their corporate directory credentials.
- Manage end user passwords of LDAP synchronized end users from within the corporate directory interface. Note that the password field is no longer displayed in Unified CM Administration for users synchronized from LDAP when authentication is enabled.
- Manage end user PINs from the Unified CM Administration web pages or from the Unified CM User Options page.
- Manage Application User passwords from the Unified CM Administration web pages. Remember that these application users facilitate communication and remote call control with other Cisco Unified Communications applications and are not associated with real people.
- Enable single logon for Unified CM administrators by adding their corresponding end user to the Unified CM Super Users user group from the Unified CM Administration web pages. Multiple levels of administrator rights can be defined by creating customized user groups and roles.

Additional Considerations for Microsoft Active Directory

In environments that employ a distributed AD topology with multiple domain controllers geographically distributed, authentication speed might be unacceptable. When the Domain Controller for the authentication agreement does not contain a user account, a search must occur for that user across other domain controllers. If this configuration applies, and login speed is unacceptable, it is possible to set the authentication configuration to use a Global Catalog Server.

An important restriction exists, however. A Global Catalog does not carry the employeeNumber attribute by default. In that case either use Domain Controllers for authentication (beware of the limitations listed above) or update the Global Catalog to include the employeeNumber attribute. Refer to Microsoft Active Directory documentation for details.

To enable queries against the Global Catalog, simply configure the LDAP Server Information in the LDAP Authentication page to point to the IP address or host name of a Domain Controller that has the Global Catalog role enabled, and configure the LDAP port as 3268.

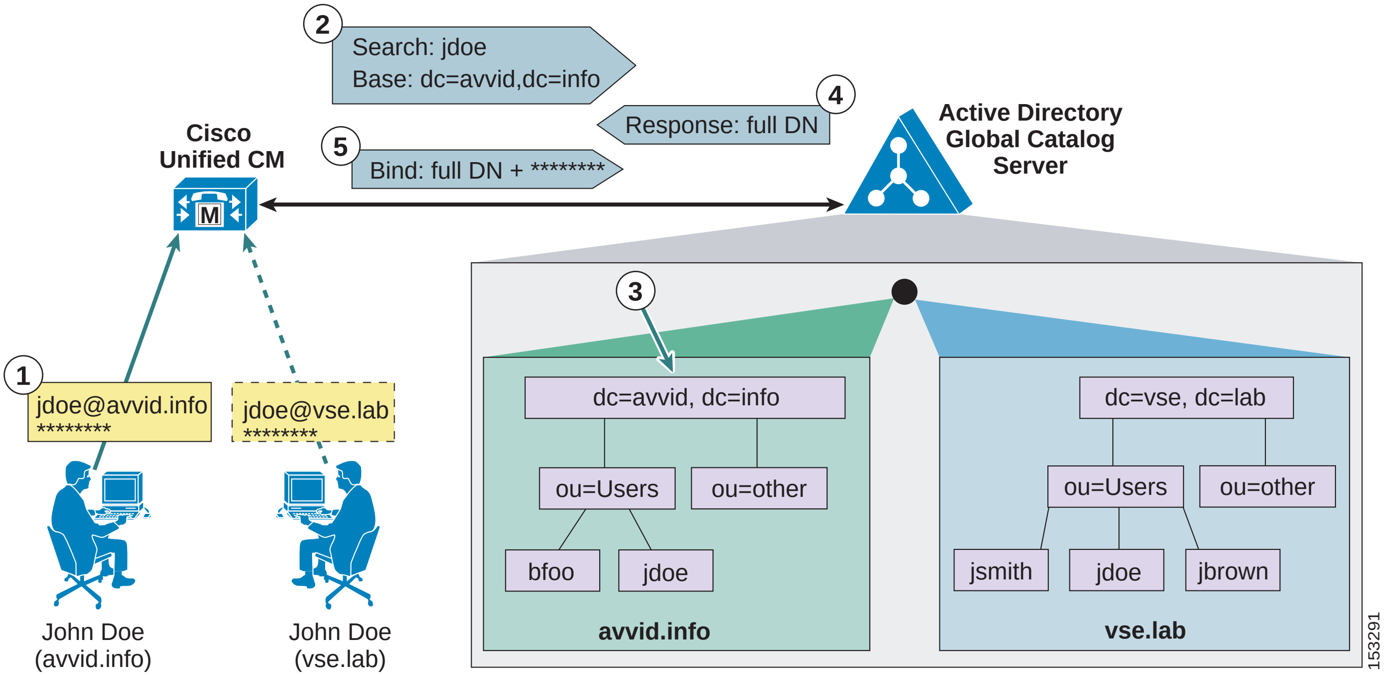
The use of Global Catalog for authentication becomes even more efficient if the users synchronized from Microsoft AD belong to multiple domains, because it allows Unified CM to authenticate users immediately without having to follow referrals. For these cases, point Unified CM to a Global Catalog server and set the LDAP User Search Base to the top of the root domain.

In the case of a Microsoft AD forest that encompasses multiple trees, some additional considerations apply. Because a single LDAP search base cannot cover multiple namespaces, Unified CM must use a different mechanism to authenticate users across these discontiguous namespaces.

As mentioned in the section on LDAP Synchronization, page 16-9, in order to support synchronization with an AD forest that has multiple trees, the UserPrincipalName (UPN) attribute must be used as the user ID within Unified CM. When the user ID is the UPN, the LDAP authentication configuration page within Unified CM Administration does not allow you to enter the LDAP Search Base field, but instead it displays the note, "LDAP user search base is formed using userid information."

In fact, the user search base is derived from the UPN suffix for each user, as shown in Figure 16-14. In this example, a Microsoft Active Directory forest consists of two trees, avvid.info and vse.lab. Because the same user name may appear in both trees, Unified CM has been configured to use the UPN to uniquely identify users in its database during the synchronization and authentication processes.

Figure 16-14 Authentication with Microsoft AD Forests with Multiple Trees



As shown in Figure 16-14, a user named John Doe exists in both the avvid.info tree and the vse.lab tree. The following steps illustrate the authentication process for the first user, whose UPN is jdoe@avvid.info:

1. The user authenticates to Unified CM via HTTPS with its user name (which corresponds to the UPN) and password.
2. Unified CM performs an LDAP query against a Microsoft Active Directory Global Catalog server, using the user name specified in the UPN (anything before the @ sign) and deriving the LDAP search base from the UPN suffix (anything after the @ sign). In this case, the user name is jdoe and the LDAP search base is "dc=avvid, dc=info".
3. Microsoft Active Directory identifies the correct Distinguished Name corresponding to the user name in the tree specified by the LDAP query. In this case, "cn=jdoe, ou=Users, dc=avvid, dc=info".
4. Microsoft Active Directory responds via LDAP to Unified CM with the full Distinguished Name for this user.
5. Unified CM attempts an LDAP bind with the Distinguished Name provided and the password initially entered by the user, and the authentication process then continues as in the standard case shown in Figure 16-13.

Note Support for LDAP authentication with Microsoft AD forests containing multiple trees relies exclusively on the approach described above. Therefore, support is limited to deployments where the UPN suffix of a user corresponds to the root domain of the tree where the user resides. AD allows the use of aliases, which allows a different UPN suffix. If the UPN suffix is disjointed from the actual namespace of the tree, it is not possible to authenticate Unified CM users against the entire Microsoft Active Directory forest. (It is, however, still possible to use a different attribute as user ID and limit the integration to a single tree within the forest.)

User Filtering for Directory Synchronization and Authentication

Unified CM provides an LDAP Query Filter to optimize directory synchronization performance. Cisco recommends importing only those directory user accounts that will be assigned to Unified Communications resources in each individual cluster. When the number of directory user accounts exceeds the number supported for an individual cluster, filtering must be used to select the subset of users that will be associated on that cluster. The Unified CM synchronization feature is not meant to replace a large-scale corporate directory.

In many cases, a unique search base is all that is needed to control which accounts are synchronized. When a unique search base is not available, a custom LDAP filter might be required. The information in the following sections addresses both methods that can be used to optimize directory synchronization. When any mechanism is used to limit the accounts imported into Unified CM, the default directory lookup configuration will list only those directory entries that exist in the Unified CM database. For directory lookup to access the entire directory, you must configure Unified CM to utilize an external web server. Details of this configuration are not discussed here but are discussed in the Unified CM product documentation available at

http://www.cisco.com/en/US/partner/products/sw/voicesw/ps556/tsd_products_support_series_home.html

Optimizing Unified CM Database Synchronization

The Unified CM Database Synchronization feature provides a mechanism for importing a subset of the user configuration data (attributes) from the LDAP directory store into the Unified CM publisher database. Once synchronization of a user account has occurred, the copy of each user's LDAP account information may then be associated to additional data required to enable specific Unified Communications features for that user. When authentication is also enabled, the user's credentials are used to bind to the LDAP store for password verification. The end user's password is never stored in the Unified CM database when enabled for synchronization and/or authentication.

User account information is cluster-specific. Each Unified CM publisher server maintains a unique list of those users receiving Unified Communications services from that cluster. Synchronization agreements are cluster-specific, and each publisher has its own unique copy of user account information. Only those users who will be assigned Unified Communications resources should be synchronized with Unified CM. The following is a partial list of common reasons why the entire set of users defined in the LDAP directory should not be imported into the Unified CM cluster:

- Importing users who will not be assigned Unified Communications resources can increase directory synchronization time.
- Importing users who will not be assigned Unified Communications resources can slow Unified CM searches and overall database performance.
- In many cases, the number of user accounts in the LDAP directory store far exceeds the total user capacity of the Unified CM database.

Unified CM has no enforced limit on the number of accounts that may be added to the system. Cisco recommends limiting the number of users to twice the supported number of endpoints. There might be cases where accounts are needed for applications, and some designs might require additional accounts.

Cisco recommends using the control mechanisms described here to minimize the number of user accounts imported, regardless of the LDAP database size. This will improve the speed of the first and subsequent periodic synchronizations and will also improve manageability of the user accounts.

Using the LDAP Structure to Control Synchronization

Many deployments of LDAP directories use the Organizational Unit Name (OU) to group users into a logical order and sometimes hierarchical order. If the LDAP directory has a structure that organizes users into multiple OUs, then it often is possible to use that structure to control the groups of users imported. Each individual Unified CM synchronization agreement specifies a single OU. All active accounts under the specified OU, even within sub-OUs, are imported. Only those users in the OU are synchronized. When multiple OUs containing users are required in a cluster, multiple synchronization agreements are required. When an OU contains users that will not be assigned Unified Communications resources, Cisco recommends omitting those OUs from the directory synchronization.

The same technique may be used with AD, which defines containers. A synchronization agreement may specify a particular container in the directory tree and thereby limit the extent of the import.

Because there are only five synchronization agreements available, LDAP deployments with many OUs or containers can quickly exhaust this technique. One possible method to synchronize users in a multi-OU environment is to control the permissions assigned to the synchronization service account. Configure the synchronization agreement to a tree node that contains a mix of users, and then restrict the system account from read access to selected parts of the subtree. Refer to your LDAP vendor documentation on how to restrict this access.

LDAP Query

Additional control over filtering might be required for any of the following reasons:

- The LDAP directory has a flat structure that does not enable adequate control by configuration of the synchronization agreements. When the aggregate number of users that are imported by all the synchronization agreements is greater than the maximum number of users supported by the Unified CM cluster, then it is necessary to control the number of users imported through filters.
- You want to import a subset of user accounts into the Unified CM cluster, for administrative segmentation of users, to control a subset of users that have access and authentication to the cluster. Any account that is imported into a cluster has some level of access to the web pages and authentication mechanisms, which might not be desirable in some cases.
- The LDAP directory structure does not have an accurate representation of how users are going to be mapped into the Unified CM clusters. For instance, if OUs are set up according to an organizational hierarchy but users are mapped to Unified CM by geography, there might be little overlap between the two.

In these cases, the LDAP Query filter may be used to provide additional control over the synchronization agreements.

LDAP Query Filter Syntax and Server-Side Filtering

Unified CM uses standard LDAP mechanisms for synchronizing data from an LDAP directory store. It utilizes the Search mechanism, as defined by RFC 2251–Lightweight Directory Access Protocol (v3), to send a request to retrieve data from the LDAP server. Also defined by that mechanism is the ability to specify a filter string inside the Search message that is used by the LDAP server to select entries in the database for which to return data. The syntax of the filter string is defined by RFC 2254, The String Representation of LDAP Search Filters. This RFC may be used as a reference for constructing more complex filter strings.

The filter string is embedded within a Search message that is sent by Unified CM to the LDAP server and is executed by the server to select which user accounts will be provided in the response.

Simple Filter Syntax

You can configure a filter by specifying standard attribute names and values that are desired for those attributes. The attributes may also be specified by DN element instead of name. The filter string that is used by Unified CM in LDAP queries is stored internally in the ldapfilter table and is the string inserted into the Search message.

A filter is a UTF-8 formatted string that has the following syntax:

(*attribute operator value***)**

or

(*operator***(***filter1***)(***filter2***))**

Where *filter1* and *filter2* have the syntax shown in first line, and the *operator* is one of those listed in Table 16-6. The *attribute* corresponds to an LDAP attribute that exists in the directory, *operator* is one of the operators listed in Table 16-6, and *value* corresponds to the actual data value that is requested for the attribute.

Table 16-6 Basic Filter String Operators

Operator	Meaning of Function
!	Logical NOT
&	Logical AND
|	Logical OR
*	Wildcard
=	Equal to
>=	Lexicographically greater than or equal to
<=	Lexicographically less than or equal to

An attribute specified in the filter can be any attribute that exists in the LDAP directory store, and it does not have to be one of the attributes that is understood and imported by Unified CM. The attribute is used only on the LDAP server to select data, and the corresponding entries will have a subset of their data imported into Unified CM.

Example 16-1 A Single Condition

```
(givenName=Jack)
```

The filter in Example 16-1 selects any user with a given name of Jack.

Example 16-2 Multiple Conditions May Be Joined with Logical Characters

```
(&(objectclass=user)(department=Engineering))
```

The filter in Example 16-2 selects all users in the engineering department.

Default Filter Strings

If no custom filter strings are defined, Unified CM uses a default LDAP filter string as follows:

- Default Active Directory (AD) filter string

  (&(objectclass=user)(!(objectclass=Computer))(!(UserAccountControl:1.2.840.113556.1.4.803:=2)))

  This default filter selects entries for which the object class is a user but not a computer, and for which the account is not flagged as disabled.

- Default SunOne or Netscape filter string

  (objectclass=inetOrgPerson)

  This default filter selects all users for which the object class is inetOrgPerson.

- Default OpenLDAP filter string

  (objectclass=inetOrgPerson)

- Default Active Directory Application Mode (ADAM) or Active Directory Lightweight Directory Services (AD LDS) filter string

  (&(objectclass=user)((objectclass=Computer))(!(msDS-UserAccountDisabled=TRUE)))

Extending the Default Filter

Cisco recommends that you use the default filter string and append additional conditions to it. For example:

(&(objectclass=user)(!(objectclass=Computer))(!(UserAccountControl:1.2.840.113556.1.4.803:=2))(telephonenumber=919*))

This filter selects only users that have a prefix of 919 in their telephonenumber field. The synchronization agreement will import only users with an area code of 919. This example assumes all entries begin with an area code.

For the search filter, you may use any existing attribute or even a custom attribute that is defined in the LDAP directory store. The filter string controls which records are selected by the LDAP server to be returned to Unified CM, but the attributes that are imported are not affected by the filter string.

Custom LDAP filter strings can be up to 2048 characters long. Custom LDAP filters first need to be created, and then existing custom LDAP filters can be assigned to LDAP synchronization agreements. Different LDAP synchronization agreements can use different custom LDAP filters.

High Availability

Unified CM LDAP Synchronization allows for the configuration of up to three redundant LDAP servers for each directory synchronization agreement. Unified CM LDAP Authentication allows for the configuration of up to three redundant LDAP servers for a single authentication agreement. You should configure a minimum of two LDAP servers for redundancy. The LDAP servers can be configured with IP addresses instead of host names to eliminate dependencies on Domain Name System (DNS) availability.

Capacity Planning for Unified CM Database Synchronization

The Unified CM Database Synchronization feature provides a mechanism for importing a subset of the user configuration data (attributes) from the LDAP store into the Unified CM publisher database. Once synchronization of a user account has occurred, the copy of each user's LDAP account information may then be associated to additional data required to enable specific Unified Communications features for that user. When authentication is also enabled, the user's credentials are used to bind to the LDAP store for password verification. The end user's password is never stored in the Unified CM database when enabled for synchronization and/or authentication.

User account information is cluster-specific. Each Unified CM publisher server maintains a unique list of those users receiving Unified Communications services from that cluster. Synchronization agreements are cluster-specific, and each publisher has its own unique copy of user account information.

The maximum number of users that a Unified CM cluster can handle is limited by the maximum size of the internal configuration database that gets replicated between the cluster members. Starting with Unified CM release 8.6(1), the maximum number of users that can be configured or synchronized was increased from 60,000 to 80,000. To optimize directory synchronization performance, Cisco recommends considering the following points:

- Directory lookup from phones and web pages may use the Unified CM database or the IP Phone Service SDK. When directory lookup functionality uses the Unified CM database, only users who were configured or synchronized from the LDAP store are shown in the directory. If a subset of users are synchronized, then only that subset of users are seen on directory lookup.
- When the IP Phone Services SDK is used for directory lookup, but authentication of Unified CM users to LDAP is needed, the synchronization can be limited to the subset of users who would log in to the Unified CM cluster.
- If only one cluster exists, and the LDAP store contains fewer than the maximum number of users supported by the Unified CM cluster, and directory lookup is implemented to the Unified CM database, then it is possible to import the entire LDAP directory.
- When multiple clusters exist and the number of users in LDAP is less than the maximum number of users supported by the Unified CM cluster, it is possible to import all users into every cluster to ensure directory lookup has all entries.
- If the number of user accounts in LDAP exceeds the maximum number of users supported by the Unified CM cluster and the entire user set should be visible to all users, it will be necessary to use the Unified IP Phone Services SDK to off-load the directory lookup from Unified CM.
- If both synchronization and authentication are enabled, user accounts that have either been configured or synchronized into the Unified CM database will be able to log in to that cluster. The decision about which users to synchronize will impact the decision on directory lookup support.

Note Cisco supports the synchronization of user accounts up to the limit mentioned above, but it does not enforce this limit. Synchronizing more user accounts can lead to starvation of disk space, slower database performance, and longer upgrade times.

CHAPTER 17

Media Resources

Revised: April 30, 2013; OL-27282-05

A media resource is a software-based or hardware-based entity that performs media processing functions on the data streams to which it is connected. Media processing functions include mixing multiple streams to create one output stream (conferencing), passing the stream from one connection to another (media termination point), converting the data stream from one compression type to another (transcoding), streaming music to callers on hold (music on hold), echo cancellation, signaling, voice termination from a TDM circuit (coding/decoding), packetization of a stream, streaming audio (annunciation), and so forth. The software-based resources are provided by the Unified CM IP Voice Media Streaming Service (IP VMS). Digital signal processor (DSP) cards provide both software and hardware based resources.

This chapter explains the overall Media Resources Architecture and Cisco IP Voice Media Streaming Application service, and it focuses on the following media resources:

Use this chapter to gain an understanding of the function and capabilities of each media resource type and to determine which resource would be required for your deployment.

For proper DSP sizing of Cisco Integrated Service Router (ISR) gateways, you can use the Cisco Unified Communications Sizing Tool (Unified CST), available to Cisco employees and partners at http://tools.cisco.com/cucst. If you are not a Cisco partner or employee, you can use the DSP Calculator at http://www.cisco.com/go/dspcalculator. For other Cisco non-ISR gateway platforms (such as the Cisco 1700, 2600, 3700, and AS5000 Series) and/or Cisco IOS releases preceding and up to 12.4 mainline, you can access the legacy DSP calculator at http://www.cisco.com/pcgi-bin/Support/DSP/cisco_dsp_calc.pl.

What's New in This Chapter

Table 17-1 lists the topics that are new in this chapter or that have changed significantly from previous releases of this document.

Table 17-1 New or Changed Information Since the Previous Release of This Document

New or Revised Topic	Described in:	Revision Date
SIP Delayed Offer trunks	DTMF Relay over SIP Trunks, page 17-14	April 30, 2013
Media resource group (MRG) allocation	Media Resource Groups and Lists, page 17-38	October 31, 2012
Music on hold (MoH) capacity planning	Capacity Planning for Music on Hold, page 17-36	August 31, 2012
No changes for Cisco Unified Communications System Release 9.0		June 28, 2012

Media Resources Architecture

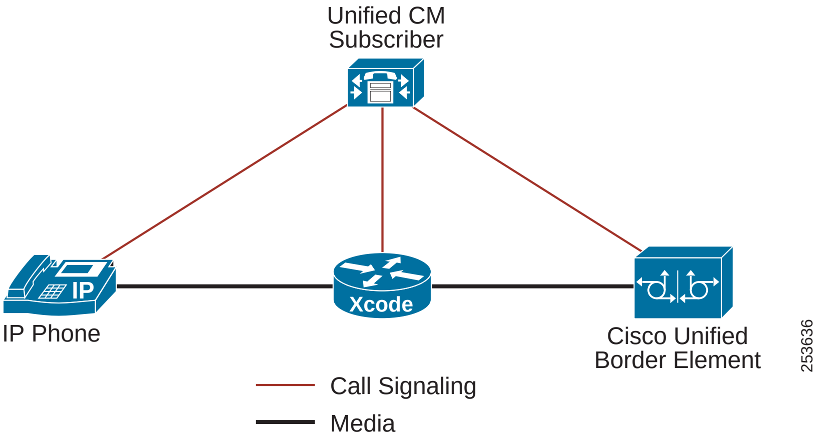
To properly design the media resource allocation strategy for an enterprise, it is critical to understand the Cisco Unified CM architecture for the various media resource components. The following sections highlight the important characteristics of media resource design with Unified CM.

Media Resource Manager

The Media Resource Manager (MRM), a software component in the Unified CM, determines whether a media resource needs to be allocated and inserted in the media path. This media resource may be provided by the Unified CM IP Voice Media Streaming Application service or by digital signal processor (DSP) cards. When the MRM decides and identifies the type of the media resource, it searches through the available resources according to the configuration settings of the media resource group list (MRGL) and media resource groups (MRGs) associated with the devices in question. MRGLs and MRGs are constructs that hold related groups of media resources together for allocation purposes and are described in detail in the section on Media Resource Groups and Lists, page 17-38.

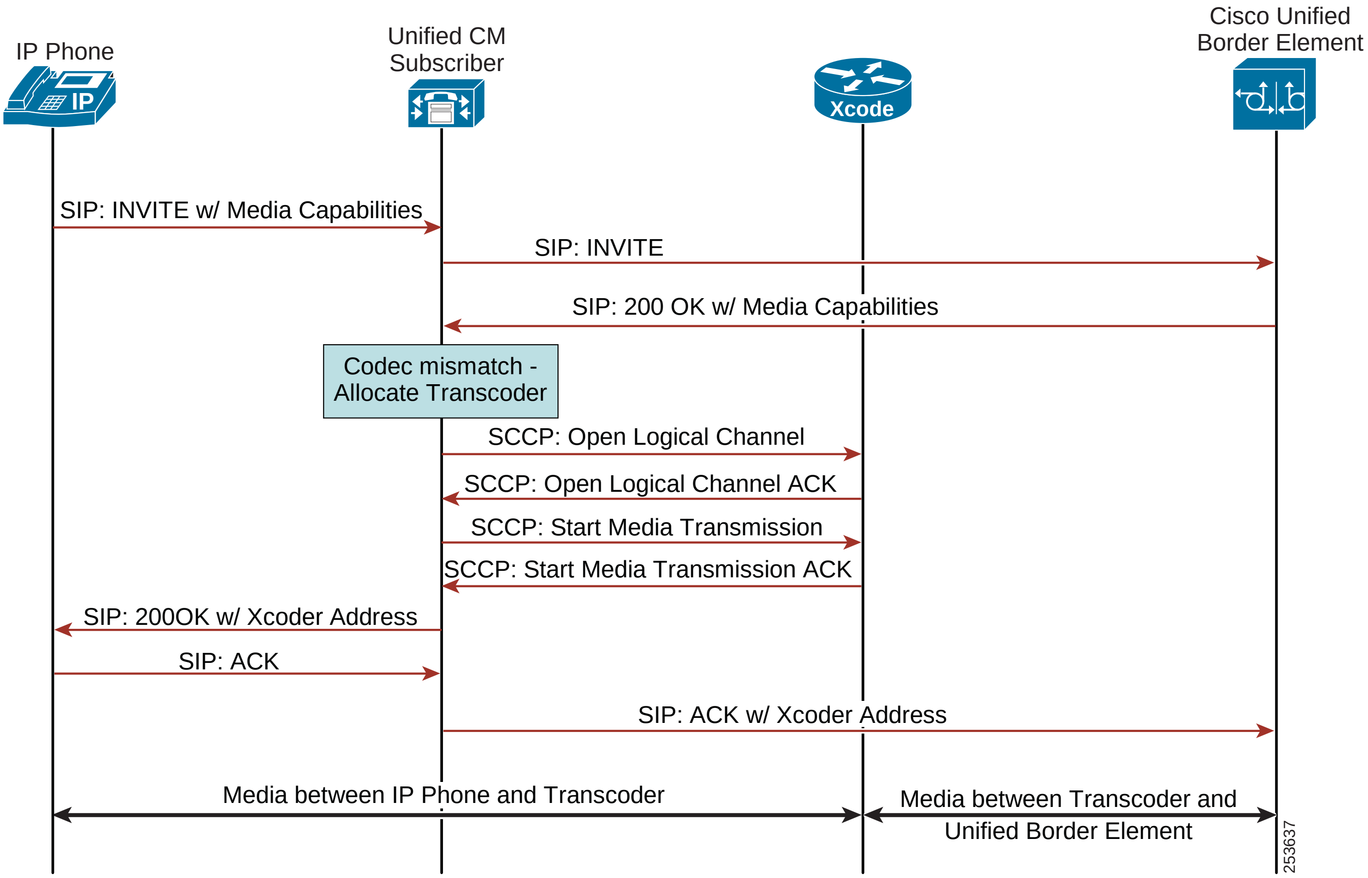
Figure 17-1 shows how a media resource such as a transcoder may be placed in the media path between an IP phone and a Cisco Unified Border Element when a common codec between the two is not available.

Figure 17-1 Use of a Transcoder Where a Common Codec Is Not Available



Unified CM communicates with media resources using Skinny Client Control Protocol (SCCP). This messaging is independent of the protocol that might be in use between Unified CM and the communicating entities. Figure 17-2 shows an example of the message flow, but it does not show all of the SCCP or SIP messages exchanged between the entities.

Figure 17-2 Message Flow Between Components


IP Phone
IP
Unified CM Subscriber
Xcode
Cisco Unified Border Element
SIP: INVITE w/ Media Capabilities
SIP: INVITE
SIP: 200 OK w/ Media Capabilities
Codec mismatch - Allocate Transcoder
SCCP: Open Logical Channel
SCCP: Open Logical Channel ACK
SCCP: Start Media Transmission
SCCP: Start Media Transmission ACK
SIP: 200OK w/ Xcoder Address
SIP: ACK
SIP: ACK w/ Xcoder Address
Media between IP Phone and Transcoder
Media between Transcoder and Unified Border Element
253637

Cisco IP Voice Media Streaming Application

The Cisco IP Voice Media Streaming Application provides the following software-based media resources:

- Conference bridge
- Music on Hold (MoH)
- Annunciator
- Media termination point (MTP)

The details of these resources are covered in the respective sections below.

When the IP Voice Media Streaming Application is activated, one of each of the above resources is automatically configured. Conferencing, annunciator, and MTP services can be disabled if required. If these resources are not needed, Cisco recommends that you disable them by modifying the appropriate service parameter in the Unified CM configuration. The service parameters have default settings for the maximum number of connections that each service can handle. For details on how to modify the service parameters, refer to the appropriate version of the *Cisco Unified Communications Manager Administration Guide*, available at

http://www.cisco.com/en/US/products/sw/voicesw/ps556/prod_maintenance_guides_list.html

Give careful consideration to situations that require multiple resources and to the load they place on the IP Voice Media Streaming Application. The media resources can reside on the same server as Unified CM or on a dedicated server not running the Unified CM call processing service. If your deployment requires more than the default number of any resource, Cisco recommends that you configure that resource to run on its own dedicated server. If heavy use of media resources is expected within a deployment, Cisco recommends deploying dedicated Unified CM media resource nodes (non-publisher nodes that do not perform call processing within the cluster) or relying on hardware-based media resources. Software-based media resources on Unified CM nodes are intended for small deployments or deployments where need for media resources is limited.

Note Cisco Business Edition 3000 provides only MoH and annunciator software-based media resources. No software-based media resources are available for conferencing and MTP. Hardware-based conferencing and MTP resources are provided by the DSP cards on board the Cisco MCS 7890 C2 platform.

Voice Termination

Voice termination applies to a call that has two call legs, one leg on a time-division multiplexing (TDM) interface and the second leg on a Voice over IP (VoIP) connection. The TDM leg must be terminated by hardware that performs encoding/decoding and packetization of the stream. This termination function is performed by a digital signal processor (DSP) resource residing in the same hardware module, blade, or platform.

All DSP hardware on Cisco TDM gateways is capable of terminating voice streams, and certain hardware is also capable of performing other media resource functions such as conferencing or transcoding (see Conferencing, page 17-6 and Transcoding, page 17-9). The DSP hardware has either fixed DSP resources that cannot be upgraded or changed, or modular DSP resources that can be upgraded.

The number of supported calls per DSP depends on the computational complexity of the codec used for a call and also on the complexity mode configured on the DSP. Cisco IOS enables you to configure a complexity mode on the hardware module. Hardware platforms such as the PVDM2 and PVDM3 DSPs support three complexity modes: medium, high and flex mode. Some of the other hardware platforms support only medium and high complexity modes.

Medium and High Complexity Mode

You can configure each DSP separately as either medium complexity, high complexity, or flex mode (PVDM3 DSPs and those based on C5510). The DSP treats all calls according to its configured complexity, regardless of the actual complexity of the codec of the call. A resource with configured complexity equal or higher than the actual complexity of the incoming call must be available, or the call will fail. For example, if a call requires a high-complexity codec but the DSP resource is configured for medium complexity mode, the call will fail. However, if a medium-complexity call is attempted on a DSP configured for high complexity mode, then the call will succeed and Cisco IOS will allocate a high-complexity mode resource.

Flex Mode

Flex mode, available on hardware platforms that use the C5510 chipset and on PVDM3 DSPs, eliminates the requirement to specify the codec complexity at configuration time. A DSP in flex mode accepts a call of any supported codec type, as long as it has available processing power.

For C5510-based DSPs, the overhead of each call is tracked dynamically via a calculation of processing power in millions of instructions per second (MIPS). Cisco IOS performs a MIPS calculation for each call received and subtracts MIPS credits from its budget whenever a new call is initiated. The number of MIPS consumed by a call depends on the codec of the call. The DSP will allow a new call as long as it has remaining MIPS credits greater than or equal to the MIPS required for the incoming call.

Similarly, PVDM3 DSP modules use a credit-based system. Each module is assigned a fixed number of "credits" that represent a measure of its capacity to process media streams. Each media operation, such as voice termination, transcoding, and so forth, is assigned a cost in terms of credits. As DSP resources are allocated for a media processing function, its cost value is subtracted from the available credits. A DSP module runs out of capacity when the available credits run out and are no longer sufficient for the requested operation. The credit allocation rules for PVDM3 DSPs are rather complex.

For proper DSP sizing of Cisco ISR gateways, you can use the Cisco Unified Communications Sizing Tool (Unified CST), available to Cisco employees and partners at http://tools.cisco.com/cucst. If you are not a Cisco partner, you can use the DSP Calculator at http://www.cisco.com/go/dspcalculator. For other Cisco non-ISR gateway platforms (such as the Cisco 1700, 2600, 3700, and AS5000 Series) and/or Cisco IOS releases preceding and up to 12.4 mainline, you can access the legacy DSP calculator at http://www.cisco.com/cgi-bin/Support/DSP/cisco_dsp_calc.pl.

Flex mode has an advantage when calls of multiple codecs must be supported on the same hardware because flex mode can support more calls than when the DSPs are configured as medium or high complexity. However, flex mode does allow oversubscription of the resources, which introduces the risk of call failure if all resources are used. With flex mode it is possible to have fewer DSP resources than with physical TDM interfaces.

Compared to medium or high complexity mode, flex mode has the advantage of supporting the most G.711 calls per DSP. For example, a PVDM2-16 DSP can support 8 G.711 calls in medium complexity mode or 16 G.711 calls in flex mode.

Conferencing

A conference bridge is a resource that joins multiple participants into a single call (audio or video). It can accept any number of connections for a given conference, up to the maximum number of streams allowed for a single conference on that device. There is a one-to-one correspondence between media streams connected to a conference and participants connected to the conference. The conference bridge mixes the streams together and creates a unique output stream for each connected party. The output stream for a given party is the composite of the streams from all connected parties minus their own input stream. Some conference bridges mix only the three loudest talkers on the conference and distribute that composite stream to each participant (minus their own input stream if they are one of the talkers).

Audio Conferencing

Audio conferencing can be performed by both software-based and hardware-based conferencing resources. A hardware conference bridge has all the capabilities of a software conference bridge. In addition, some hardware conference bridges can support multiple low bit-rate (LBR) stream types such as G.729 or G.723. This capability enables some hardware conference bridges to handle mixed-mode conferences. In a mixed-mode conference, the hardware conference bridge transcodes G.729 and G.723 streams into G.711 streams, mixes them, and then encodes the resulting stream into the appropriate stream type for transmission back to the user. Some hardware conference bridges support only G.711 conferences.

All conference bridges that are under the control of Cisco Unified Communications Manager (Unified CM) use Skinny Client Control Protocol (SCCP) to communicate with Unified CM.

Unified CM allocates a conference bridge from a conferencing resource that is registered with the Unified CM cluster. Both hardware and software conferencing resources can register with Unified CM at the same time, and Unified CM can allocate and use conference bridges from either resource. Unified CM does not distinguish between these types of conference bridges when it processes a conference allocation request.

The number of individual conferences that may be supported by the resource varies, and the maximum number of participants in a single conference varies, depending on the resource.

The following types of conference bridge resources may be used on a Unified CM system:

- Software Audio Conference Bridge (Cisco IP Voice Media Streaming Application), page 17-6
- Hardware Audio Conference Bridge (Cisco NM-HDV2, NM-HD-1V/2V/2VE, PVDM2, and PVDM3 DSPs), page 17-7
- Hardware Audio Conference Bridge (Cisco WS-SVC-CMM-ACT), page 17-7
- Hardware Audio Conference Bridge (Cisco NM-HDV and 1700 Series Routers), page 17-7

Software Audio Conference Bridge (Cisco IP Voice Media Streaming Application)

A software unicast conference bridge is a standard conference mixer that is capable of mixing G.711 audio streams and Cisco Wideband audio streams. Any combination of Wideband or G.711 a-law and mu-law streams may be connected to the same conference. The number of conferences that can be supported on a given configuration depends on the server where the conference bridge software is running and on what other functionality has been enabled for the application. The Cisco IP Voice Media Streaming Application is a resource that can also be used for several functions, and the design must consider all functions together (see Cisco IP Voice Media Streaming Application, page 17-4).

Hardware Audio Conference Bridge (Cisco NM-HDV2, NM-HD-1V/2V/2VE, PVDM2, and PVDM3 DSPs)

DSPs that are configured through Cisco IOS as conference resources will load firmware into the DSPs that are specific to conferencing functionality only, and these DSPs cannot be used for any other media feature. Any PVDM2 or PVDM3 based hardware, such as the NM-HDV2, may be used simultaneously in a single chassis for voice termination but may not be used simultaneously for other media resource functionality. The DSPs based on PVDM-256K and PVDM2 have different DSP farm configurations, and only one may be configured in a router at a time. DSPs on PVDM2 hardware are configured individually as voice termination, conferencing, media termination, or transcoding, so that DSPs on a single PVDM may be used as different resource types. Allocate DSPs to voice termination first, then to other functionality as needed.

Starting with Cisco IOS Release 12.4(15)T, the limit on the maximum number of participants has been increased to 32. A conference based on these DSPs can be configured to have a maximum of 8, 16, or 32 participants. The DSP resources for a conference are reserved during configuration, based on the profile attributes and irrespective of how many participants actually join. Refer to the following module data sheets for accurate information on module capacity and capabilities:

- For capacity information on PVDM2 modules, refer to the *High-Density Packet Voice Digital Signal Processor Module for Cisco Unified Communications Solutions* data sheet, available at

  http://www.cisco.com/en/US/prod/collateral/routers/ps5854/product_data_sheet0900aecd8016e845_ps3115_Products_Data_Sheet.html

- For capacity information on PVDM3 modules, refer to the *High-Density Packet Voice Video Digital Signal Processor Module for Cisco Unified Communications Solutions* data sheet, available at

  http://www.cisco.com/en/US/prod/collateral/modules/ps3115/data_sheet_c78-553971.html

Note The integrated gateway on the Cisco MCS 7890 C2 platform for Cisco Business Edition 3000 supports up to 24 conference streams.

Hardware Audio Conference Bridge (Cisco WS-SVC-CMM-ACT)

The following guidelines and considerations apply to this DSP resource:

- DSPs on this hardware are configured individually as voice termination, conferencing, media termination, or transcoding, so that DSPs on a single module may be used as different resource types. Allocate DSPs to voice termination first.
- Each ACT Port Adaptor contains 4 DSPs that are individually configurable. Each DSP can support 32 conference participants. You can configure up to 4 ACT Port Adaptors per CMM Module.
- This Cisco Catalyst-based hardware provides DSP resources that can provide conference bridges of up to 128 participants per bridge. A conference bridge may span multiple DSPs on a single ACT Port Adaptor; but conference bridges cannot span across multiple ACT Port Adaptors.
- The G.711 and G.729 codecs are supported on these conference bridges without extra transcoder resources. However, transcoder resources would be necessary if other codecs are used.

Hardware Audio Conference Bridge (Cisco NM-HDV and 1700 Series Routers)

The following guidelines and considerations apply to these DSP resources:

- This hardware utilizes the PVDM-256K type modules that are based on the C549 DSP chipset.
- Conferences using this hardware provide bridges that allow up to 6 participants in a single bridge.
- The resources are configured on a per-DSP basis as conference bridges.

- The NM-HDV may have up to 5 PVDM-256K modules, while the Cisco 1700 Series Routers may have 1 or 2 PVDM-256K modules.
- Each DSP provides a single conference bridge that can accept G.711 or G.729 calls.
- The Cisco 1751 is limited to 5 conference calls per chassis, and the Cisco 1760 can support 20 conference calls per chassis.

Note Any PVDM2-based hardware, such as the NM-HDV2, may be used simultaneously in a single chassis for voice termination but may not be used simultaneously for other media resource functionality. The DSPs based on PVDM-256K and PVDM2 have different DSP farm configurations, and only one may be configured in a router at a time.

Video Conferencing

Video-capable endpoints provide the capability to conduct video conferences that function similar to audio conferences. Video conferences can be invoked as ad-hoc conferences from a Skinny Client Control Protocol (SCCP) device through the use of Conf, Join, or cBarge softkeys.

The video portion of the conference can operate in either of two modes:

- Voice activation

  In this mode, the video endpoints display the dominant participant (the one speaking most recently or speaking the loudest). In this way, the video portion follows or tracks the audio portion. This mode is optimal when one participant speaks most of the time, as with an instructor teaching or training a group.

- Continuous presence

  In this mode, input from all (or selected) video endpoints is displayed simultaneously and continuously. The audio portion of the conference follows or tracks the dominant speaker. Continuous presence is more popular, and it is optimal for conferences or discussions between speakers at various sites.

Videoconferencing resources are of two types:

- Software videoconferencing bridges

  Software videoconferencing bridges process video and audio for the conference using just software. Cisco Unified MeetingPlace Express Media Server is a software videoconferencing bridge that can support ad-hoc video conferences. Cisco Unified MeetingPlace Express Media Server supports only voice activation mode for video conferences.

- Hardware videoconferencing bridges

  Hardware videoconferencing bridges have hardware DSPs that are used for the video conferences. The Cisco 3500 Series Multipoint Control Units (MCUs) and, starting with Cisco IOS Release 15.1.4M, the PVDM3 DSPs provide this type of videoconferencing bridge. Most hardware videoconferencing bridges can also be used as audio-only conference bridges. Hardware videoconferencing bridges provide the advantages of video transrating, higher video resolution, and scalability.

Videoconferencing bridges can be configured in a manner similar to audio conferencing resources, with similar characteristics for media resource groups (MRGs) and media resource group lists (MRGLs) for the device pools or endpoints.

Cisco Unified CM includes the Intelligent Bridge Selection feature, which provides a method for selecting conference resources based on the capabilities of the endpoints in the conference. For additional details on this functionality, see Intelligent Bridge Selection, page 12-16.

Secure Conferencing

Secure conferencing is a way to use regular conferencing to ensure that the media for the conference is secure and cannot be compromised. There are various security levels that a conference can have, such as authenticated or encrypted. With secure conferencing, the devices and conferencing resource can be authenticated to be trusted devices, and the conference media can then be encrypted so that every authenticated participant sends and received encrypted media for that conference. In most cases the security level of the conference will depend on the lowest security level of the participants in the conference. For example, if there is one participant who is not using a secure endpoint, then the entire conference will be non-secure. As another example, if one of the endpoints is authenticated but does not do encryption, then the conference will be in authenticated mode.

Secure conferencing provides conferencing functionality at an enhanced security level and prevents unauthorized capture and decryption of conference calls.

Consider the following factors when designing secure conferencing:

- Security levels of devices (phones and conferencing resources)
- Security overhead for call signaling and secure (SRTP) media
- Bandwidth utilization impact if secure participants are across the WAN
- Any intermediate devices such as NAT and firewalls that might not support secure calls across them

Secure conferencing is subject to the following restrictions and limitations:

- Secure conferencing is supported only for audio conferencing; video conferencing is not supported.
- With secure conferencing, Cisco IOS DSPs support a maximum of 8 participants in a conference.
- Secure conferencing may also use more DSP resources than non-secure conferencing, so DSPs must be provisioned according to the DSP Calculator.
- Some protocols may rely on IPSec to secure the call signaling.
- Secure conferencing cannot be cascaded between Unified CM and Unified CM Express.
- MTPs and transcoders do not support secure calls. Therefore, a conference might no longer be secure if any call into that conference invokes an MTP or a transcoder.
- An elaborate security policy might be needed.
- Secure conferencing might not be available for all codecs.

Transcoding

A transcoder is a device that converts an input stream from one codec into an output stream that uses a different codec. Starting with Cisco IOS Release 15.0.1M, a transcoder also supports transrating, whereby it connects two streams that utilize the same codec but with a different packet size.

Transcoding from G.711 to any other codec is referred to as traditional transcoding. Transcoding between any two non-G.711 codecs is called universal transcoding and requires Universal Cisco IOS transcoders. Universal transcoding is supported starting with Cisco IOS Release 12.4.20T. Universal transcoding has a lower DSP density than traditional transcoding.

In a Unified CM system, the typical use of a transcoder is to convert between a G.711 voice stream and the low bit-rate compressed voice stream G729a. The following cases determine when transcoder resources are needed:

- Single codec for the entire system

  A single codec is generally used in a single-site deployment that usually has no need for conserving bandwidth. When a single codec is configured for all calls in the system, then no transcoder resources are required. In this scenario, G.711 is the most common choice that is supported by all vendors.

- Multiple codecs in use in the system, with all endpoints capable of all codec types

  The most common reason for multiple codecs is to use G.711 for LAN calls to maximize the call quality and to use a low-bandwidth codec to maximize bandwidth efficiency for calls that traverse a WAN with limited bandwidth. Cisco recommends using G.729a as the low-bandwidth codec because it is supported on all Cisco Unified IP Phone models as well as most other Cisco Unified Communications devices, therefore it can eliminate the need for transcoding. Although Unified CM allows configuration of other low-bandwidth codecs between regions, some phone models do not support those codecs and therefore would require transcoders. They would require one transcoder for a call to a gateway and two transcoders if the call is to another IP phone. The use of transcoders is avoided if all devices support and are configured for both G.711 and G.729 because the devices will use the appropriate codec on a call-by-call basis.

- Multiple codecs in use in the system, and some endpoints support or are configured for G.711 only

  This condition exists when G.729a is used in the system but there are devices that do not support this codec, or a device with G.729a support may be configured to not use it. In this case, a transcoder is also required. Devices from some third-party vendors may not support G.729.

Note Cisco Unified MeetingPlace Express prior to release 2.0 supported G.711 only. In an environment where G.729 is configured for a call into earlier versions of Cisco Unified MeetingPlace Express, transcoder resources are required.

A transcoder is also capable of performing the same functionality as a media termination point (MTP). In cases where transcoder functionality and MTP functionality are both needed, a transcoder is allocated by the system. If MTP functionality is required, Unified CM will allocate either a transcoder or an MTP from the resource pool, and the choice of resource will be determined by the media resource groups, as described in the section on Media Resource Groups and Lists, page 17-38.

To finalize the design, it is necessary to know how many transcoders are needed and where they will be placed. For a multi-site deployment, Cisco recommends placing a transcoder local at each site where it might be required. If multiple codecs are needed, it is necessary to know how many endpoints do not support all codecs, where those endpoints are located, what other groups will be accessing those resources, how many maximum simultaneous calls these device must support, and where those resources are located in the network.

Transcoding Resources

DSP resources are required to perform transcoding. Those DSP resources can be located in the voice modules and the hardware platforms for transcoding that are listed in the following sections.

Hardware Transcoder (Cisco NM-HDV2, NM-HD-1V/2V/2VE, and PVDM2 DSPs)

The number of sessions supported on each DSP is determined by the codecs used in universal transcoding mode. The following guidelines and considerations apply to these DSP resources:

- Transcoding is available between G.711 mu-law or a-law and G.729a, G.729ab, G.722, and iLBC. A single PVDM2-16 can support 8 sessions for transcoding between low and medium complexity codecs (such as G.711 and G.729a or G.722) or 6 sessions for transcoding between low and high complexity codecs (such as G.711 and G.729 or iLBC).

Note If transcoding is not required between G.711 and G.722, Cisco recommends that you do not include G.722 in the Cisco IOS configuration of the dspfarm profile. This is to preclude Unified CM from selecting G.722 as the codec for a call in which transcoding is required. DSP resources configured as Universal Transcoders are required for transcoding between G.722 and other codecs.

- Cisco Unified IP Phones use only the G.729a variants of the G.729 codecs. The default for a new DSP farm profile is G.729a/G.729ab/G.711u/G.711a. Because a single DSP can provide only one function at a time, the maximum sessions configured on the profile should be specified in multiples of 8 to prevent wasted resources.

For capacity information on PVDM2 modules, refer to the *High-Density Packet Voice Digital Signal Processor Module for Cisco Unified Communications Solutions* data sheet, available at

http://www.cisco.com/en/US/prod/collateral/routers/ps5854/product_data_sheet0900aecd8016e845_ps3115_Products_Data_Sheet.html

Hardware Transcoder (Cisco WS-SVC-CMM-ACT)

The following guidelines and considerations apply to this DSP resource:

- Transcoding is available between G.711 mu-law or a-law and G.729, G.729b, or G.723.
- There are 4 DSPs per ACT that may be allocated individually to DSP pools.
- The CCM-ACT can have 16 transcoded calls per DSP or 64 per ACT. The ACT reports resources as streams rather than calls, and a single transcoded call consists of two streams.

Hardware Transcoder (Cisco NM-HDV and 1700 Series Routers)

The following guidelines and considerations apply to these DSP resources:

This hardware utilizes the PVDM-256K type modules, and each DSP provides 2 transcoding sessions.

- The NM-HDV may have up to 4 PVDM-256K modules, and the Cisco 1700 Series Routers may have 1 or 2 PVDM-256K modules. The Cisco 1751 Router has a chassis limit of 16 sessions, and the Cisco 1760 Router has a chassis limit of 20 sessions.
- NM-HDV and NM-HDV2 modules may be used simultaneously in a single chassis for voice termination but may not be used simultaneously for other media resource functionality. Only one type of DSP farm configuration may be active at one time (either the NM-HDV or the HM-HDV2) for conferencing, MTP, or transcoding.

- Transcoding is supported from G.711 mu-law or a-law to any of G.729, G.729a, G.729b, or G.729ab codecs.

Hardware Transcoder (PVDM3 DSP)

PVDM3 DSPs are hosted by Cisco 2900 Series and 3900 Series Integrated Services Routers, and they support both secure and non-secure transcoding from any and to any codec. As with voice termination and conferencing, each transcoding session debits the available credits for each type of PVDM3 DSPs. The available credits determine the total capacity of the DSP.

For example, a PVDM3-16 can support 12 sessions for transcoding between low and medium complexity codecs (such as G.711 and G.729a or G.722) or 10 sessions for transcoding between low and high complexity codecs (such as G.711 and G.729 or iLBC).

Note For Cisco Business Edition 3000, the default gateway configuration will support only 10 transcoding sessions per Cisco MCS 7890 appliance.

For capacity information on PVDM3 modules, refer to the *High-Density Packet Voice Video Digital Signal Processor Module for Cisco Unified Communications Solutions* data sheet, available at

http://www.cisco.com/en/US/prod/collateral/modules/ps3115/data_sheet_c78-553971.html

Media Termination Point (MTP)

A media termination point (MTP) is an entity that accepts two full-duplex media streams. It bridges the streams together and allows them to be set up and torn down independently. The streaming data received from the input stream on one connection is passed to the output stream on the other connection, and vice versa. MTPs have many possible uses, such as:

- Re-Packetization of a Stream, page 17-12
- DTMF Conversion, page 17-12
- Protocol-specific usage
  - DTMF Relay over SIP Trunks, page 17-14
  - H.323 Supplementary Services, page 17-17
  - H.323 Outbound Fast Connect, page 17-17

Re-Packetization of a Stream

An MTP can be used to transcode G.711 a-law audio packets to G.711 mu-law packets and vice versa, or it can be used to bridge two connections that utilize different packetization periods (different sample sizes). Note that re-packetization requires DSP resources in a Cisco IOS MTP.

DTMF Conversion

DTMF tones are used during a call to signal to a far-end device for purposes of navigating a menu system, entering data, or other manipulation. They are processed differently than DTMF tones sent during a call setup as part of the call control. There are several methods for sending DTMF over IP, and two communicating endpoints might not support a common procedure. In these cases, Unified CM may dynamically insert an MTP in the media path to convert DTMF signals from one endpoint to the other. Unfortunately, this method does not scale because one MTP resource is required for each such call. The following sections help determine the optimum amount of MTP resources required, based on the combination of endpoints, trunks, and gateways in the system.

If Unified CM determines that an MTP needs to be inserted but no MTP resources are available, it uses the setting of the service parameter **Fail call if MTP allocation fails** to decide whether or not to allow the call to proceed.

DTMF Relay Between Endpoints

The following methods are used to relay DTMF from one endpoint to another.

Named Telephony Events (RFC 2833)

Named Telephony Events (NTEs) defined by RFC 2833 are a method of sending DTMF from one endpoint to another after the call media has been established. The tones are sent as packet data using the already established RTP stream and are distinguished from the audio by the RTP payload type field. For example, the audio of a call can be sent on a session with an RTP payload type that identifies it as G.711 data, and the DTMF packets are sent with an RTP payload type that identifies them as NTEs. The consumer of the stream utilizes the G.711 packets and the NTE packets separately.

Key Press Markup Language (RFC 4730)

The Key Press Markup Language (KPML) is defined in RFC 4730. Unlike NTEs, which is an in-band method of sending DTMF, KPML uses the signaling channel (out-of-band, or OOB) to send SIP messages containing the DTMF digits.

KPML procedures use a SIP SUBSCRIBE message to register for DTMF digits. The digits themselves are delivered in NOTIFY messages containing an XML encoded body.

Unsolicited Notify (UN)

Unsolicited Notify procedures are used primarily by Cisco IOS SIP Gateways to transport DTMF digits using SIP NOTIFY messages. Unlike KPML, these NOTIFY messages are unsolicited, and there is no prior registration to receive these messages using a SIP SUBSCRIBE message. But like KPML, Unsolicited Notify messages are out-of-band.

Also unlike KPML, which has an XML encoded body, the message body in these NOTIFY messages is a 10-character encoded digit, volume, and duration, describing the DTMF event.

H.245 Signal, H.245 Alphanumeric

H.245 is the media control protocol used in H.323 networks. In addition to its use in negotiating media characteristics, H.245 also provides a channel for DTMF transport. H.245 utilizes the signaling channel and, hence, provides an out-of-band (OOB) way to send DTMF digits. The Signal method carries more information about the DTMF event (such as its actual duration) than does Alphanumeric.

Cisco Proprietary RTP

This method sends DTMF digits in-band, that is, in the same stream as RTP packets. However, the DTMF packets are encoded differently than the media packets and use a different payload type. This method is not supported by Unified CM but is supported on Cisco IOS Gateways.

Skinny Client Control Protocol (SCCP)

The Skinny Client Control Protocol is used by Unified CM for controlling the various SCCP-based devices registered to it. SCCP defines out-of-band messages that transport DTMF digits between Unified CM and the controlled device.

DTMF Relay Between Endpoints in the Same Unified CM Cluster

The following rules apply to endpoints registered to Unified CM servers in the same cluster:

- Calls between two non-SIP endpoints do not require MTPs.

  All Cisco Unified Communications endpoints other than SIP send DTMF to Unified CM via various signaling paths, and Unified CM forwards the DTMF between dissimilar endpoints. For example, an IP phone may use SCCP messages to Unified CM to send DTMF, which then gets sent to an H.323 gateway via H.245 signaling events. Unified CM provides the DTMF forwarding between different signaling types.

- Calls between two Cisco SIP endpoints do not require MTPs.

  All Cisco SIP endpoints support NTE, so DTMF is sent directly between endpoints and no conversion is required.

- A combination of a SIP endpoint and a non-SIP endpoint might require MTPs.

  To determine the support for NTE in your devices, refer to the product documentation for those devices. Support of NTE is not limited to SIP and can be supported in devices with other call control protocols. Unified CM has the ability to allocate MTPs dynamically on a call-by-call basis, based on the capabilities of the pair of endpoints.

DTMF Relay over SIP Trunks

A SIP trunk configuration is used to set up communication with a SIP User Agent such as another Cisco Unified CM cluster or a SIP gateway.

SIP negotiates media exchange via Session Description Protocol (SDP), where one side offers a set of capabilities to which the other side answers, thus converging on a set of media characteristics. SIP allows the initial offer to be sent either by the caller in the initial INVITE message (Early Offer) or, if the caller chooses not to, the called party can send the initial offer in the first reliable response (Delayed Offer).

By default, Unified CM SIP trunks send the INVITE without an initial offer (Delayed Offer). Unified CM has two configurable options to enable a SIP trunk to send the offer in the INVITE (Early Offer):

- Media Termination Point Required

  Checking this option on the SIP trunk assigns an MTP for every outbound call. This statically assigned MTP supports only the G.711 codec or the G.729 codec, thus limiting media to voice calls only.

- Early Offer support for voice and video calls (insert MTP if needed)

  Checking this option on the SIP Profile associated with the SIP Trunk inserts an MTP only if the calling device cannot provide Unified CM with the media characteristics required to create the Early Offer (for example, where an inbound call to Unified CM is received on a Delayed Offer SIP trunk or a Slow Start H.323 trunk). This option is available only with Unified CM 8.5 and later releases.

In general, Cisco recommends **Early Offer support for voice and video calls (insert MTP if needed)** because this configuration option reduces MTP usage. Calls from older SCCP phones registered to Unified CM over SIP Early Offer trunks use an MTP to create the Offer SDP. These calls support voice, video, and encryption. Inbound calls to Unified CM from SIP Delayed Offer trunks or H.323 Slow Start trunks that are extended over a SIP Early Offer trunk use an MTP to create the Offer SDP. However, these calls support audio only in the initial call setup, but they can be escalated to support video mid-call if the called or calling device invokes it.

Also note that MTP resources are not required for incoming INVITE messages, whether or not they contain an initial offer.

Whether or not an MTP will be allocated by Unified CM depends on the capabilities of the communicating endpoints and the configuration on the intermediary device, if any. For example, the SIP trunk may be configured to handle DTMF exchange in one of several ways: a SIP trunk can carry DTMF using KPML or it can instruct the communicating endpoints to use NTE.

Note As described in this section, SIP Early Offer can also be enabled by checking the **Media Termination Point Required** option on the SIP trunk. However, this option increases MTP usage because an MTP is assigned for every outbound call rather than on an as-needed basis.

SIP Trunk MTP Requirements

By default, the SIP trunk parameter **Media Termination Point Required** and the SIP Profile parameter **Early Offer support for voice and video calls (insert MTP if needed)** are not selected.

Use the following steps to determine whether MTP resources are required for your SIP trunks.

1. Is the far-end SIP device defined by this SIP trunk capable of accepting an inbound call without a SIP Early Offer?

   If not, then on the SIP Profile associated with this trunk, check the box to enable **Early Offer support for voice and video calls (insert MTP if needed)**. For outbound SIP trunk calls, an MTP will be inserted only if the calling device cannot provide Unified CM with the media characteristics required to create the Early Offer, or if DTMF conversion is needed.

   If yes, then do not check the **Early Offer support for voice and video calls (insert MTP if needed)** box, and use Step 2. to determine whether an MTP is inserted dynamically for DTMF conversion. Note that DTMF conversion can be performed by the MTP regardless of the codec in use.

   Note The option for **Early Offer support for voice and video calls (insert MTP if needed)** is available only with Unified CM 8.5 and later releases.

2. Select a Trunk DTMF Signaling Method, which controls the behavior of DTMF selection on that trunk. Available MTPs will be allocated based on the requirements for matching DTMF methods for all calls.

   a. DTMF Signaling Method: No Preference

      In this mode, Unified CM attempts to minimize the usage of MTP by selecting the most appropriate DTMF signaling method.

      If both endpoints support NTE, then no MTP is required.

      If both devices support any out-of-band DTMF mechanism, then Unified CM will use KPML or Unsolicited Notify over the SIP trunk. For example, this is the case if a Cisco Unified IP Phone 7936 using SCCP (which supports DTMF using only SCCP messaging) communicates with a Cisco Unified IP Phone 7970 using SIP (which supports DTMF using NTE and KPML) over a SIP trunk configured as described above. The only case where MTP is required is when one of the endpoints supports out-of-band only and the other supports NTE only (for example, an SCCP Cisco Unified IP Phone 7936 communicating with a SIP Cisco Unified IP Phone 7970).

b. DTMF Signaling Method: RFC 2833

By placing a restriction on the DTMF signaling method across the trunk, Unified CM is forced to allocate an MTP if any one or both the endpoints do not support NTE. In this configuration, the only time an MTP will not be allocated is when both endpoints support NTE.

c. DTMF Signaling Method: OOB and RFC 2833

In this mode, the SIP trunk signals both KPML (or Unsolicited Notify) and NTE-based DTMF across the trunk, and it is the most intensive MTP usage mode. The only cases where MTP resources will not be required is when both endpoints support both NTE and any OOB DTMF method (KPML or SCCP).

Note Cisco Unified IP Phones play DTMF to the end user when DTMF is received via SCCP, but they do not play tones received by NTE. However, there is no requirement to send DTMF to another end user. It is necessary only to consider the endpoints that originate calls combined with endpoints that might need DTMF, such as PSTN gateways, application servers, and so forth.

DTMF Relay on SIP Gateways and Cisco Unified Border Element

Cisco SIP Gateways support KPML, NTE, or Unsolicited Notify as the DTMF mechanism, depending on the configuration. Because there may be a mix of endpoints in the system, multiple methods may be configured on the gateway simultaneously in order to minimize MTP requirements.

On Cisco SIP Gateways, configure both **sip-kpml** and **rtp-nte** as DTMF relay methods under SIP dial peers. This configuration will enable DTMF exchange with all types of endpoints, including those that support only NTE and those that support only OOB methods, without the need for MTP resources. With this configuration, the gateway will negotiate both NTE and KPML with Unified CM. If NTE is not supported by the Unified CM endpoint, then KPML will be used for DTMF exchange. If both methods are negotiated successfully, the gateway will rely on NTE to receive digits and will not subscribe to KPML.

Cisco SIP gateways also have the ability to use proprietary Unsolicited Notify (UN) method for DTMF. The UN method sends a SIP Notify message with a body that contains text describing the DTMF tone. This method is also supported on Unified CM and may be used if **sip-kpml** is not available. Configure **sip-notify** as the DTMF relay method. Note that this method is Cisco proprietary.

SIP gateways that support only NTE require MTP resources to be allocated when communicating with endpoints that do not support NTE.

H.323 Trunks and Gateways

For the H.323 gateways and trunks there are three reasons for invoking an MTP:

- H.323 Supplementary Services, page 17-17
- H.323 Outbound Fast Connect, page 17-17
- DTMF Conversion, page 17-17

H.323 Supplementary Services

MTPs can be used to extend supplementary services to H.323 endpoints that do not support the H.323v2 OpenLogicalChannel and CloseLogicalChannel request features of the Empty Capabilities Set (ECS). This requirement occurs infrequently. All Cisco H.323 endpoints support ECS, and most third-party endpoints have support as well. When needed, an MTP is allocated and connected into a call on behalf of an H.323 endpoint. When an MTP is required on an H.323 call and none is available, the call will proceed but will not be able to invoke supplementary services.

H.323 Outbound Fast Connect

H.323 defines a procedure called Fast Connect, which reduces the number of packets exchanged during a call setup, thereby reducing the amount of time for media to be established. This procedure uses Fast Start elements for control channel signaling, and it is useful when two devices that are utilizing H.323 have high network latency between them because the time to establish media depends on that latency. Unified CM distinguishes between inbound and outbound Fast Start based on the direction of the call setup, and the distinction is important because the MTP requirements are not equal. For inbound Fast Start, no MTP is required. Outbound calls on an H.323 trunk do require an MTP when Fast Start is enabled. Frequently, it is only inbound calls that are problematic, and it is possible to use inbound Fast Start to solve the issue without also enabling outbound Fast Start.

DTMF Conversion

An H.323 trunk supports the signaling of DTMF by means of H.245 out-of-band methods. H.323 intercluster trunks also support DTMF by means of NTE. There are no DTMF configuration options for H.323 trunks; Unified CM dynamically chooses the DTMF transport method.

The following scenarios can occur when two endpoints on different clusters are connected with an H.323 trunk:

- When both endpoints are SIP, then NTE is used. No MTP is required for DTMF.
- When one endpoint is SIP and supports both KPML and NTE, but the other endpoint is not SIP, then DTMF is sent as KPML from the SIP endpoint to Unified CM, and H.245 is used on the trunk. No MTP is required for DTMF.
- If one endpoint is SIP and supports only NTE but the other is not SIP, then H.245 is used on the trunk. An available MTP is allocated for the call. The MTP will be allocated on the Unified CM cluster where the SIP endpoint is located.

For example: A Cisco Unified IP Phone 7970 using SIP to communicate with a Cisco Unified IP Phone 7970 running SCCP, will use NTE when connected via a SIP trunk but will use OOB methods when communicating over an H.323 trunk (with the trunk using the H.245 method).

When a call is inbound from one H.323 trunk and is routed to another H.323 trunk, NTE will be used for DTMF when both endpoints are SIP. H.245 will be used if either endpoint is not SIP. An MTP will be allocated if one side is a SIP endpoint that supports only NTE and the other side is non-SIP.

DTMF Relay on H.323 Gateways and Cisco Unified Border Element

H.323 gateways support DTMF relay via H.245 Alphanumeric, H.245 Signal, NTE, and audio in the media stream. The NTE option must not be used because it is not supported on Unified CM for H.323 gateways at this time. The preferred option is H.245 Signal. MTPs are required for establishing calls to an H.323 gateway if the other endpoint does not have signaling capability in common with Unified CM. For example, a Cisco Unified IP Phone 7960 running the SIP stack supports only NTEs, so an MTP is needed with an H.323 gateway.

CTI Route Points

A CTI Route Point uses CTI events to communicate with CTI applications. For DTMF purposes, the CTI Route Point can be considered as an endpoint that supports all OOB methods and does not support RFC 2833. For such endpoints, the only instance where an MTP will be required for DTMF conversion would be when it is communicating with another endpoint that supports only RFC 2833.

CTI Route Points that have first-party control of a phone call will participate in the media stream of the call and require an MTP to be inserted. When the CTI has third-party control of a call so that the media passes through a device that is controlled by the CTI, then the requirement for an MTP is dependent on the capabilities of the controlled device.

Example 17-1 Call Flow that Requires an MTP for NTE Conversion

Assume the example system has CTI route points with first-party control (the CTI port terminates the media), which integrate to a system that uses DTMF to navigate an IVR menu. If all phones in the system are running SCCP, then no MTP is required. In this case Unified CM controls the CTI port and receives DTMF from the IP phones via SCCP. Unified CM provides DTMF conversion.

However, if there are phones running a SIP stack (that support only NTE and not KPML), an MTP is required. NTEs are part of the media stream; therefore Unified CM does not receive them. An MTP is invoked into the media stream and has one call leg that uses SCCP, and the second call leg uses NTEs. The MTP is under SCCP control by Unified CM and performs the NTE-to-SCCP conversion. Note that the newer phones that do support KPML will not need an MTP.

MTP Usage with a Conference Bridge

MTPs are utilized in a conference call when one or more participant devices in the conference use RFC 2833. When the conference feature is invoked, Unified CM allocates MTP resources for every conference participant device in the call that supports only RFC 2833. This is regardless of the DTMF capabilities of the conference bridge used.

MTP Resources

The following types of devices are available for use as an MTP:

Software MTP (Cisco IP Voice Media Streaming Application)

A software MTP is a device that is implemented by enabling the Cisco IP Voice Media Streaming Application on a Unified CM server. When the installed application is configured as an MTP application, it registers with a Unified CM node and informs Unified CM of how many MTP resources it supports. A software MTP device supports only G.711 streams. The IP Voice Media Streaming Application is a resource that may also be used for several functions, and the design guidance must consider all functions together (see Cisco IP Voice Media Streaming Application, page 17-4).

Software MTP (Based on Cisco IOS)

- The capability to provide a software-based MTP on the router is available beginning with Cisco IOS Release 12.3(11)T for the Cisco 3800 Series Routers; Release 15.0(1)M for the Cisco 2900 Series and 3900 Series Routers; Release IOS-XE for ASR1002, 1004, and 1006 Routers; Release IOS-XE 3.2 for ASR1001 Routers; and Release 12.3(8)T4 for other router models.
- This MTP allows configuration of any of the following codecs, but only one may be configured at a given time: G.711 mu-law and a-law, G.729a, G.729, G.729ab, G.729b, and passthrough. Some of these are not pertinent to a Unified CM implementation.
- Router configurations permit up to 1,000 individual streams, which support 500 transcoded sessions. This number of G.711 streams generates 10 Mbytes of traffic. The Cisco ISR G2s and ASR routers can support significantly higher numbers than this.

Hardware MTP (PVDM2, Cisco NM-HDV2 and NM-HD-1V/2V/2VE)

- This hardware uses the PVDM-2 modules for providing DSPs.
- Each DSP can provide 16 G.711 mu-law or a-law, 8 G.729a or G.722, or 6 G.729 or G.729b MTP sessions.

Hardware MTP (Cisco 2900 and 3900 Series Routers with PVDM3)

- These routers use the PVDM3 DSPs natively on the motherboards or PVDM2 with an adaptor on the motherboard or on service modules.
- The capacity of each of the DSP type varies from 16 G.711 a-law or mu-law sessions for the PVDM3-16 to 256 G.711 sessions for the PVDM3-256.

Note You cannot configure G.729 or G.729b codecs when configuring hardware MTP resources in Cisco IOS. However, Unified CM can use hardware transcoding resources as MTPs if all other MTP resources are exhausted or otherwise unavailable.

Trusted Relay Point

A Trusted Relay Point (TRP) is a device that can be inserted into a media stream to act as a control point for that stream. It may be used to provide further processing on that stream or as a method to ensure that the stream follows a specific desired path. There are two components to the TRP functionality, the logic utilized by Unified CM to invoke the TRP and the actual device that is invoked as the anchor point of the call. The TRP functionality can invoke an MTP device to act as that anchor point.

Unified CM provides a new configuration parameter for individual phone devices, which invokes a TRP for any call to or from that phone. The system utilizes the media resource pool mechanisms to manage the TRP resources. The media resource pool of that device must have an available device that will be invoked as a TRP.

See the chapter on Network Infrastructure, page 3-1, for an example of a use case for the TRP as a QoS enforcement mechanism, and see the chapter on Unified Communications Security, page 4-1, for an example of utilizing the TRP as an anchor point for media streams in a redundant data center with firewall redundancy.

Annunciator

An annunciator is a software function of the Cisco IP Voice Media Streaming Application that provides the ability to stream spoken messages or various call progress tones from the system to a user. It uses SCCP messages to establish RTP streams, and it can send multiple one-way RTP streams to devices such as Cisco IP phones or gateways. The device must be capable of SCCP to utilize this feature. SIP phones and devices are still able to receive all the various messages provided by the annunciator. For SIP devices, all these messages and tones are downloaded (pushed) to the device at registration so that they can be invoked as needed by SIP signaling messages from Unified CM.

Tones and announcements are predefined by the system. The announcements support localization and may also be customized by replacing the appropriate .wav file. The annunciator is capable of supporting G.711 a-law and mu-law, G.729, and Wideband codecs without any transcoding resources.

The following features require an annunciator resource:

- Cisco Multilevel Precedence Preemption (MLPP)

  This feature has streaming messages that it plays in response to the following call failure conditions.

  – Unable to preempt due to an existing higher-precedence call.

  – A precedence access limitation was reached.

  – The attempted precedence level was unauthorized.

  – The called number is not equipped for preemption or call waiting.

- Integration via SIP trunk

  SIP endpoints have the ability to generate and send tones in-band in the RTP stream. Because SCCP devices do not have this ability, an annunciator is used in conjunction with an MTP to generate or accept DTMF tones when integrating with a SIP endpoint. The following types of tones are supported:

  – Call progress tones (busy, alerting, and ringback)

  – DTMF tones

- Cisco IOS gateways and intercluster trunks

  These devices require support for call progress tone (ringback tone).

- System messages

  During the following call failure conditions, the system plays a streaming message to the end user:

  – A dialed number that the system cannot recognize

  – A call that is not routed due to a service disruption

  – A number that is busy and not configured for preemption or call waiting

- Conferencing

  During a conference call, the system plays a barge-in tone to announce that a participant has joined or left the bridge.

An annunciator is automatically created in the system when the Cisco IP Voice Media Streaming Application is activated on a server. If the Media Streaming Application is deactivated, then the annunciator is also deleted. A single annunciator instance can service the entire Unified CM cluster if it meets the performance requirements (see Annunciator Performance, page 17-21); otherwise, you must configure additional annunciators for the cluster. Additional annunciators can be added by activating the Cisco IP Voice Media Streaming Application on other servers within the cluster.

The annunciator registers with a single Unified CM at a time, as defined by its device pool. It will automatically fail over to a secondary Unified CM if a secondary is configured for the device pool. Any announcement that is playing at the time of an outage will not be maintained.

An annunciator is considered a media device, and it can be included in media resource groups (MRGs) to control which annunciator is selected for use by phones and gateways.

Annunciator Performance

By default, the annunciator is configured to support 48 simultaneous streams, which is the maximum recommended for an annunciator running on the same server (co-resident) with the Unified CM service. If the server has only 10 Mbps connectivity, lower the setting to 24 simultaneous streams.

A standalone server without the Cisco CallManager Service can support up to 255 simultaneous announcement streams, and a high-performance server with dual CPUs and a high-performance disk system can support up to 400 streams. You can add multiple standalone servers to support the required number of streams.

Cisco RSVP Agent

In order to provide topology-aware call admission control, Unified CM invokes one or two RSVP Agents during the call setup to perform an RSVP reservation across the IP WAN. These agents are MTP or transcoder resources that have been configured to provide RSVP functionality. RSVP resources are treated the same way as regular MTPs or transcoders from the perspective of allocation of an MTP or transcoder resource by Unified CM.

The Cisco RSVP Agent feature was first introduced in Cisco IOS Release 12.4(6)T. For details on RSVP and Cisco RSVP Agents, refer to the chapter on Call Admission Control, page 11-1.

Music on Hold

The Music on Hold (MoH) feature requires that each MoH server must be part of a Unified CM cluster and participate in the data replication schema. Specifically, the MoH server must share the following information with the Unified CM cluster through the database replication process:

- Audio sources - The number and identity of all configured MoH audio sources
- Multicast or unicast - The transport nature (multicast or unicast) configured for each of these sources
- Multicast address - The multicast base IP address of those sources configured to stream as multicast

To configure a MoH server, enable the Cisco IP Voice Media Streaming Application Service on one or more Unified CM nodes. An MoH server can be deployed along with Unified CM on the same server or in standalone mode.

Unicast and Multicast MoH

Unified CM supports unicast and multicast MoH transport mechanisms.

A unicast MoH stream is a point-to-point, one-way audio Real-Time Transport Protocol (RTP) stream from the MoH server to the endpoint requesting MoH. It uses a separate source stream for each user or connection. Thus, if twenty devices are on hold, then twenty streams are generated over the network between the server and these endpoint devices. Unicast MoH can be extremely useful in those networks where multicast is not enabled or where devices are not capable of multicast, thereby still allowing an administrator to take advantage of the MoH feature. However, these additional MoH streams can potentially have a negative effect on network throughput and bandwidth.

A multicast MoH stream is a point-to-multipoint, one-way audio RTP stream between the MoH server and the multicast group IP address. The endpoints requesting an MoH audio stream can join the multicast group as needed. This mode of MoH conserves system resources and bandwidth because it enables multiple users to use the same audio source stream to provide music on hold. For this reason, multicast is an extremely attractive transport mechanism for the deployment of a service such as MoH because it greatly reduces the CPU impact on the source device and also greatly reduces the bandwidth consumption for delivery over common paths. However, multicast MoH can be problematic in situations where a network is not enabled for multicast or where the endpoint devices are not capable of handling multicast.

There are distinct differences between unicast and multicast MoH in terms of call flow behavior. A unicast MoH call flow is initiated by a message from Unified CM to the MoH server. This message tells the MoH server to send an audio stream to the holdee device's IP address. On the other hand, a multicast MoH call flow is initiated by a message from Unified CM to the holdee device. This message instructs the endpoint device to join the multicast group address of the configured multicast MoH audio stream.

For a detailed look at MoH call flows, see the section on MoH Call Flows, page 17-27.

Supported Unicast and Multicast Gateways

The following gateways support both unicast and multicast MoH:

- Cisco 2900 Series and Cisco 3900/3900E Series ISR G2 Routers with Cisco IOS 15.0.1M or later release
- Cisco 2800 Series and 3800 Series Routers with MGCP or H.323 and Cisco IOS 12.3.14T or later release
- Cisco 2800 Series and 3800 Series Routers with SIP and Cisco IOS 12.4(24)T or later release
- Cisco VG224 Analog Voice Gateways with MGCP and Cisco IOS 12.3.14T or later release
- Cisco VG204 and VG202 Analog Voice Gateways with MGCP or SCCP and Cisco IOS 12.4(22)T or later release
- Cisco VG248 Analog Phone Gateways
- Cisco ASR 1000 Series Aggregation Services Routers

Note Cisco 2800 Series, 3800 Series, and VG248 gateways are End of Sale (EoS). There are other legacy gateways that also support unicast and multicast MoH.

Note The Cisco Unified Border Element on Cisco ASR 1000 Series Aggregation Services Routers might not support one-way streaming of music or announcements by the Cisco Unified Communications Manager Music on Hold (MoH) feature. For more information, refer to the release notes for your version of Cisco Unified Communications Manager, available at http://www.cisco.com/en/US/products/sw/voicesw/ps556/prod_release_notes_list.html.

MoH Selection Process

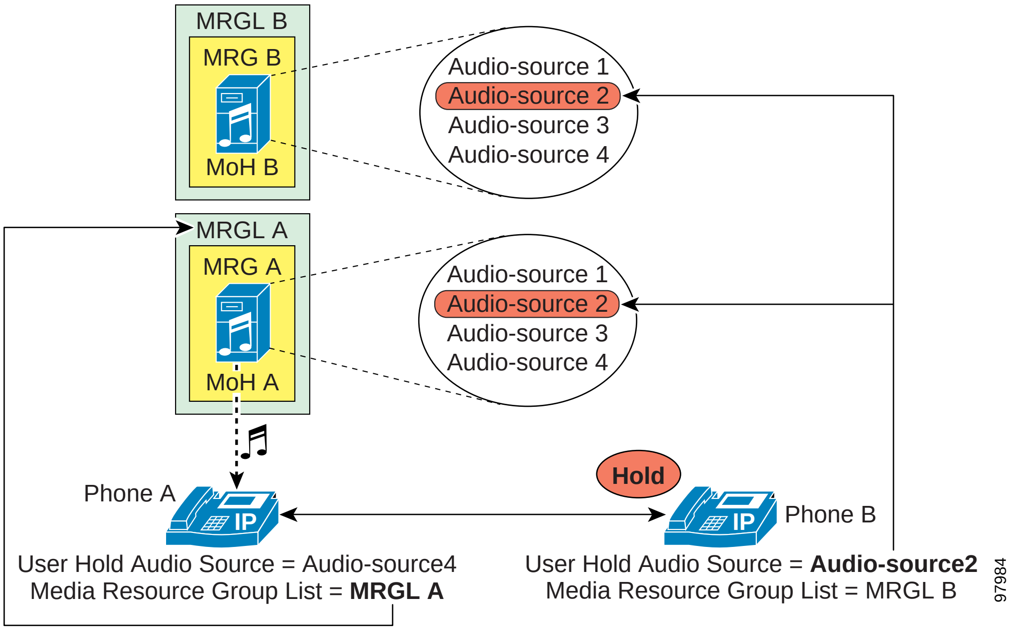
This section describes the MoH selection process as implemented in Unified CM.

The basic operation of MoH in a Cisco Unified Communications environment consists of a holder and a holdee. The *holder* is the endpoint user or network application placing a call on hold, and the *holdee* is the endpoint user or device placed on hold.

The MoH stream that an endpoint receives is determined by a combination of the User Hold MoH Audio Source of the device placing the endpoint on hold (holder) and the configured media resource group list (MRGL) of the endpoint placed on hold (holdee). The User Hold MoH Audio Source configured for the holder determines the audio file that will be streamed when the holder puts a call on hold, and the holdee's configured MRGL indicates the resource or server from which the holdee will receive the MoH stream.

As illustrated by the example in Figure 17-3, if phones A and B are on a call and phone B (holder) places phone A (holdee) on hold, phone A will hear the MoH audio source configured for phone B (Audio-source2). However, phone A will receive this MoH audio stream from the MRGL (resource or server) configured for phone A (MRGL A).

Figure 17-3 User Hold Audio Source and Media Resource Group List (MRGL)



Because the configured MRGL determines the server from which a unicast-only device will receive the MoH stream, you must configure unicast-only devices with an MRGL that points to a unicast MoH resource or media resource group (MRG). Likewise, a device capable of multicast should be configured with an MRGL that points to a multicast MRG containing a MoH server configured for multicast.

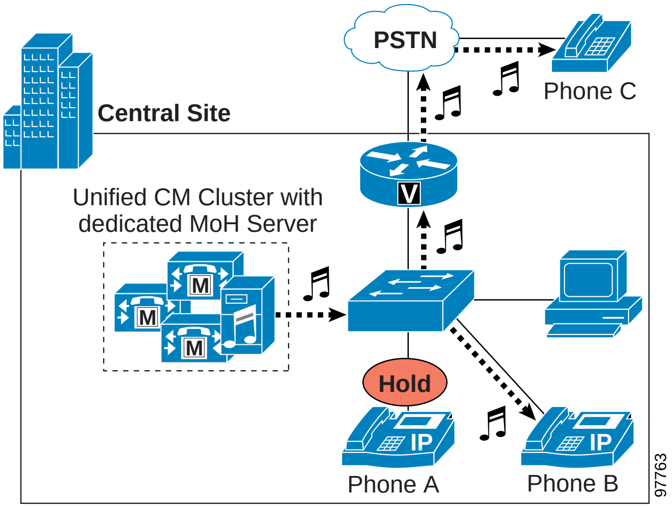
User and Network Hold

User hold includes the following types:

- User on hold at an IP phone or other endpoint device
- User on hold at the PSTN, where MoH is streamed to the gateway

Figure 17-4 shows these two types of call flows. If phone A is in a call with phone B and phone A (holder) pushes the Hold softkey, then a music stream is sent from the MoH server to phone B (holdee). The music stream can be sent to holdees within the IP network or holdees on the PSTN, as is the case if phone A places phone C on hold. In the case of phone C, the MoH stream is sent to the voice gateway interface and converted to the appropriate format for the PSTN phone. When phone A presses the Resume softkey, the holdee (phone B or C) disconnects from the music stream and reconnects to phone A.

Figure 17-4 Basic User Hold Example



Network hold can occur in following scenarios:

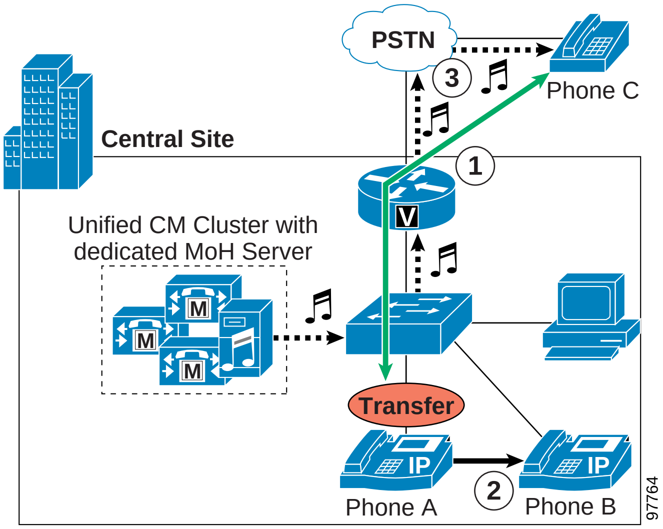
- Call transfer
- Call Park
- Conference setup
- Application-based hold

Figure 17-5 illustrates an example of network hold during a call transfer. The call flow involves the following steps:

1. Phone A receives a call from PSTN phone C.
2. Phone A answers the call and then transfers it to phone B. During the transfer process, phone C is put on network hold.
3. Phone C receives an MoH stream from the MoH server via the gateway. After phone A completes the transfer action, phone C disconnects from the music stream and gets redirected to phone B.

This process is the same for other network hold operations such as call park and conference setup.

Figure 17-5 Basic Network Hold Example for Call Transfer



MoH Sources

A Unified CM MoH server can generate a MoH stream from two types of sources, audio file and fixed source, either of which can be transmitted as unicast or multicast. You can configure a maximum of 51 MoH audio sources per Unified CM cluster, of which up to 50 can be audio files but only one can be a fixed source.

Audio File

Audio files (.wav format) can be uploaded to Unified CM, which then automatically generates MoH audio files for the specified codecs. Unified CM supports G711 (a-law and mu-law), G.729 Annex A, and Wideband codecs for MoH streams.

Note Before configuring a MoH audio source, you must upload the .wav formatted audio source file to every MoH server within the cluster using the upload file function in the Unified CM Administration interface. Cisco recommends that you first upload the audio source file onto each MoH server in the cluster, then upload it onto the publisher (even if not an MoH server), and finally assign an MoH Audio Stream Number and configure the MoH audio source in the Unified CM Administration interface on the publisher.

Fixed Source

If recorded or live audio is needed, MoH can be generated from a fixed source connected to the audio input of the local sound card. The Cisco MoH USB audio sound card (MOH-USB-AUDIO=) must be used for connecting a fixed or live audio source to the MoH server. This USB sound card is compatible with all Cisco MCS platforms that support Cisco Unified CM.

This mechanism enables you to use radios, CD players, or any other compatible sound source to stream MoH. The stream from the fixed audio source is transcoded in real-time to support the codec that was configured through Unified CM Administration. The fixed audio source can be transcoded into G.711 (A-law or mu-law), G.729 Annex A, and Wideband, and it is the only audio source that is transcoded in real-time.

Note Prior to using a fixed audio source to transmit music on hold, you should consider the legalities and the ramifications of re-broadcasting copyrighted audio materials. Consult your legal department for potential issues.

MoH Selection

To determine which User and Network Audio Source configuration setting to apply in a particular case, Unified CM interprets these settings for the *holder* device in the following priority order:

1. Directory or line setting (Devices with no line definition, such as gateways, do not have this level.)
2. Device setting
3. Common Device Configuration setting
4. Cluster-wide default setting

Unified CM also interprets the MRGL configuration settings of the *holdee* device in the following priority order:

1. Device setting
2. Device pool setting
3. System default MoH resources

Note that system default MoH resources are resources that are not assigned to any MRG and they are always unicast.

MoH Call Flows

The following sections provide detailed illustrations and explanations of unicast and multicast MoH call flows for both SCCP and SIP endpoints.

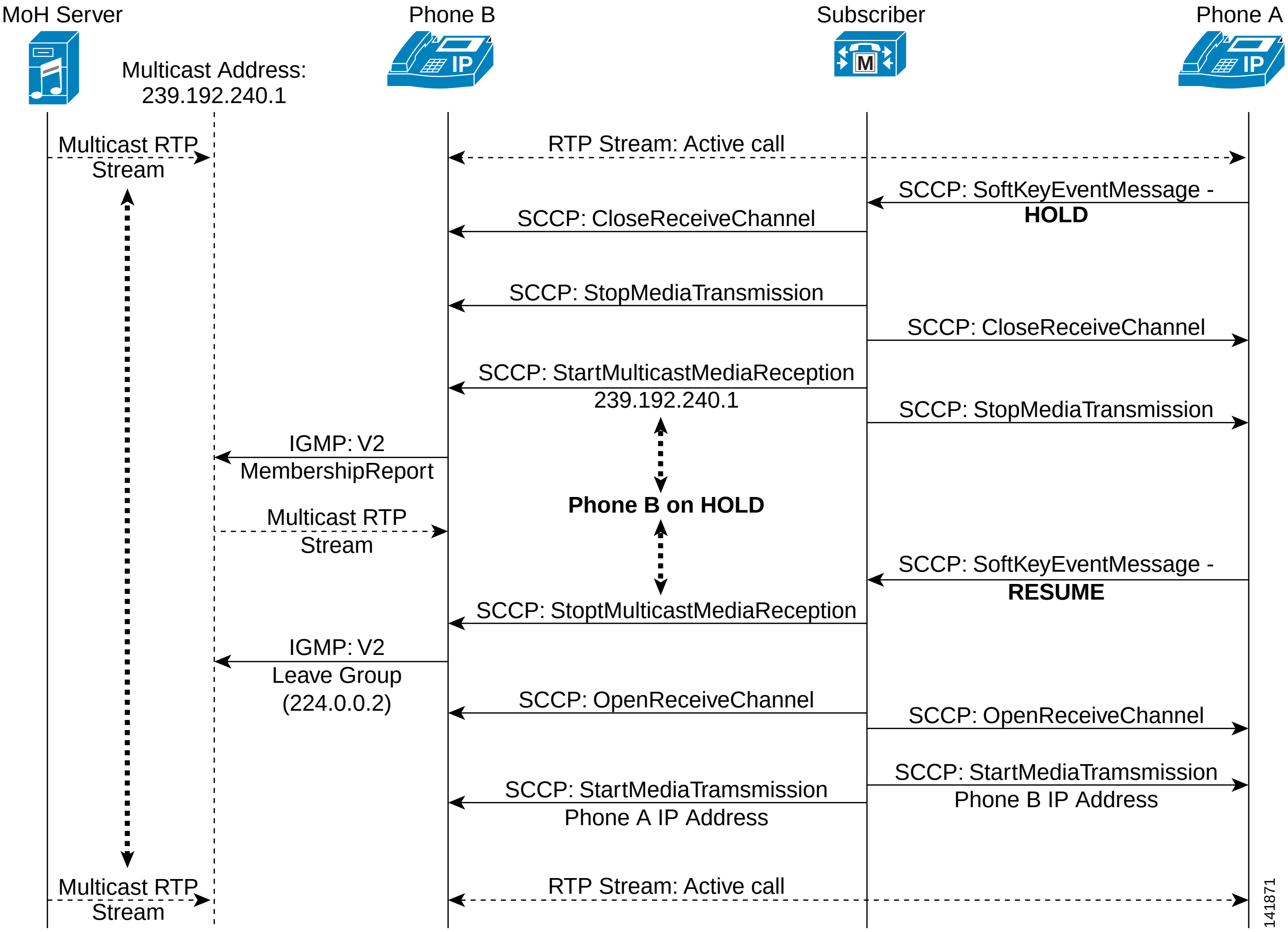
SCCP Call Flows

This section describes the multicast and unicast call flows for music on hold with Skinny Client Control Protocol (SCCP) endpoints.

SCCP Multicast Call Flow

Figure 17-6 illustrates a typical SCCP multicast call flow. As shown in the diagram, when the Hold softkey is pressed at phone A, Unified CM instructs both phone A and phone B to Close Receive Channel and Stop Media Transmission. This action effectively stops the RTP two-way audio stream. Next, Unified CM tells phone B (the holdee) to Start Multicast Media Reception from multicast group address 239.192.240.1. The phone then issues an Internet Group Management Protocol (IGMP) V2 Membership Report message indicating that it is joining this group.

Figure 17-6 Detailed SCCP Multicast MoH Call Flow



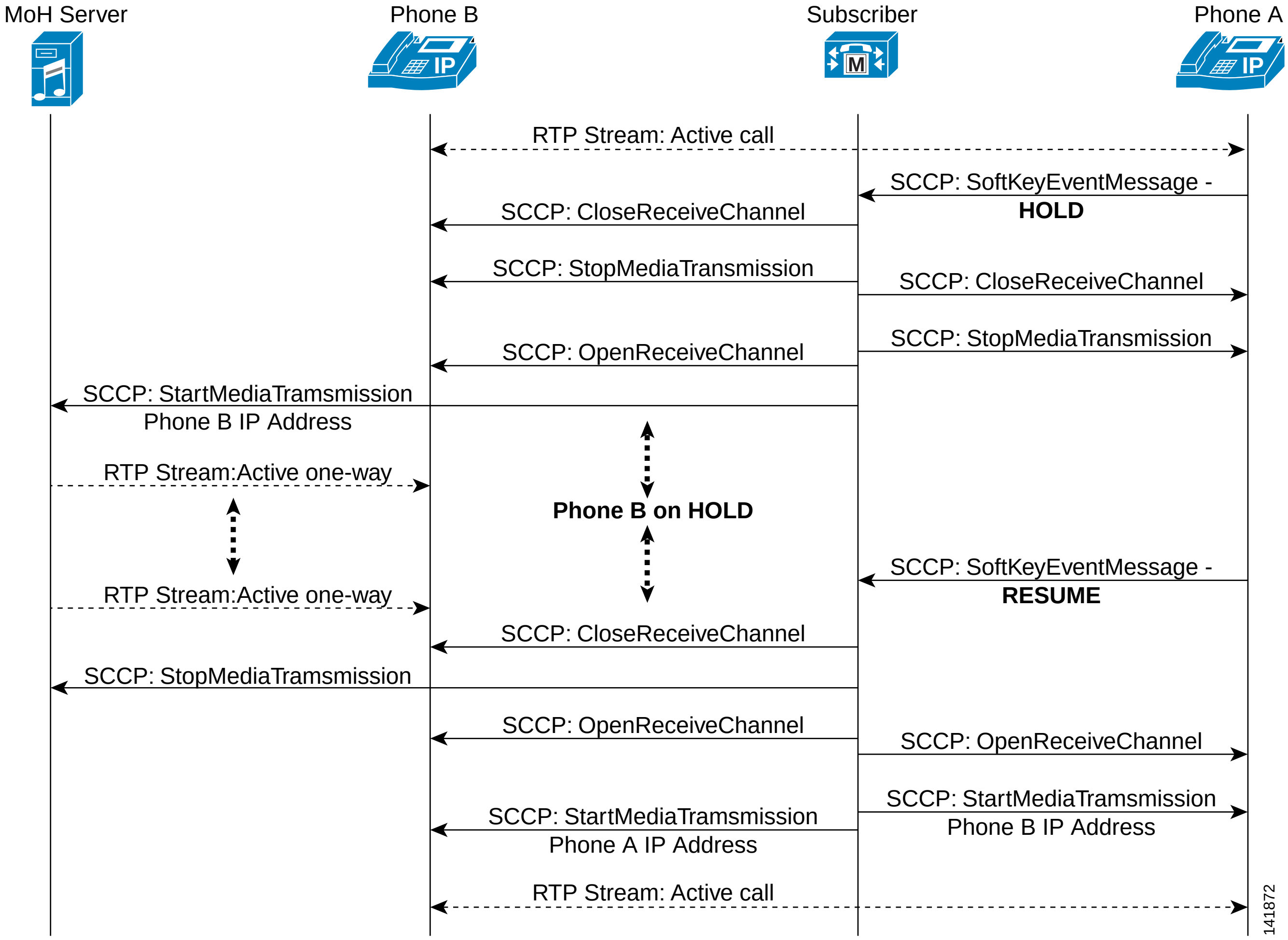
Meanwhile, the MoH server has been sourcing RTP audio to this multicast group address and, upon joining the multicast group, phone B begins receiving the MoH stream. Once phone A presses the Resume softkey, Unified CM instructs phone B to Stop Multicast Media Reception. Phone B then sends an IGMP V2 Leave Group message to 224.0.0.2 to indicate that the multicast stream is no longer needed. This effectively ends the MoH session. Next, Unified CM sends a series of Open Receive Channel messages to phones A and B, just as would be sent at the beginning of a phone call between the two phones. Soon afterwards, Unified CM instructs both phones to Start Media Transmission to each other's IP addresses. The phones are once again connected by means of an RTP two-way audio stream.

Note The call flow diagrams in Figure 17-6 and Figure 17-7 assume that an initial call exists between phones A and B, with a two-way RTP audio stream. These diagrams are representative of call flows and therefore include only the pertinent traffic required for proper MoH operation. Thus, keep-alives, acknowledgements, and other miscellaneous traffic have been eliminated to better illustrate the interaction. The initial event in each diagram is the Hold softkey action performed by phone A.

SCCP Unicast Call Flow

Figure 17-7 depicts an SCCP unicast MoH call flow. In this call flow diagram, when the Hold softkey is pressed at phone A, Unified CM instructs both phone A and phone B to Close Receive Channel and Stop Media Transmission. This action effectively stops the RTP two-way audio stream. Up to this point, unicast and multicast MoH call flows behave exactly the same way.

Figure 17-7 Detailed SCCP Unicast MoH Call Flow


MoH Server
Phone B
Subscriber
Phone A
IP
M
IP
RTP Stream: Active call
SCCP: SoftKeyEventMessage - HOLD
SCCP: CloseReceiveChannel
SCCP: StopMediaTransmission
SCCP: CloseReceiveChannel
SCCP: StopMediaTransmission
SCCP: OpenReceiveChannel
SCCP: StartMediaTramsmission
Phone B IP Address
RTP Stream:Active one-way
Phone B on HOLD
SCCP: SoftKeyEventMessage - RESUME
RTP Stream:Active one-way
SCCP: CloseReceiveChannel
SCCP: StopMediaTramsmission
SCCP: OpenReceiveChannel
SCCP: OpenReceiveChannel
SCCP: StartMediaTramsmission
Phone B IP Address
SCCP: StartMediaTramsmission
Phone A IP Address
RTP Stream: Active call
141872

Next, Unified CM tells phone B (the holdee) to Open Receive Channel. (This is quite different from the multicast case, where Unified CM tells the holdee to Start Multicast Media Reception.) Then Unified CM tells the MoH server to Start Media Transmission to the IP address of phone B. (This too is quite different behavior from the multicast MoH call flow, where the phone is prompted to join a multicast group address.) At this point, the MoH server is sending a one-way unicast RTP music stream to phone B. When phone A presses the Resume softkey, Unified CM instructs the MoH server to Stop Media Transmission and instructs phone B to Close Receive Channel, effectively ending the MoH session. As with the multicast scenario, Unified CM sends a series of Open Receive Channel messages and Start Media Transmissions messages to phones A and B with each other’s IP addresses. The phones are once again connected by means of an RTP two-way audio stream.

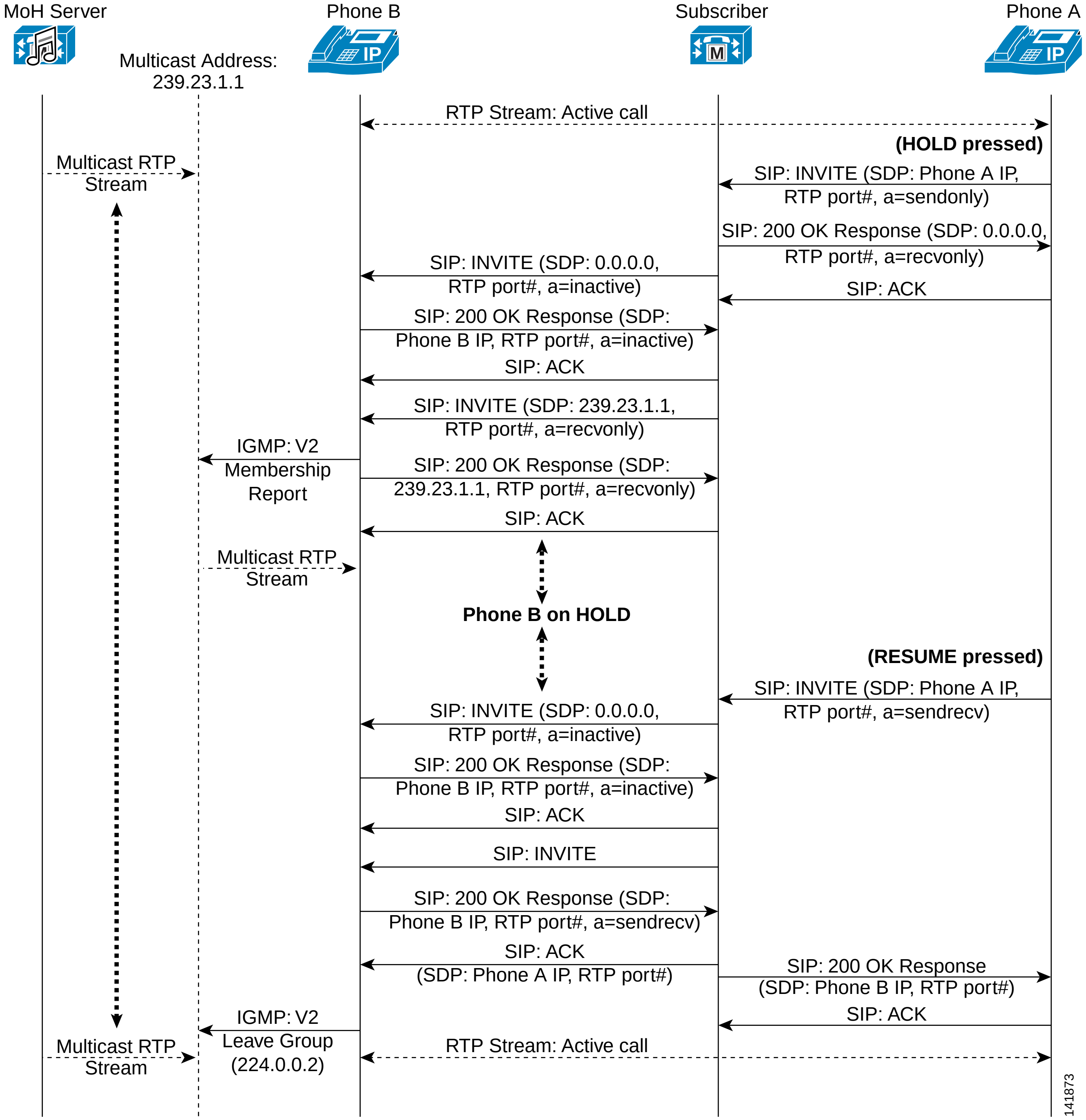
SIP Call Flows

This section describes the multicast and unicast call flows for music on hold with Session Initiation Protocol (SIP) endpoints.

SIP Multicast Call Flow

Figure 17-8 illustrates a typical SIP multicast call flow. As shown in the diagram, when the Hold softkey is pressed at phone A, phone A sends a SIP INVITE with a Session Description Protocol (SDP) connection information indication of phone A's IP address and a media attribute indication of sendonly. Unified CM then instructs phone A to disconnect the RTP stream by means of a SIP 200 OK Response with an SDP connection information indication of 0.0.0.0 and a media attribute indication of recvonly. Phone B is then told to disconnect the RTP stream by means of a SIP INVITE from Unified CM with an SDP connection information indication of 0.0.0.0 and a media attribute of inactive. After a SIP 200 OK Response is sent back from phone B to Unified CM indicating an SDP media attribute of inactive, Unified CM then sends a SIP INVITE to phone B with an SDP connection information indication of the MoH multicast group address (in this case 239.23.1.1) and a media attribute of recvonly.

Figure 17-8 Detailed SIP Multicast MoH Call Flow



Next, phone B in Figure 17-8 issues an IGMP V2 Membership Report message indicating that it is joining this multicast group. In addition, phone B sends a SIP 200 OK Response back to Unified CM indicating an SDP media attribute of recvonly in response to the previous SIP INVITE. Meanwhile, the MoH server has been sourcing RTP audio to this MoH multicast group address and, upon joining the multicast group, phone B begins receiving the one-way MoH stream.

When the user at phone A presses the Resume softkey, phone A sends a SIP INVITE with an SDP connection information indication of phone A's IP address and media attribute indications of phone A's receiving RTP port and sendrecv. Unified CM then instructs phone B to disconnect from the multicast MoH stream by means of a SIP INVITE with an SDP connection information indication of 0.0.0.0 and a media attribute indication of inactive. A SIP 200 OK Response is sent back from phone B to Unified CM, indicating an SDP media attribute of inactive.

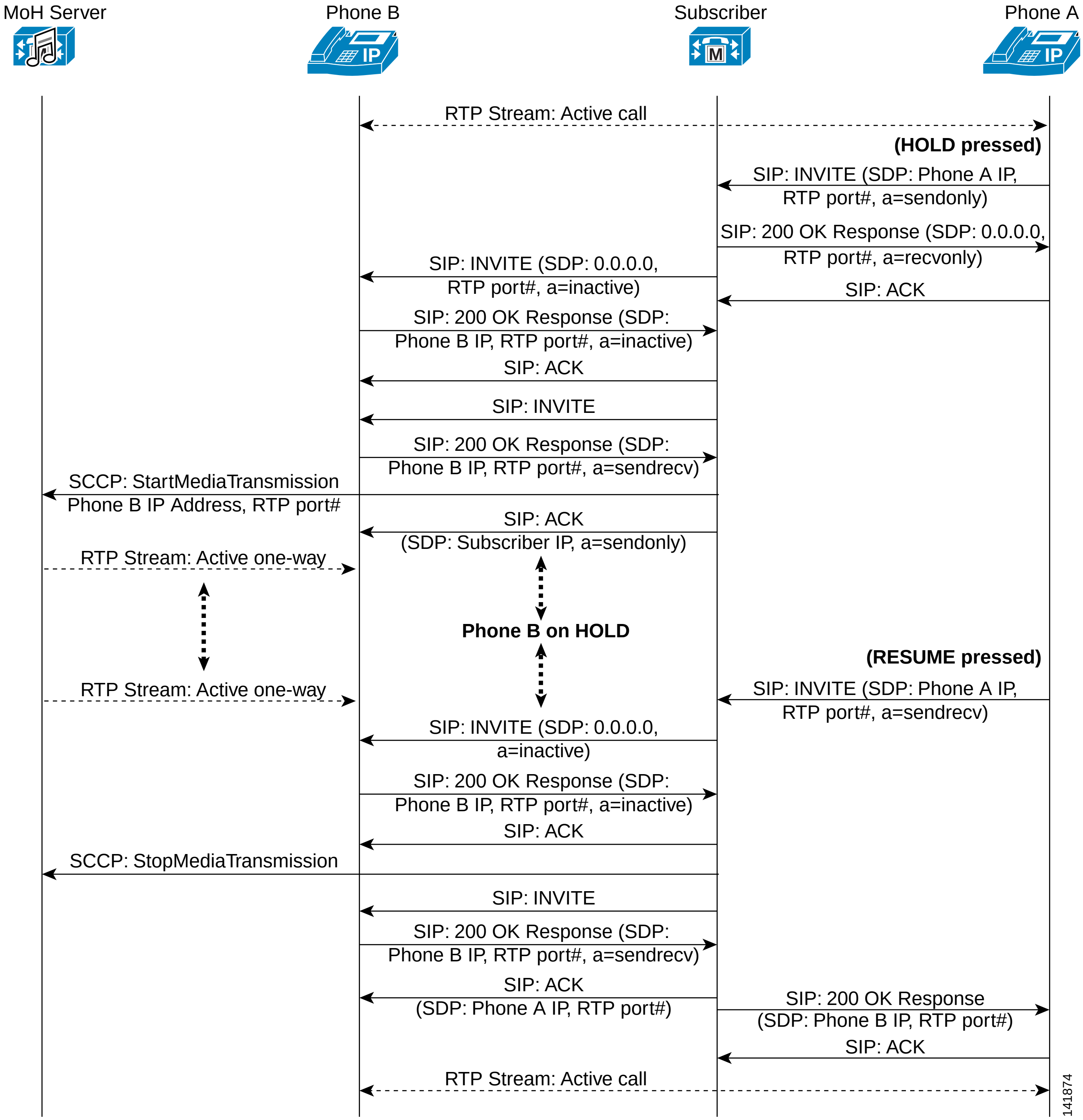
Next Unified CM sends a SIP INVITE to phone B, and phone B responds with a SIP 200 OK Response with an SDP connection information indication of phone B's IP address and media attribute indications of phone B's receiving RTP port and sendrecv. Unified CM responds by sending a SIP ACK to phone B with an SDP connection information indication of phone A's IP address and a media attribute of phone A's receiving RTP port number. Likewise, Unified CM forwards a SIP 200 OK Response to phone A's original resuming SIP INVITE, with an SDP connection information indication of phone B's IP address and a media attribute of phone B's receiving RTP port number. Phone B then sends an IGMP V2 Leave Group message to 224.0.0.2 to indicate that the multicast stream is no longer needed. Finally, the RTP two-way audio stream between phones A and B is reestablished.

Note The call flow diagrams in Figure 17-8 and Figure 17-9 assume that an initial call exists between phones A and B, with a two-way RTP audio stream. These diagrams are representative of call flows and therefore include only the pertinent traffic required for proper MoH operation. Thus, keep-alives, some acknowledgements, progression indications, and other miscellaneous traffic have been eliminated to better illustrate the interaction. The initial event in each diagram is the Hold softkey action performed by phone A.

SIP Unicast Call Flow

Figure 17-9 depicts a SIP unicast MoH call flow. As shown in the diagram, when the Hold softkey is pressed at phone A, phone A sends a SIP INVITE with an SDP connection information indication of phone A's IP address and a media attribute indication of sendonly. Unified CM then instructs phone A to disconnect the RTP stream by means of a SIP 200 OK Response with an SDP connection information indication of 0.0.0.0 and a media attribute indication of recvonly. Phone B is then told to disconnect the RTP stream by means of a SIP INVITE from Unified CM, with an SDP connection information indication of 0.0.0.0 and a media attribute of inactive. Next a SIP 200 OK Response is sent back from phone B to Unified CM, indicating an SDP media attribute of inactive. Up to this point, unicast and multicast MoH call flows are exactly the same.

Figure 17-9 Detailed SIP Unicast MoH Call Flow


MoH Server
Phone B
Subscriber
Phone A
IP
M
IP
RTP Stream: Active call
(HOLD pressed)
SIP: INVITE (SDP: Phone A IP, RTP port#, a=sendonly)
SIP: 200 OK Response (SDP: 0.0.0.0, RTP port#, a=recvonly)
SIP: INVITE (SDP: 0.0.0.0, RTP port#, a=inactive)
SIP: ACK
SIP: 200 OK Response (SDP: Phone B IP, RTP port#, a=inactive)
SIP: ACK
SIP: INVITE
SIP: 200 OK Response (SDP: Phone B IP, RTP port#, a=sendrecv)
SCCP: StartMediaTransmission Phone B IP Address, RTP port#
SIP: ACK (SDP: Subscriber IP, a=sendonly)
RTP Stream: Active one-way
Phone B on HOLD
(RESUME pressed)
SIP: INVITE (SDP: Phone A IP, RTP port#, a=sendrecv)
RTP Stream: Active one-way
SIP: INVITE (SDP: 0.0.0.0, a=inactive)
SIP: 200 OK Response (SDP: Phone B IP, RTP port#, a=inactive)
SIP: ACK
SCCP: StopMediaTransmission
SIP: INVITE
SIP: 200 OK Response (SDP: Phone B IP, RTP port#, a=sendrecv)
SIP: ACK (SDP: Phone A IP, RTP port#)
SIP: 200 OK Response (SDP: Phone B IP, RTP port#)
SIP: ACK
RTP Stream: Active call
141874

Unified CM then sends a SIP INVITE to phone B, and phone B responds back with a SIP 200 OK Response indicating SDP connection information with phone B's IP address and media attribute indications of phone B's receiving RTP port number and sendrecv. Unified CM then sends a SCCP StartMediaTransmission message to the MoH server, with phone B's address and receiving RTP port number. This is followed by a SIP ACK from Unified CM to phone B indicating SDP connection information of the Unified CM IP address and a media attribute of sendonly. Meanwhile, the MoH server begins sourcing RTP audio to phone B, and phone B begins receiving the one-way MoH stream.

When the user at phone A presses the Resume softkey, phone A sends a SIP INVITE with an SDP connection information indication of phone A's IP address and media attribute indications of phone A's receiving RTP port and sendrecv. Unified CM then instructs phone B to disconnect from the multicast MoH stream by means of a SIP INVITE with an SDP connection information indication of 0.0.0.0 and a media attribute indication of inactive. A SIP 200 OK Response is sent back from phone B to Unified CM, indicating an SDP media attribute of inactive. Then Unified CM sends an SCCP StopMediaTransmission message to the MoH server, causing the MoH server to stop forwarding the MoH stream to phone B.

Next Unified CM sends a SIP INVITE to phone B, and phone B responds with a SIP 200 OK Response with an SDP connection information indication of phone B's IP address and media attribute indications of phone B's receiving RTP port and sendrecv. Unified CM responds by sending a SIP ACK to phone B, with an SDP connection information indication of phone A's IP address and a media attribute of phone A's receiving RTP port number. Likewise, Unified CM forwards a SIP 200 OK Response to phone A's original resuming SIP INVITE with an SDP connection information indication of phone B's IP address and a media attribute of phone B's receiving RTP port. Finally, the RTP two-way audio stream between phones A and B is reestablished.

Capacity Planning for Media Resources

This section provides information on the capacities of various network modules and chassis that carry DSPs, the capacities of the chassis to carry network modules, and software dependencies of the hardware.

For all Cisco ISR G1 and G2 capacity planning, use the DSP Calculator available at http://www.cisco.com/go/dspcalculator. For other platforms (such as the Cisco 1700, 2600, and 3700 Series Routers), use the legacy DSP calculator at http://www.cisco.com/cgi-bin/Support/DSP/cisco_dsp_calc.pl.

The DSP resources for Unified Communications solutions are provided by NM-HD, NM-HDV, and PVDM modules. NM-HD and NM-HDV2 modules are supported on Cisco ISR G1 and G2 Series platforms. Refer to the respective product data sheets for capacity information for these modules.

PVDM modules are available in three models: PVDM-256K, PVDM2, and the newer PVDM3. Each of the models has several modules with different density support. For example, a PVDM-256K-4 and a PVDM2-16 are single DSP modules in their respective category. PVDM2 modules are supported on Cisco ISR G1 and ISR G2 platforms (minimum of Cisco IOS Release 15.0(1)M is required for the Cisco ISR G2 Series). The PVDM3 DSP modules are supported on the Cisco 2900 Series and 3900 Series platforms and require a minimum Cisco IOS Release of 15.0(1) M. PVDM3 modules provide DSP resources for both voice and video. The PVDM3 modules are newer than the PVDM2 and PVDM-256K modules, and the three types are not interchangeable.

Some things to consider when doing capacity planning for hardware-based media resources include the density of the module, the underlying platform (Cisco ISR G1 or G2), and the minimum Cisco IOS version required.

For capacity information on PVDM2 modules, refer to the *High-Density Packet Voice Digital Signal Processor Module for Cisco Unified Communications Solutions* data sheet, available at

> http://www.cisco.com/en/US/prod/collateral/routers/ps5854/product_data_sheet0900aecd8016e845_ps3115_Products_Data_Sheet.html

For capacity information on PVDM3 modules, refer to the *High-Density Packet Voice Video Digital Signal Processor Module for Cisco Unified Communications Solutions* data sheet, available at

http://www.cisco.com/en/US/prod/collateral/modules/ps3115/data_sheet_c78-553971.html

Considerations for Cisco 2900 and 3900 Series Platforms

The following guidelines and considerations apply to the DSP resources hosted by these platforms:

- The Cisco 2900 and 3900 Series Routers support only the PVDM3 DSPs in the on-board (motherboard) DSP slots. PVDM2 DSPs may be used in those slots by using an adaptor card. NM-HD and NM-HDV2 cards can be used in Service Module slots with an adaptor card.
- PVDM2 and PVDM3 modules cannot be used at the same time on the same motherboard.
- DSP sharing can be done only between the same DSP types. For example, if the motherboard is populated with PVDM3 DSPs and the Service Modules are populated with PVDM2 DSPs, then the DSPs in the Service Modules may be shared with each other but DSPs on the motherboard may not be shared with those in the Service Modules.
- PVDM3 DSPs support all the functions that the PVDM2 DSPs support except for Cisco Fax Relay.

Unlike the PVDM2, the PVDM3 DSPs have a single software image for all media functions.

Cisco 2800 and 3800 Series Platforms

The following guidelines and considerations apply to the DSP resources hosted by these platforms:

- Although the Cisco 2800 and 3800 Series Routers all have two AIM slots, they do not support the AIM-VOICE-30 or AIM-ATM-VOICE-30 cards because PVDM2 modules that are installed on the motherboard provide that functionality.

You can install the NM-HDV2, NM-HD-*xx*, and NM-HDV modules in the Cisco IOS platforms as indicated in the product data sheets.

All three families of modules may be installed in a single chassis. However, the conferencing and transcoding features cannot be used simultaneously on both the NM-HDV family and either of the other two families (NM-HD-*xx* or NM-HDV2). In addition, the NM-HDV (TI-549), NM-HD-*xx*, and NM-HDV2 (TI-5510) cannot be used simultaneously for conferencing and transcoding within a single chassis.

You can mix NM-HDV and NM-HDV-FARM modules in the same chassis, but not all chassis can be completely populated by these modules.

Capacity Planning for Music on Hold

It is important to be aware of the hardware capacity for MoH resources and to consider the implications of multicast and unicast MoH in relation to this capacity when doing capacity planning for MoH resources. The capacity of the MoH server depends on several factors such as deployment model (co-resident or standalone), underlying server platform, and so forth.

Co-resident and Standalone MoH

The MoH feature requires the use of a server that is part of a Unified CM cluster. You can configure the MoH server in either of the following ways:

- Co-resident deployment

  The term *co-resident* refers to two or more services or applications running on the same server. In a co-resident deployment, the MoH feature runs on any server (either publisher or subscriber) in the cluster that is also running the Unified CM software.

- Standalone deployment

  A standalone deployment, places the MoH feature on a dedicated media resource server node within the Unified CM cluster. This server acts as neither a publisher or a subscriber. That is, the Cisco IP Voice Media Streaming Application service is the only service enabled on the server. The only function of this dedicated server is to send MoH streams to devices within the network.

Server Platform Limits

Starting with Cisco Unified Communications Manager Release 9.0, a maximum of 1,000 MoH streams is supported across MCS 7835 and MCS 7845 servers or OVA equivalent platforms for both standalone and co-resident deployments. For other platforms, Unified CM can support half that amount or less, depending upon what other services are active on the server. Ensure that network call volumes do not exceed these limits because, once MoH sessions have reached these limits, additional load could result in poor MoH quality, erratic MoH operation, or even loss of MoH functionality. Note that you can configure a maximum of 51 unique audio sources per Unified CM cluster.

For more information on supported server platforms, refer to *Supported Servers for Releases of Cisco Unified Communications Manager (Including Business Edition 3000/5000/6000 and Session Manager Edition) and Cisco Intercompany Media Engine*, available at

http://www.cisco.com/en/US/prod/collateral/voicesw/ps6790/ps5748/ps378/prod_brochure0900aecd8062a4f9.html

The following two MoH Server Configuration parameters affect MoH server capacity:

- **Maximum Half Duplex Streams**

  This parameter determines the number of devices that can be placed on unicast MoH. By default this value is set to 250.

  The Maximum Half Duplex Streams parameter should be set to the value derived from the following formula:

  > (Server and deployment capacity) – ((Number of multicast MoH sources) ∗ (Number of MoH codecs enabled))

  For example:

MCS-7835 standalone MoH server (or OVA equivalent)	Multicast MoH audio sources	MoH codecs enabled (G.711 mu-law and G.729)	**Maximum half-duplex streams**
1,000	- (12	∗ 2)	= **976**

  Therefore, in this example, the Maximum Half Duplex Streams parameter would be configured with a value of no more than 976.

- **Maximum Multicast Connections**

  This parameter determines the number of devices that can be placed on multicast MoH.

  The Maximum Multicast Connections parameter should be set to a number that ensures that all devices can be placed on multicast MoH if necessary. Although the MoH server can generate only a finite number of multicast streams, a large number of held devices can join each multicast stream. This parameter should be set to a number that is greater than or equal to the number of devices that might be placed on multicast MoH at any given time. Typically multicast traffic is accounted for based on the number of streams being generated; however, Unified CM maintains a count of the actual number of devices placed on multicast MoH or joined to each multicast MoH stream. Although this method is different than the way multicast traffic is normally tracked, it is important to configure this parameter appropriately.

  Note Because you can configure only 51 unique audio sources per Unified CM cluster and because there are only four possible codecs for MoH streams, the maximum number of multicast streams per MoH server is 204.

Failure to configure these parameters properly could lead to under-utilization of MoH server resources or failure of the server to handle the network load. For details on how to configure the service parameters, refer to the *Cisco Unified Communications Manager Administration Guide*, available at

http://www.cisco.com/en/US/products/sw/voicesw/ps556/prod_maintenance_guides_list.html

Note The maximum limit of 1,000 sessions per MoH server applies to unicast, multicast, or simultaneous unicast and multicast sessions. The limit represents the recommended maximum number of sessions a platform can support, irrespective of the transport mechanism.

Resource Provisioning

When provisioning for co-resident or standalone MoH server configurations, network administrators should consider the type of transport mechanism used for the MoH audio streams. If using unicast MoH, each device on hold requires a separate MoH stream. However, if using multicast MoH and only a single audio source, then only a single MoH stream is required for each configured codec type, no matter how many devices of that type are on hold.

For example, given a cluster with 30,000 phones and a 2% hold rate (only 2% of all endpoint devices are on hold at any given time), 600 MoH streams or sessions would be required. Given a unicast-only MoH environment, one co-resident (or standalone) MoH server running on an MCS 7845 (or OVA equivalent) would be required to handle this load.

By comparison, a multicast-only MoH environment with 36 unique MoH audio streams, for example, would require one co-resident MoH server (MCS 7816, 7825, or 7878). These 36 unique multicast streams could be provisioned in any one of the following ways:

- 36 unique audio sources streamed using a single codec
- 18 unique audio sources streamed using only 2 codecs
- 12 unique audio sources streamed using only 3 codecs
- 9 unique audio source streamed using all 4 codecs

In the preceding examples, the 2% hold rate is based on 30,000 phones and does not take into account gateways or other endpoint devices in the network that are also capable of being placed on hold. You should consider these other devices when calculating a hold rate because they could potentially be placed on hold just as the phones can.

The preceding calculations also do not provide for MoH server redundancy. If an MoH server fails or if more than 2% of the users go on hold at the same time, there are no other MoH resources in this scenario to handle the overflow or additional load. Your MoH resource calculations should include enough extra capacity to provide for redundancy. Additional MoH servers can be provisioned for redundancy or high availability as explained in the section on High Availability for Media Resources, page 17-38.

High Availability for Media Resources

The Unified CM constructs of media resource groups (MRGs) and media resource group lists (MRGLs) are used to control how the resources described in this chapter are organized and accessed. This section discusses considerations for how to utilize these constructs effectively.

Media Resource Groups and Lists

Media resource groups (MRGs) and media resource lists (MRGLs) provide a method to control how resources are allocated that could include rights to resources, location of resources, or resource type for specific applications. This section assumes you have an understanding of media resource groups and lists, and it highlights the following design considerations:

- The system defines a default media resource group that is not visible in the user interface. All resources are members of this default MRG when they are created. When using MRGs to control access to resources, it is necessary to move the resources out of the default MRG by explicitly configuring them in some other MRG. If the desired effect is for resources to be available only as a last resort for all calls, then the resources may remain in the default group. Also, if no control over resources is necessary, they may remain in the default group.

- Consumers of media resources use resources first from any media resource group (MRG) or media resource group list (MRGL) that their configuration specifies. If the required resource is not available, the default MRG is searched for the resource. For simple deployments, the default MRG alone may be used.
- Use media resource groups (MRGs) and media resource group lists (MRGLs) to provide sharing of resources across multiple Unified CMs. If you do not use MRGs and MRGLs, the resources are available to a single Unified CM only.
- MRGLs will use MRGs in the order that they are listed in the configuration. If one MRG does not have the needed resource, the next MRG is searched. If all MRGs are searched and no resource is found, the search terminates.
- Within an MRG, resources are allocated based on their order in their configuration even though Unified CM Administration displays the devices in an MRG in alphabetical order. If you want media resources to be allocated in a specific order, Cisco recommends that you create a separate MRG for each individual resource and use MRGLs to specify the order of allocation.
- When there are multiple devices providing the same type of resource within an MRG, the algorithm for allocating that resource load-balances across all those devices. Cisco Unified CM uses a throttling mechanism to load balance across MTP and transcoder resources using the **MTP and Transcoder Resource Throttling Percentage** service parameter, which defines a percentage of the configured number of MTP or transcoder resources. When the number of active MTP or transcoder resources is equal to or greater than the percentage that is configured for this parameter, Cisco Unified CM stops sending calls to this resource and hunts through the MRGL (including the default MRG) one time to find a resource that uses matching codecs on both sides of the call. If Cisco Unified CM cannot find an available resource with matching codecs, it returns to the top of the MRGL to repeat the search, which then includes those resources that are in a throttled state and that match a smaller subset of capabilities for the call. Cisco Unified CM extends the call to the resource that is the best match for the call when such a resource is available. The call fails when Cisco Unified CM cannot allocate a resource for the call.
- Unified CM server-based software MTPs are pass-through enabled by default. Cisco IOS Enhanced MTP devices can be configured to support codec pass-through or non-codec pass-through modes. If a codec pass-through MTP is required and if, after the first iteration through the MRGL (including the default MRG), a codec pass-through MTP is not found, then there will be a second iteration that will ignore codec pass-through capabilities.
- An MRG may contain multiple types of resources, and the appropriate resource will be allocated from the group based on the feature needed. MTPs and transcoders are a special case because a transcoder may also be used as an MTP. For example, when both MTPs and transcoders exist in the same MRG and an MTP is required, the allocation is done based on the order in which the resources appear in the MRG. If transcoder devices appear earlier than MTPs in the MRG, transcoder resources will be allocated for the MTP requirement until the transcoder resources are exhausted and then the system will start allocating MTPs. For this reason, it is important to consider the order of resources when creating MRGs and MRGLs.
- MRGs can also be used to group resources of similar types. As explained in the example above, because a transcoder is a more expensive resource, Cisco recommends grouping transcoders and MTPs into separate MRGs and invoking the right resource by adding MRGs to the MRGL in appropriate order. Another example involves conference bridges. Conference bridge resources vary in the number of participants they support, and different MRGs could be used to group the conference resources by conference bridge size.

- You can also use MRGs and MRGLs to separate resources based on geographical location, thereby conserving WAN bandwidth whenever possible.
- Ensure that the media resources themselves have configurations that prevent further invocation of other media resources. For example, if an MTP is inserted into a call and the codec configured on that MTP does not match the one needed by Unified CM for the call, then a transcoder may also be invoked. A frequent mistake is to configure an MTP for G.729 or G.729b when Unified CM needs G.729a.

Redundancy and Failover Considerations for Cisco IOS-Based Media Resources

A high availability design with media resources must include redundant media resources. When these resources are Cisco IOS-based, they can be distributed on more than one Cisco IOS platform to guard against failure of a single platform and they can be registered to different primary Unified CM servers.

Cisco IOS supports two modes of failover capability: graceful and immediate. The default failover method is graceful, in which the resources register to a backup Unified CM server only after all media activity has ceased. The immediate method, on the other hand, makes the resources register to the backup Unified CM server as soon as failure of the primary is detected. In situations where there is only one set of media resources with no redundancy, Cisco recommends use of the immediate failover method.

High Availability for Music on Hold

Cisco recommends that you configure and deploy multiple MoH servers for completely redundant MoH operation. If the first MoH server fails or becomes unavailable because it no longer has the resources required to service requests, the second server can provide continued MoH functionality. For proper redundant configuration, assign resources from at least two MoH servers to each MRG in the cluster.

In environments where both multicast and unicast MoH are required, be sure to provide redundancy for both transport types to ensure MoH redundancy for all endpoints in the network.

Design Considerations for Media Resources

This section discusses specific considerations for deploying media resources for use with the various Unified CM deployment models. It also highlights the configuration considerations and best practices to help you design a robust solution for media resource allocation in your Unified CM implementation.

Deployment Models

This section examines where and when the MTP and transcoding resources are used within the following three enterprise IP Telephony deployment models:

- Single-Site Deployments, page 17-41
- Multisite Deployments with Centralized Call Processing, page 17-41
- Multisite Deployments with Distributed Call Processing, page 17-42

Single-Site Deployments

In a single-site deployment, there is no need for transcoding because there are no low-speed links to justify the use of a low bit-rate (LBR) codec. Some MTP resources might be required in the presence of a significant number of devices that are not compliant with H.323v2, such as older versions of Microsoft NetMeeting or certain video devices. MTP resources may be required for DTMF conversion if SIP endpoints are present (see Named Telephony Events (RFC 2833), page 17-13.)

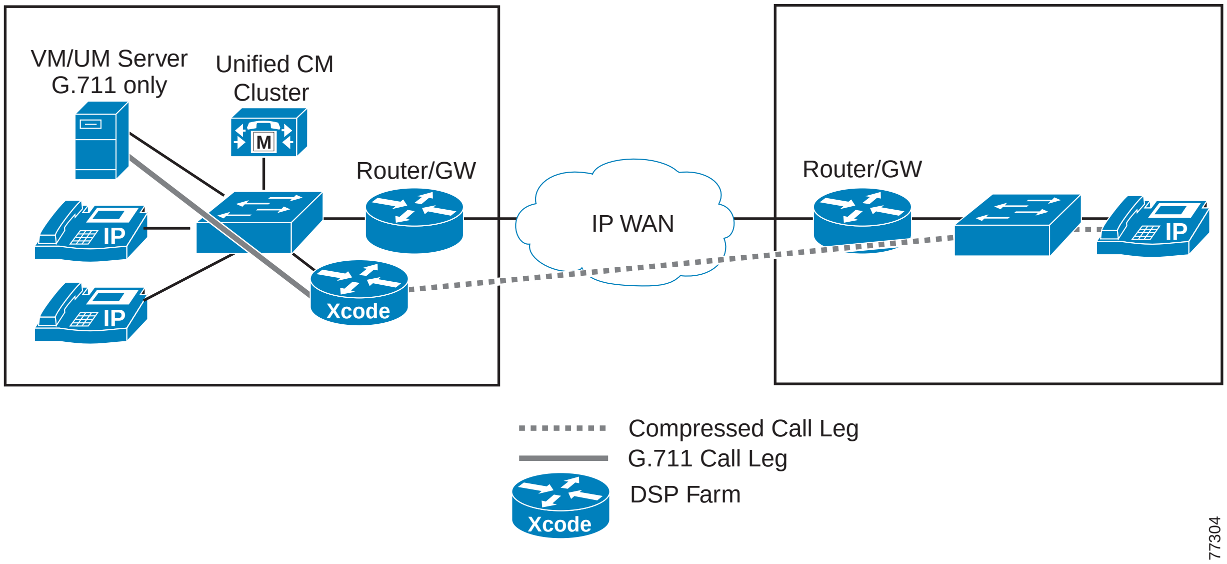
Multisite Deployments with Centralized Call Processing

In a centralized call processing deployment, the Unified CM cluster and the applications (such as voice mail and IVR) are located at the central site, while several remote sites are connected through an IP WAN. The remote sites rely on the centralized Unified CMs to handle their call processing.

Because WAN bandwidth is typically limited, calls are configured to use a low bit-rate codec such as G.729 when traversing the WAN. (See Figure 17-10.)

Voice compression between IP phones is easily configured through the use of *regions* and *locations* in Unified CM. A region defines the type of compression (for example, G.711 or G.729) used by the devices in that region, and a location specifies the total amount of bandwidth available for calls to and from devices at that location.

Figure 17-10 Transcoding for the WAN with Centralized Call Processing



Unified CM uses media resource groups (MRGs) to enable sharing of MTP and transcoding resources among the Unified CM servers within a cluster. In addition, when using an LBR codec (for example, G.729a) for calls that traverse different regions, the transcoding resources are used only if one (or both) of the endpoints is unable to use the LBR codec.

In Figure 17-10, Unified CM knows that a transcoder is required and allocates one based on the MRGL and/or MRG of the device that is using the higher-bandwidth codec. In this case it is the VM/UM server that determines which transcoder device is used. This behavior of Unified CM is based on the assumption that the transcoder resources are actually located close to the higher-bandwidth device. If this system was designed so that the transcoder for the VM/UM server was located at the remote site, then G.711 would be sent across the WAN, which would defeat the intended design. As a result, if there are multiple sites with G.711-only devices, then each of these sites would need transcoder resources when an LBR is run on the WAN.

The placement of other resources is also important. For example, if a conference occurs with three phones at a remote site and the conference resource is located in the central (call processing) site, then three media streams are carried over the WAN. If the conference resource were local, then the calls would not traverse the WAN. It is necessary to consider this factor when designing the bandwidth and call admission control for your WAN.

Multisite Deployments with Distributed Call Processing

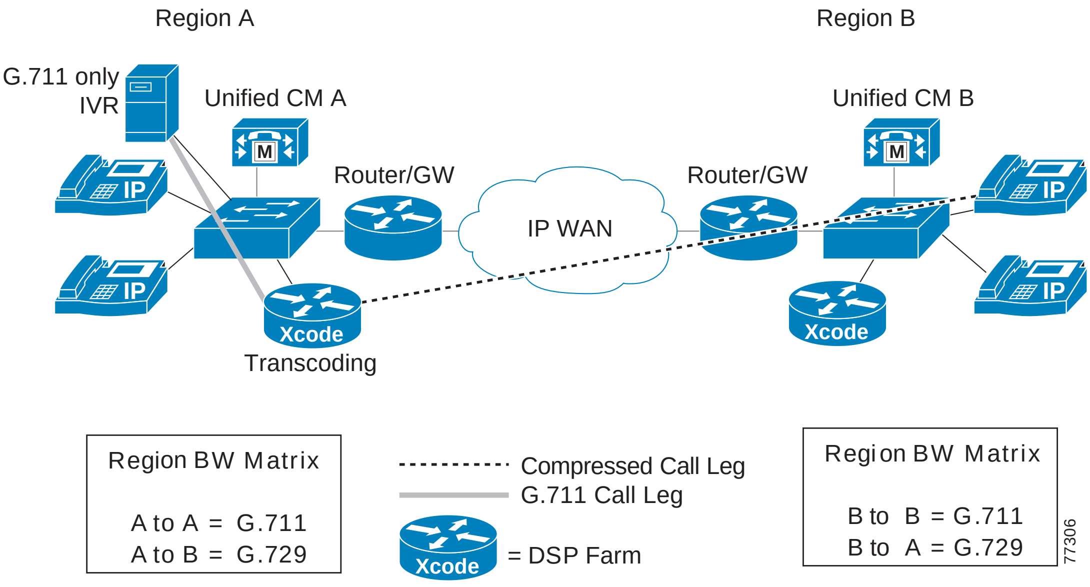
In distributed call processing deployments, several sites are connected through an IP WAN. Each site contains a Unified CM cluster that can, in turn, follow the single-site model or the centralized call processing model. A gatekeeper may be used for call admission control between sites.

Because WAN bandwidth is typically limited, calls between sites may be configured to use an LBR codec (such as G.729a) when traversing the WAN. H.323v2 intercluster trunks are used to connect Unified CM clusters. Unified CM also supports compressed voice call connections through the MTP service if a hardware MTP is used. (See Figure 17-11.)

A distributed call processing deployment might need transcoding and MTP services in the following situations:

- With current versions of Cisco applications, it is possible and recommended to avoid the use of transcoding resources. There might be specific instances where G.711 on a specific device cannot be avoided.
- Some endpoints (for example, video endpoints) do not support the H.323v2 features.

Figure 17-11 Intercluster Call Flow with Transcoding



Unified CM uses media resource groups (MRGs) to enable sharing of MTP and transcoding resources among the Unified CM servers within a cluster. In addition, for calls across intercluster trunks, MTP and transcoding resources are used only when needed, thus eliminating the need to configure the MTP service for applications that do not support LBR codecs.

The following characteristics apply to distributed call processing deployments:

- Only the intercluster calls that require transcoding will use the MTP service. For example, if both endpoints of a call are capable of using a G.729 codec, no transcoding resources will be used.
- Sharing MTP resources among servers within a cluster provides more efficient resource utilization.

Media Functions and Voice Quality

Any process that manipulates media can degrade the quality of the media. For example, encoding a voice stream for transmission across any network (IP or TDM) and decoding it at the other end will result in a loss of information, and the resulting voice stream will not be an exact reproduction of the original. If there are media traversal paths through the network that involve multiple encoding and decoding steps of the same voice stream, then each successive encoding/decoding operation will further degrade the voice quality. In general, such paths should be avoided. This is especially true for low-bandwidth codecs (LBC) such as G.729.

If such paths cannot be avoided, voice quality can generally be improved by using a higher bandwidth, low-compression codec, such as the G.711 or G.722 codecs, which are recommended wherever such paths are anticipated. Use of lower bandwidth, higher compression codecs in such scenarios is not recommended.

Music on Hold Design Considerations

This section highlights some MoH configuration considerations and best practice to help you design a robust MoH solution.

Codec Selection

If you need multiple codecs for MoH deployment, configure them in the IP Voice Media Streaming Application service parameter **Supported MoH Codecs** under the Clusterwide Unified CM Service Parameters Configuration. From the Supported MoH Codecs list under the Clusterwide Parameters, select all the desired codec types that should be allowed for MoH streams. By default, only G.711 mu-law is selected. To select another codec type, click on it in the scrollable list. For multiple selections, hold down the CTRL key and use the mouse to select multiple codecs from the scrollable list. The actual codec used for a MoH event is determined by the Region settings of the MoH server and the device being put on hold (IP phone, gateway, and so forth). Therefore, assign the proper Region setting to your MoH servers and configure the desired Region Relationships to control the codec selection of MoH interactions.

Note If you are using the G.729 codec for MoH audio streams, be aware that this codec is optimized for speech and it provides only marginal audio fidelity for music.

Multicast Addressing

Proper IP addressing is important for configuring multicast MoH. Addresses for IP multicast range from 224.0.1.0 to 239.255.255.255. The Internet Assigned Numbers Authority (IANA), however, assigns addresses in the range 224.0.1.0 to 238.255.255.255 for public multicast applications. Cisco strongly discourages using public multicast addresses for music on hold. Instead, Cisco recommends that you configure multicast MoH audio sources to use IP addresses in the range 239.1.1.1 to 239.255.255.255, which is reserved for administratively controlled applications on private networks.

Furthermore, you should configure multicast audio sources to increment on the IP address and not the port number, for the following reasons:

- IP phones placed on hold join multicast IP addresses, not port numbers.

  Cisco IP phones have no concept of multicast port numbers. Therefore, if all the configured codecs for a particular audio stream transmit to the same multicast IP address (even on different port numbers), all streams will be sent to the IP phone even though only one stream is needed. This has the potential of saturating the network with unnecessary traffic because the IP phone is capable of receiving only a single MoH stream.

- IP network routers route multicast based on IP addresses, not port numbers.

  Routers have no concept of multicast port numbers. Thus, when it encounters multiple streams sent to the same multicast group address (even on different port numbers), the router forwards all streams of the multicast group. Because only one stream is needed, network bandwidth is over-utilized and network congestion can eventually result.

MoH Audio Sources

Configured audio sources are shared among *all* MoH servers in the Unified CM cluster, requiring each audio source file to be uploaded to every MoH server within the cluster. You can configure up to 51 unique audio sources per cluster (50 audio file sources and one fixed/live source via a sound card). For methods of providing additional sources, refer to the sections on Using Multiple Fixed (Live) Audio Sources, page 17-44, and Multicast MoH from Branch Routers, page 17-49.

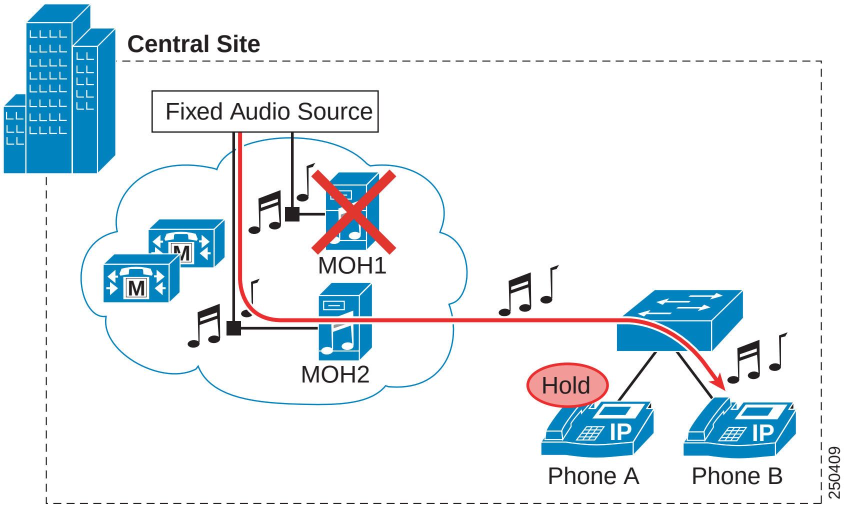
For those audio sources that will be used for multicast streaming, ensure that **Allow Multicasting** and **Play continuously (repeat)** are enabled. If continuous play of an audio source is not specified, only the first party placed on hold, not additional parties, will receive the MoH audio source.

Using Multiple Fixed (Live) Audio Sources

It is important to remember that only a single fixed audio source can be configured within Unified CM. However, each MoH server in the Unified CM cluster is capable of streaming a single fixed audio source by means of a Cisco MoH USB audio sound card (MOH-USB-AUDIO). When multiple fixed audio sources are needed, additional MoH servers can be added to provide these multiple sources. The audio supplied to each MoH server sound card can be the same or different, and the administrator can determine which MoH server is selected based on MRG and MRGL selections. When multiple audio sources are done in this manner, the holder's **User/Network Hold MoH Audio Source** should be configured for the fixed audio source (the single fixed audio source that is configured in Unified CM), and the MoH server to stream that fixed audio source to the device is then determined by the MRGL of the holdee.

In the case where the audio source is the same, this method also allows for redundancy of the fixed audio source. For example, in Figure 17-12 there are two MoH servers, each with an MOH-USB-AUDIO sound card connected to an audio source streaming audio derived from a live radio station feed. Phone B's MRGL contains first an MRG that contains the MOH1 server and second an MRG that contains the MOH2 server. Assuming the User/Network Hold Audio Source at Phone A has been set to the fixed audio source, after a call is established between Phone A and Phone B, and Phone B is placed on hold by Phone A, Phone B will receive the live feed audio source from the MOH1 server. In the case where the MOH1 server is down (or has no available capacity) when Phone A puts Phone B on hold, Phone B will receive the live feed audio source from the MOH2 server.

Figure 17-12 Fixed Audio Source Redundancy Example



Note Using live radio broadcasts as multicast audio sources can have legal ramifications. Consult your legal department for potential issues.

Unicast and Multicast in the Same Unified CM Cluster

In some cases, administrators might want to configure a single Unified CM cluster to handle both unicast and multicast MoH streams. This configuration might be necessary because the telephony network contains devices or endpoint that do not support multicast or because some portions of the network are not enabled for multicast.

Use one of the following methods to enable a cluster to support both unicast and multicast MoH audio streams:

- Deploy separate MoH servers, with one server configured as a unicast MoH server and the second server configured as a multicast MoH server.
- Deploy a single MoH server with two media resource groups (MRGs), each containing the same MoH server, with one MRG configured to use multicast for audio streams and the second MRG configured to use unicast.

In either case, you must configure at least two MRGs and at least two media resource group lists (MRGLs). Configure one unicast MRG and one unicast MRGL for those endpoints requiring unicast MoH. Likewise, configure one multicast MRG and one multicast MRGL for those endpoints requiring multicast MoH.

When deploying separate MoH servers, configure one server without multicast enabled (unicast-only) and configure a second MoH server with multicast enabled. Assign the unicast-only MoH media resource and the multicast-enabled MoH media resource to the unicast and multicast MRGs, respectively. Ensure that the **Use Multicast for MoH Audio** box is checked for the multicast MRG but not for the unicast MRG. Also assign these unicast and multicast MRGs to their respective MRGLs. In this case, an MoH stream is unicast or multicast based on whether the MRG is configured to use multicast and then on the server from which it is served.

When deploying a single MoH server for both unicast and multicast MoH, configure the server for multicast. Assign this same MoH media resource to both the unicast MRG and the multicast MRG, and check the **Use Multicast for MoH Audio** box for the multicast MRG. In this case, an MoH stream is unicast or multicast based solely on whether the MRG is configured to use multicast.

Note When configuring the unicast MRG, do not be confused by the fact that the MoH media resource you are adding to this MRG has [Multicast] appended to the end of the resource name even though you are adding it to the unicast MRG. This label is simply an indication that the resource is capable of being multicast, but the **Use Multicast for MoH Audio** box determines whether the resource will use unicast or multicast.

In addition, you must configure individual devices or device pools to use the appropriate MRGL. You can place all unicast devices in a device pool or pools and configure those device pools to use the unicast MRGL. Likewise, you can place all multicast devices in a device pool or pools and configure those device pools to use the multicast MRGL. Optionally, you can configure individual devices to use the appropriate unicast or multicast MRGL. Lastly, configure a User Hold Audio Source and Network Hold Audio Source for each individual device or (in the case of phone devices) individual lines or directory numbers to assign the appropriate audio source to stream.

When choosing a method for deploying both multicast and unicast MoH in the same cluster, an important factor to consider is the number of servers required. When using a single MoH server for both unicast and multicast, fewer MoH servers are required throughout the cluster. Deploying separate multicast and unicast MoH servers will obviously require more servers within the cluster.

Quality of Service (QoS)

Convergence of data and voice on a single network requires adequate QoS to ensure that time-sensitive and critical real-time applications such as voice are not delayed or dropped. To ensure proper QoS for voice traffic, the streams must be marked, classified, and queued as they enter and traverse the network to give the voice streams preferential treatment over less critical traffic. MoH servers automatically mark audio stream traffic the same as voice bearer traffic, with a Differentiated Services Code Point (DSCP) value of 46 or a Per Hop Behavior (PHB) value of EF (ToS of 0xB8). Therefore, as long as QoS is properly configured on the network, MoH streams will receive the same classification and priority queueing treatment as voice RTP media traffic.

Call signaling traffic between MoH servers and Unified CM servers is automatically marked with a DSCP value of 24 or a PHB value of CS3 (ToS of 0x60) by default. Therefore, as long as QoS is properly configured on the network, this call signalling traffic will be properly classified and queued within the network along with all other call signalling traffic.

Call Admission Control and MoH

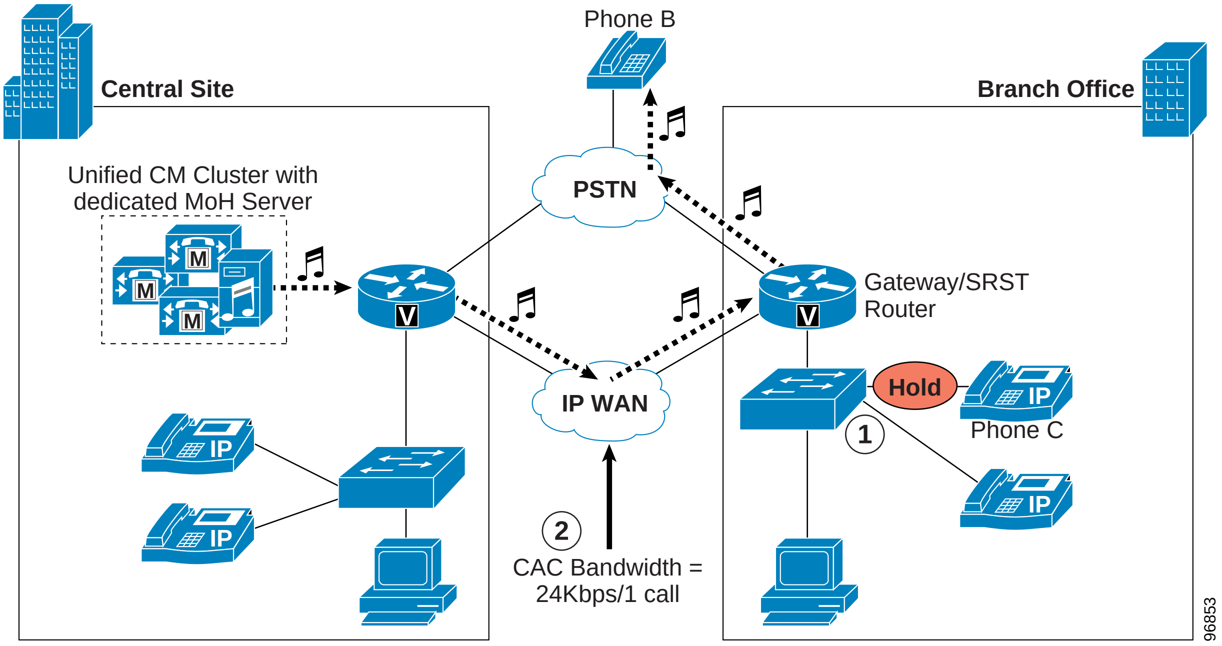
Call admission control (CAC) is required when IP telephony traffic is traveling across WAN links. Due to the limited bandwidth available on these links, it is highly probable that voice media traffic might get delayed or dropped without appropriate call admission control. For additional information, see Call Admission Control, page 11-1.

Call admission control for Unified CM (based on either static locations or RSVP-enabled locations) is capable of tracking unicast MoH streams traversing the WAN but not multicast MoH streams. Thus, even if WAN bandwidth has been fully subscribed, a multicast MoH stream will not be denied access to the WAN by call admission control. Instead, the stream will be sent across the WAN, likely resulting in poor audio stream quality and poor quality on all other calls traversing the WAN. To ensure that multicast MoH streams do not cause this over-subscription situation, you should over-provision the QoS configuration on all downstream WAN interfaces by configuring the low-latency queuing (LLQ) voice priority queue with additional bandwidth. Because MoH streams are uni-directional, only the voice priority queues of the downstream interfaces (from the central site to remote sites) must be over-provisioned. Add enough bandwidth for every unique multicast MoH stream that might traverse the WAN link. For example, if there are four unique multicast audio streams that could potentially traverse the WAN, then add 96 kbps to the voice priority queue (4 * 24 kbps per G.729 audio stream = 96 kbps).

Figure 17-13 shows an example of call admission control and MoH in a centralized multisite deployment. For this example, assume that IP phone C is in a call with a PSTN phone (phone B). At this point, no bandwidth has been consumed on the WAN. When phone C pushes the Hold softkey (step 1), phone B receives an MoH stream from the central-site MoH server by way of the WAN, thereby consuming bandwidth on the link. Whether or not this bandwidth is taken into consideration by call admission control depends on the type of MoH stream. If multicast MoH is streamed, then call admission control will not consider the 24 kbps being consumed (therefore, QoS on the downstream WAN interfaces should be provisioned accordingly). However, if unicast MoH is streamed, call admission control will subtract 24 kbps from the available WAN bandwidth (step 2).

Note The preceding example might seem to imply that unicast MoH should be streamed across the WAN. However, this is merely an example used to illustrate locations-based call admission control with MoH and is not intended as a recommendation or endorsement of this configuration. As stated earlier, multicast MoH is the recommended transport mechanism for sending MoH audio streams across the WAN.

Figure 17-13 Locations-Based Call Admission Control and MoH



Deployment Models for Music on Hold

The various Unified Communications call processing deployment models introduce additional considerations for MoH configuration design. Which deployment model you choose can also affect your decisions about MoH transport mechanisms (unicast or multicast), resource provisioning, and codecs. This section discusses these issues in relation to the various deployment models.

For more detailed information about the deployment models, see the chapter on Unified Communications Deployment Models, page 5-1.

Single-Site Campus (Relevant to All Deployments)

Single-site campus deployments are typically based on a LAN infrastructure and provide sufficient bandwidth for large amounts of traffic. Because bandwidth is typically not limited in a LAN infrastructure, Cisco recommends the use of the G.711 (A-law or mu-law) codec for all MoH audio streams in a single-site deployment. G.711 provides the optimal voice and music streaming quality in an IP Telephony environment.

MoH server redundancy should also be considered. In the event that an MoH server becomes overloaded or is unavailable, configuring multiple MoH servers and assigning them in preferred order to MRGs ensures that another server can take over and provide the MoH streams.

With the increasing diversity of network technologies, in a large single-site campus it is likely that some endpoint devices or areas of the network will be unable to support multicast. For this reason, you might have to deploy both unicast and multicast MoH resources. For more information, see the section on Unicast and Multicast in the Same Unified CM Cluster, page 17-45.

To ensure that off-net calls and application-handled calls receive expected MoH streams when placed on hold, configure all gateways and other devices with the appropriate MRGLs and audio sources, or assign them to appropriate device pools.

Centralized Multisite Deployments

Multisite IP telephony deployments with centralized call processing typically contain WAN connections to multiple non-central sites. These WAN links usually cause bandwidth and throughput bottlenecks. To minimize bandwidth consumption on these links, Cisco recommends the use of the G.729 codec for all MoH audio streams traversing the WAN. Because the G.729 codec is optimized for voice and not music applications, you should use G.729 only across the WAN, where the bandwidth savings far outweighs the lower quality afforded by G.729 for MoH transport. Likewise, because multicast traffic provides significant bandwidth savings, you should always use multicast MoH when streaming audio to endpoints across the WAN.

If the sound quality of an MoH stream becomes an issue when using the G.729 codec across the WAN, you can use the G.711 codec for MoH audio streams across the WAN while still using G.729 for voice calls. In order to send MoH streams across the WAN with the G.711 codec but voice calls across the WAN with the G.729 codec, place all MoH servers in a Unified CM region by themselves, and configure that region to use G.711 between itself and all other regions. Thus, when a call is placed between two phones on either side of a WAN, the G.729 codec is used between their respective regions. However, when the call is placed on hold by either party, the MoH audio stream is encoded using G.711 because G.711 is the configured codec to use between the MoH server's region and the region of the phone placed on hold.

Multicast MoH from Branch Routers

Branch routers deployed with the Cisco Unified Survivable Remote Site Telephony (SRST) feature can provide multicast MoH in a remote or branch site, with the MoH streaming from the branch SRST router's flash or from a live feed connected to an analog port. Multicast MoH from a branch router via these two methods enhances the Cisco Unified Communications MoH feature in both of the following scenarios:

- Non-Fallback Mode

  When the WAN is up and the phones are controlled by Unified CM, this configuration can eliminate the need to forward MoH across the WAN to remote branch sites by providing locally sourced MoH.

- Fallback Mode

  When SRST is active and the branch devices have lost connectivity to the central-site Unified CM, the branch router can continue to provide multicast MoH.

When using the live feed option in either scenario, the SRST router provides redundancy by monitoring the live feed input, and it will revert to streaming MoH from a file in flash if the live feed is disconnected. You can use only a single multicast address and port number per SRST router to provide multicast MoH; therefore, the SRST router does not support streaming from both the live feed and the flash file at the same time. In addition, the SRST router can stream only a single audio file from flash.

Note An SRST license is required regardless of whether the SRST functionality will actually be used. The license is required because the configuration for streaming MoH from branch router flash is done under the SRST configuration mode and, even if SRST functionality will not be used, at least one **max-ephones** and one **max-dn** must be configured.

Non-Fallback Mode

During non-fallback mode (when the WAN is up and SRST is not active), the branch SRST router can provide multicast MoH to all local Cisco Unified Communications devices. To accomplish this, you must configure a Unified CM MoH server with an audio source that has the same multicast IP address and port number as configured on the branch router. In this scenario, because the multicast MoH audio stream is always coming from the SRST router, it is not necessary for the central-site MoH server audio source to traverse the WAN.

To prevent the central-site audio stream(s) from traversing the WAN, use one of the following methods:

- Configure a maximum hop count

  Configure the central-site MoH audio source with a maximum hop count (or TTL) low enough to ensure that it will not stream further than the central-site LAN.

- Configure an access control list (ACL) on the WAN interface

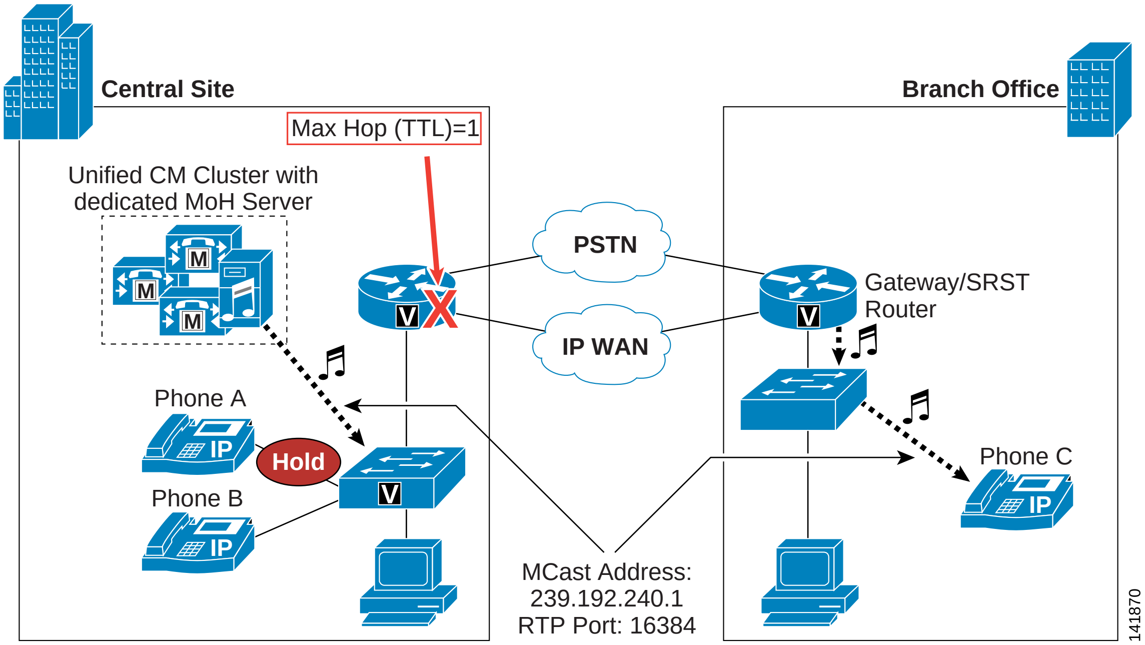
  Configure an ACL on the central-site WAN interface to disallow packets destined to the multicast group address(es) from being sent out the interface.

- Disable multicast routing on the WAN interface

  Do not configure multicast routing on the WAN interface, thus ensuring that multicast streams are not forwarded into the WAN.

Figure 17-14 illustrates streaming multicast MoH from a branch router when it is not in fallback mode. After phone A places phone C on hold, phone C receives multicast MoH from the local SRST router. In this example, the MoH server is streaming a multicast audio source to 239.192.240.1 (on RTP port 16384); however, this stream has been limited to a maximum hop of one (1) to ensure that it will not travel off the local MoH server's subnet and across the WAN. At the same time, the branch office SRST router/gateway is multicasting an audio stream from either flash or a live feed. This stream is also using 239.192.240.1 as its multicast address and 16384 as the RTP port number. When phone A presses the Hold softkey, phone C receives the MoH audio stream sourced by the SRST router.

Figure 17-14 Multicast MoH from Branch Router



When using this method for delivering multicast MoH, configure all devices within the Unified CM cluster to use the same user hold and network hold audio source and configure all branch routers with the same multicast group address and port number. Because the user or network hold audio source of the holder is used to determine the audio source, if you configure more than one user or network hold audio source within the cluster, there is no way to guarantee that a remote holdee will always receive the local MoH stream. For example, suppose a central-site phone is configured with an audio source that uses group address 239.192.254.1 as its user and network hold audio source. If this phone places a remote device on hold, the remote device will attempt to join 239.192.254.1 even if the local router flash MoH stream is sending to multicast group address 239.192.240.1. If instead all devices in the network are configured to use the user/network hold audio source with multicast group address 239.192.240.1 and all branch routers are configured to multicast from flash on 239.192.240.1, then every remote device will receive the MoH from its local router.

In networks with multiple branch routers configured to stream multicast MoH, this allows for more than 51 unique MoH audio sources within the Unified CM cluster. Each branch site router can multicast a unique audio stream, although all routers must multicast this audio on the same multicast group address. In addition, the central-site MoH server can multicast a MoH stream on this same multicast group address. Thus, if there are 100 branch sites each multicasting audio, then the cluster actually contains 101 unique MoH audio sources (100 branch streams and one central-site stream). If you want more than 51 unique audio streams in the central site, see the methods described in Using Multiple Fixed (Live) Audio Sources, page 17-44.

Fallback Mode

During fallback mode (when the WAN is down and SRST is active), the branch SRST router can stream multicast MoH to all analog and digital ports within the chassis, thereby providing MoH to analog phones and PSTN callers.

The branch router's configuration for fallback mode multicast MoH is the same as the normal operation configuration. However, which multicast address you configure on the router depends on the intended operation. If you want the branch router to provide multicast MoH to devices only in fallback mode (for example, if MoH received by remote devices is to be sourced from the central-site MoH server during non-fallback mode), then the multicast address and port number configured on the SRST router should not overlap with any of the central-site MoH server audio sources. Otherwise, remote devices might continue to receive MoH from the local router flash, depending on the configured user/network hold audio sources.

Note that, once the branch SRST/gateway router is configured to provide multicast MoH, the router will continue to multicast the MoH stream even when not in fallback mode.

It is also possible to configure the fallback mode to use Cisco Unified Communications Manager Express (Unified CME) in SRST mode. Fallback mode behavior is still the same, but the configuration commands are slightly different. SRST commands are entered under the Cisco IOS **call-manager-fallback** construct, while the commands for Unified CME in SRST mode are entered under **telephony-service**.

There are four methods of providing multicast MoH via SRST:

- SRST multicast MoH from branch router flash
- SRST multicast MoH from a live feed
- Unified CME in SRST mode with multicast MoH from branch router flash
- Unified CME in SRST mode with multicast MoH from a live feed

For more details on configuration of Cisco Unified SRST and Unified CME, refer to the following documentation:

- *Cisco Unified SRST System Administrator Guide*, available at

  http://www.cisco.com/en/US/products/sw/voicesw/ps2169/products_installation_and_configuration_guides_list.html

- *Cisco Unified Communications Manager Express System Administrator Guide*, available at

  http://www.cisco.com/en/US/products/sw/voicesw/ps4625/products_installation_and_configuration_guides_list.html

Distributed Multisite Deployments

Multisite IP telephony deployments with distributed call processing typically contain WAN or MAN connections between the sites. These lower-speed links usually cause bandwidth and throughput bottlenecks. To minimize bandwidth consumption on these links, Cisco recommends use of the G.729 codec for all MoH audio streams traversing them. Because the G.729 codec is optimized for voice and not music applications, you should use G.729 only across the WAN/MAN links, where the bandwidth savings far outweighs the lower quality afforded by G.729 for MoH transport.

Unlike with centralized multisite deployments, in situations where G.711 might be required for MoH audio streams traveling across a WAN, MoH audio streams cannot be forced to G.711 in a distributed multisite deployment. Even when MoH servers are placed in a separate Unified CM region and the G.711 codec is configured between this region and the intercluster or SIP trunk's region, the codec of the original voice call is maintained when a call between the two clusters is placed on hold by either phone. Because these intercluster calls are typically encoded using G.729 for bandwidth savings, a MoH stream from either cluster will also be encoded using G.729.

Another option is to provision multicast MoH for intercluster calls across an intercluster trunk (ICT) or SIP trunk. This allows endpoints in one Unified CM cluster to hear multicast MoH streamed from another Unified CM cluster, while making more efficient use of intercluster bandwidth. A properly designed IP Multicast environment is required to take advantage of this feature. For more information on IP Multicast, refer to the documentation available at

http://www.cisco.com/en/US/products/ps6552/products_ios_technology_home.html

Proper multicast address management is another important design consideration in the distributed intercluster environment. All MoH audio source multicast addresses must be unique across all Unified CM clusters in the deployment to prevent possible overlap of streaming resources throughout the distributed network.

Clustering Over the WAN

As its name suggests, clustering-over-the-WAN deployments also contain the same type of lower-speed WAN links as other multisite deployments and therefore are subject to the same requirements for G.729 codec, multicast transport mechanism, and solid QoS for MoH traffic traversing these links.

In addition, you should deploy MoH server resources at each side of the WAN in this type of configuration. In the event of a WAN failure, devices on each side of the WAN will be able to continue to receive MoH audio streams from their locally deployed MoH server. Furthermore, proper MoH redundancy configuration is extremely important. The devices on each side of the WAN should point to an MRGL whose MRG has a priority list of MoH resources with at least one local resource as the highest priority. Additional MoH resources should be configured for this MRG in the event that the primary server becomes unavailable or is unable to process requests. At least one other MoH resource in the list should point to an MoH resource on the remote side of the WAN in the event that resources at the local side of the WAN are unavailable.

CHAPTER 18

Unified Communications Endpoints

Revised: April 30, 2013; OL-27282-05

A variety of endpoints can be used in a Cisco Unified Communications deployment. These endpoints range from gateways that support ordinary analog phones in an IP environment to an extensive set of native IP phones offering a range of capabilities.

When deploying endpoints, you need to consider several factors, including authentication, upgrades, signaling protocol, Quality of Service (QoS), and so forth. The Unified Communications system must be designed appropriately to accommodate these factors.

This chapter summarizes various types of Unified Communications endpoints and covers design and deployment considerations including high availability and capacity planning. The Unified Communications endpoints covered in this chapter can be categorized into the following major types:

The sections listed above provide information about each endpoint type, including deployment considerations. That information is followed by a discussion related to high availability, capacity planning, and design considerations for effectively deploying endpoints.

Use this chapter to understand the range of available endpoint types and the high-level design considerations that go along with their deployment.

What's New in This Chapter

Table 18-1 lists the topics that are new in this chapter or that have changed significantly from previous releases of this document.

Table 18-1 New or Changed Information Since the Previous Release of This Document

New or Revised Topic	Described in:	Revision Date
A few minor updates	Various sections	April 30, 2013
Minor updates related to QoS, WLAN video call capacity, and lack of Cisco EX Series endpoint registration redundancy	Various sections throughout this chapter	October 31, 2012
Removal of product-specific content such as product support lists, QoS example configurations, and endpoint feature summary tables	Various sections throughout this chapter	June 28, 2012

Unified Communications Endpoints Architecture

Call signaling in Cisco Unified Communications Manager (Unified CM), Cisco Business Edition, and Cisco Unified Communications Manager Express (Unified CME) distinguishes between line-side signaling and trunk-side signaling. Whereas trunk-side signaling is used for connecting the entire call processing cluster or router to other servers and gateways, the line side is used for connecting end-user devices to the call processing platform. The two interfaces are distinct in the services they offer, with the line side offering a rich set of user-oriented features.

Session Initiation Protocol (SIP) and Skinny Client Control Protocol (SCCP) are the two main line-side signaling protocols supported by Cisco call processing platforms. All Cisco endpoints support either or both of these protocols. The set of features supported in both protocols is roughly equivalent, and the choice of which protocol to use is essentially a personal preference in a deployment. Cisco endpoints must be configured with several operating parameters before they can be used to make or receive calls or to run applications. This configuration must be performed in advance on the call processing server or router. Once configured, the call processing platform generates a configuration file for the endpoint to use, and it stores that file on a Trivial File Transfer Protocol (TFTP) server. The endpoints themselves go through a boot-up sequence when powered on. They retrieve this configuration file before they register with the appropriate server, and then they are ready to be used. The endpoints execute the following steps as part of the boot-up sequence:

1. When connected to the access switch, if the endpoint is not plugged in to a power source, it attempts to obtain power from the switch (Power over Ethernet).
2. Once power is obtained, if device security is enabled, the endpoint presents its credentials to the security server.
3. If it is allowed to use the network, the endpoint obtains its network parameters such as IP address, Domain Name Service (DNS) servers, gateway address, and so forth, either through static provisioning in the endpoint or through Dynamic Host Control Protocol (DHCP).
4. The endpoint also obtains a TFTP server address either through static provisioning or through DHCP options.
5. The endpoint then uses the TFTP server address to obtain its configuration files that, among other parameters, details the call processing server(s) or router(s) that the endpoint may associate with, the directory numbers that the endpoint must support, and so forth.
6. The endpoint registers with the call processing platform and is available for use.

Analog Gateways

An analog gateway typically is used to connect analog devices such as fax machines, modems, telecommunications device for the deaf (TDD)/teletypewriter (TTY), and analog phones, to the VoIP network so that the analog signal can be packetized and transmitted over the IP network. Analog gateways also provide physical connectivity to the PSTN and other traditional telephony equipment such as PBXs and key systems. Analog gateways include Cisco IOS router-based analog interface or service modules as well as fixed-port standalone gateways Generally analog gateways rely on Cisco Unified CM, Cisco Business Edition, Unified CM Express, and even Survivable Remote Site Telephony (SRST) for call control, supplementary services, and in some cases interface registration and configuration. Call control protocol support across Cisco analog gateways includes SIP, H.323, SCCP, and Media Gateway Control Protocol (MGCP).

Standalone Analog Gateways

Cisco standalone analog gateways, including the Cisco Analog Telephony Adapter (ATA) and Cisco VG200 Series Gateway, provide connectivity for analog devices such as fax machines, modems, TDD/TTY and analog phones, as well as one or more Ethernet ports for connecting to the IP network. Cisco standalone analog gateways support the FXS analog telephony interface port type only.

For more information on Cisco ATAs, refer to the data sheets and documentation at

http://www.cisco.com/en/US/products/hw/gatecont/ps514/index.html

For more information on Cisco VG200 Series Gateways, refer to the data sheets and documentation at

http://www.cisco.com/en/US/products/hw/gatecont/ps2250/prod_literature.html

Analog Interface Module

Cisco IOS router-based analog interface modules, including network modules (NMs) and voice interface cards (VICs), connect the PSTN and other legacy telephony equipment, including PBXs, analog telephones, fax machines, and key systems, to Cisco multiservice access routers such as the Cisco Integrated Services Router (ISR). Cisco IOS analog interface modules support a wide range of analog telephony interface port types, including FXS, FXO, T1/E1, E&M, and BRI.

Cisco IOS version support is critical for successful deployment of analog interface modules. For more information on Cisco IOS-based analog interface modules, including interface port type and Cisco IOS version support, refer to the data sheets and documentation listed at

http://www.cisco.com/en/US/products/ps10537/products_relevant_interfaces_and_modules.html#analogdigital

Analog Gateway Quality of Service Considerations

When configuring network-level quality of service (QoS), Cisco analog gateways such as the standalone Cisco VG200 Series and the Cisco IOS-based analog interface modules can be trusted and their packet markings honored. By default they mark their voice media and signaling packets with appropriate Layer 3 values (voice media as DSCP 46 or PHB EF; call signaling as DSCP 24 or PHB CS3), which match Cisco QoS recommendations for appropriate voice media and signaling marking, so as to ensure end-to-end voice quality on a converged network.

Desk Phones

The Cisco Unified IP Phone portfolio includes the following family of desk phones:

- Cisco Unified IP Phone 7900 Series, page 18-4
- Cisco Unified IP Phone 6900 Series, page 18-5
- Cisco Unified IP Phone 8900 and 9900 Series, page 18-5
- Cisco Unified SIP Phone 3900 Series, page 18-6

Cisco Unified IP Phone 7900 Series

The Cisco Unified IP Phone 7900 Series of endpoints consists of a wide range of models and feature sets. Models range from small single-line phones such as the Cisco Unified IP Phone 7911G to large eight-line phones such as the Cisco Unified IP Phone 7975G. In addition to the obvious and expected size differences between the various phone models, they vary in many other ways including: whether they have an LCD display and, if so, what size; whether they have a built-in speakerphone; what speed the network port(s) supports and whether there is a port for PC network attachment; how many phone lines they support; how many fixed feature keys they have and whether they are programmable; and so forth. In general, all phones in the Unified IP Phone 7900 Series provide the same basic set of enterprise IP telephony features such as call hold, call transfer, call forwarding, and so forth. However, some phone models provide features and functions well beyond the traditional enterprise IP telephony feature set, including support for IP-based phone services to enable presence, messaging, mobility, security, and other network-based applications and services. The Cisco Unified IP 7900 Series supports both SCCP and SIP signaling protocols for registering and communicating with the Cisco call processing platforms.

In some cases additional line keys can be added to Unified IP Phone 7900 Series devices by physically attaching a key expansion module such the Cisco Unified IP Phone Expansion Module 7916. This gives administrative assistants and other users the ability to answer and/or determine the status of a number of lines beyond the current line capability of their desk phone. Some Unified IP Phone 7900 Series models are capable of supporting up to two Cisco Unified IP Phone Expansion Modules, but the use of an external power adaptor may be required.

Note When two Expansion Modules are used with a single phone, the second module must be the same model as the first one.

The Cisco Unified IP Conference Station 7937G, with its 360-degree room coverage, provides conference room speaker-phone technology for use in conferencing environments. The Unified IP Conference Station provides an external speaker and built-in microphones. Optional extension microphones can be added to extend microphone coverage in larger rooms. These devices support SCCP signaling protocols for registering and communicating with the Cisco call processing platforms.

For more information about the Cisco Unified IP Phone 7900 Series, refer to the data sheets and documentation at

http://www.cisco.com/en/US/products/hw/phones/ps379/index.html

Cisco Unified IP Phone 6900 Series

The Cisco Unified IP Phone 6900 Series of endpoints includes a number of models. Models range from small, basic single-line phones such as the Cisco Unified IP Phone 6901 to larger, more advanced 12-line phones such as the Cisco Unified IP Phone 6961. In addition to the obvious and expected size differences between these various phone models, they vary in many other ways, including whether they have LCD displays, built-in speakerphone, PC port, and so forth. In general, all of the phones in the Unified IP Phone 6900 Series provide the same basic set of enterprise IP telephony features like such as hold, call transfer, call forwarding, and so forth. The Cisco Unified IP 6900 Series supports both SCCP and SIP signaling protocols for registering and communicating with the Cisco call processing platforms.

For more information about the Cisco Unified IP Phone 6900 Series, refer to the data sheets and product documentation at

http://www.cisco.com/en/US/products/ps10326/index.html

Deployment Considerations for the Cisco Unified IP Phone 6900 Series

The Cisco Unified IP Phone 6900 Series provides call features such as Direct Transfer and Direct Transfer Across Lines as well as the Join and Join Across Lines features. These features can operate over calls spanning multiple lines, and their operation can be opaque to CTI applications that monitor only the primary line on the phone. Therefore, in order for these applications to work properly and maintain control over the phone functions, it might be necessary to disable the call features. These features may be disabled, in decreasing priority order, in either the specific phone configuration, the Common Device Profile configuration applicable to a group of phones that share the profile, or the enterprise-wide phone configuration.

Cisco Unified IP Phone 8900 and 9900 Series

The Cisco Unified IP Phone 8900 and 9900 Series of endpoints provide a wide range of form-factors and physical characteristics, including models with and without LCD displays and speaker phones and with varying numbers of line keys. Likewise, some models in this series provide support for Bluetooth and/or 802.11, such as the Cisco Unified IP Phone 9971, while others do not. In general, all of the phones in the Unified IP Phone 8900 and 9900 Series provide the same set of enterprise IP telephony features such as call hold, call transfer, call forwarding, and so forth. The Cisco Unified IP Phone 8900 Series supports either SIP only or both SCCP and SIP signaling protocols (dependent on phone model) for registering and communicating with Cisco call processing platforms. The Cisco Unified IP Phone 9900 Series models support only SIP signaling when registering and communicating with call control.

For more information about the Cisco Unified IP Phone 8900 Series, refer to the data sheets and product documentation at

http://www.cisco.com/en/US/products/ps10451/index.html

For more information about the Cisco Unified IP Phone 9900 Series, refer to the data sheets and product documentation at

http://www.cisco.com/en/US/products/ps10453/index.html

The Cisco Unified IP 8900 and 9900 Series devices may also be equipped with up to three (dependent on phone model) Cisco Unified IP Color Key Expansion Modules for administrative assistants and others user who need to answer and/or determine the status of a number of lines beyond the current line capability of their phone. These modules extend the capability of the Cisco Unified IP Phone 8900 and 9900 Series desk phones by adding an additional LCDs and buttons.

Some Cisco Unified IP Phone 8900 and 9900 Series models provide video capabilities either through a built-in camera for the Cisco Unified IP Phone 8900 Series or through the Cisco Unified Video Camera add-on accessory for the Cisco Unified IP Phone 9900 Series.

For more information about the Cisco Unified IP Phone 9900 and 8900 Series accessories, refer to the data sheets and product documentation at

http://www.cisco.com/en/US/products/ps10655/index.html

Deployment Considerations for the Cisco Unified IP Phone 8900 and 9900 Series

The Cisco Unified IP Phone 8900 and 9900 Series provide call capabilities that generate JTAPI events that must be handled by applications that monitor the phone through CTI. These call features allow the user to cancel an in-progress transfer or conference, or to perform a join or direct transfer of calls across the same or different lines. If the monitoring applications have not been upgraded to versions that properly handle these events, unexpected application behavior could result, including applications that no longer have their view of the phone or call state in synchronization with the phone itself. Therefore, by default, all applications are restricted from monitoring or controlling these phones.

For applications that have been upgraded to properly handle these new events, or for applications that have verified that they are not impacted by these events, the administrator may enable the role of **Standard CTI Allow Control of Phones supporting Connected Xfer and conf** in the application or end-user configuration associated with the application. Only after this role has been enabled can the application monitor or control these phones.

Cisco Unified SIP Phone 3900 Series

The Cisco Unified SIP Phone 3900 Series provides cost-effective, entry-level endpoints that support a single line and provide a basic set of enterprise IP telephony capabilities and basic supplementary features such as mute, call hold, and call transfer. The Cisco Unified SIP Phone 3900 Series has a two-line liquid crystal display (LCD) screen and a half-duplex or full-duplex speakerphone (depending on the model). The Cisco Unified SIP Phone 3900 Series supports the SIP signaling protocols for registering and communicating with Cisco call processing platforms. For more information about the Cisco Unified SIP Phone 3900 Series, refer to the data sheets and documentation at

http://www.cisco.com/en/US/products/ps7193/index.html

Deployment Considerations for Cisco Desk Phones

The following sections list important design considerations when deploying Cisco desk phones.

For the latest information on features and functions supported for Cisco desk phone models, refer to the product data sheets and documentation available at

http://www.cisco.com/go/ipphones

Firmware Upgrades

Most commonly, and by default, IP phones upgrade their images using TFTP, which is a UDP-based protocol, from TFTP servers integrated into one or more of the call processing platforms. With this arrangement, all the phones obtain their images directly from these TFTP servers. This method works well for a relatively small number of phones or if all of the phones are located in a single campus region that has a LAN environment with essentially unlimited bandwidth.

For larger deployments that use centralized call processing, upgrading phones in branch offices that are connected to the central data center by low-speed WAN links, can require a large amount of data traffic over the WAN. The same set of files will have to traverse the WAN multiple times, once for each phone. Transferring this amount of data is not only wasteful of the WAN bandwidth but can also take a long time as each data transfer competes with the others for bandwidth. Moreover, due to the nature of TFTP protocol, some phones might be forced to abort their upgrades and fall back to the existing version of the code.

Note During the upgrade, the Cisco Unified IP Phones 9900 and 8900 Series stay in service, unlike the 7900 Series phones. The 9900 and 8900 Series phones download and store the new firmware in their memory while still maintaining their active status, and they reboot with the new firmware only after a successful download.

Two methods are available to alleviate problems created by the need to upgrade phones over the WAN. One method is to use a local TFTP server just for the upgrades. The administrator can place a TFTP server in branch offices (particularly in branches that have a larger number of phones, or whose WAN link is not speedy or robust), and can configure the phones in those offices to use that particular TFTP server just for new firmware. With this change, phones will retrieve new firmware locally. This upgrade method would require the administrator to pre-load the phone firmware on the TFTP server in the branch and manually configure the TFTP server address in the **load server** parameter in the affected phone configurations. Note that the branch router may be used as a TFTP server.

The second method to upgrade phones without using the WAN resources excessively is to use the Peer File Sharing (PFS) feature. In this feature, typically only one phone of each model in the branch downloads each new firmware file from the central TFTP server. Once the phone downloads the firmware file, it distributes that file to other phones in the branch. This method avoids the manual loading and configuration required for the load server method.

The PFS feature works when the same phone models in the same branch subnet arrange themselves in a hierarchy (chain) when asked to upgrade. They do this by exchanging messages between themselves and selecting the "root" phone that will actually perform the download. The root phone sends the firmware file to the second phone in the chain using a TCP connection; the second phone sends the firmware file to the third phone in the chain, and so on until all of the phones in the chain are upgraded. Note that the root phone may be different for different files that make up the complete phone firmware.

Power Over Ethernet

Deploying desk phones with inline power-capable switches enables these endpoints to derive power over the Ethernet network connection, thus eliminating the need for an external power supply as well as a wall power outlet. Inline power-capable switches with uninterruptible power supplies (UPS) ensures that power over Ethernet (PoE) capable IP desk phones continue to receive power during power failure situations. Provided the rest of the telephony network is available during these periods of power failure, then IP phones should be able to continue making and receiving calls.

Depending on the type of desk phone and the PoE standard supported by both the desk phone and the inline power-capable switch, in some cases the power budget of the inline powered switch port may be exceeded. This typically occurs when attaching key extension modules or other power consuming attachments such as USB cameras. In these situations, the phone may need to be powered using a wall outlet and external power supply or else the switch providing the power may need to be upgraded.

Note In addition to using the inline power from the access switch or local wall power, a Cisco Unified IP Phone can also be supplied power by a Cisco Unified IP Phone power injector. The Cisco Unified IP Phone power injector connects Cisco Unified IP Phones to Cisco switches that do not support inline power or to non-Cisco switches. The Cisco Unified IP Phone power injector is compatible with most Cisco Unified IP Phones. It has two 10/100/1000 Base-T Ethernet ports. One Ethernet port connects to the switch access port and the other connects to the Cisco Unified IP Phone.

Quality of Service

When configuring network-level quality of service (QoS), Cisco desk phones such as the Cisco Unified IP Phone 7900, 8900, and 9900 Series can be trusted and their packet markings honored. By default these endpoints mark their voice media and signaling packets with appropriate Layer 3 values (voice media as DSCP 46 or PHB EF; call signaling as DSCP 24 or PHB CS3), which match Cisco QoS recommendations for appropriate voice media and signaling marking, to ensure end-to-end voice quality on a converged network. While many Cisco desk phones support the attachment of a desktop computer, Cisco desk phones are capable of separating the voice and data traffic, placing voice traffic onto the voice VLAN and data traffic from the desktop onto the data VLAN. This enables the network to extend trust to the phone but not to the PC port of the phone, as per Cisco recommendations.

Note While many Cisco desk phones support Link Layer Discovery Protocol for Media Endpoint Devices (LLDP-MED), they do so only for VLAN and Power over Ethernet negotiation. Cisco Unified IP Phones do not honor DSCP and CoS markings provided by LLDP-MED.

SRST and Unified CME as SRST

When deploying Cisco desk phones in branch locations separated from a centralized call processing platform by a low-speed or unreliable WAN link, it is important to consider local call processing redundancy. By leveraging Survivable Remote Site Telephony (SRST) or Cisco Unified Communications Manager Express (Unified CME) as SRST on a Cisco IOS router in each branch location, basic IP telephony services can be maintained for the desk phones when connectivity to the centralized call processing platform is lost. However, the set of available user-facing features is much smaller when a device is registered to SRST than when the phone is registered to Unified CM.

Software-Based Endpoints

A software-based endpoint is an application installed on a client desktop computer that registers and communicates with Cisco call processing platforms for voice and video services. In addition, these endpoint software client applications may provide collaboration features and services such as messaging, presence, directory access, and conferencing. Software-based endpoint desktop client applications include Cisco IP Communicator as well as Cisco WebEx Connect and Cisco Jabber, all of which use the Cisco Unified Client Services Framework (CSF).

Cisco IP Communicator

Cisco IP Communicator is a Microsoft Windows-based application that provides enterprise IP phone functionality to desktop computers. This application provides enterprise-class IP voice calling for remote users, telecommuters, and other mobile users. Cisco IP Communicator supports both SCCP and SIP signaling protocols for registering and communicating with Cisco call processing platforms. For more information about Cisco IP Communicator, refer to the data sheets and product documentation at

http://www.cisco.com/en/US/products/sw/voicesw/ps5475/index.html

Cisco Unified Client Services Framework

Cisco Unified Client Services Framework (CSF) is a software application that provides an underlying framework for integration of Unified Communications services, including audio, video, web collaboration, visual voicemail, and so forth, into a software-based desktop application. The Cisco Unified Client Services Framework allows desktop application users to access a variety of communication and collaboration services as provided by back-end collaboration application servers such as Cisco Unified Communications Manager (Unified CM), Cisco Unity Connection, Cisco WebEx, and Lightweight Directory Access Protocol (LDAP)-compliant directories. The Cisco Unified Client Services Framework is a device type in Cisco Unified CM that enables phone registration and communication for Cisco Unified Communications Integration for Cisco WebEx Connect and Cisco Jabber desktop applications, and it operates in either softphone mode or deskphone mode to control a Cisco Unified IP Phone.

Softphone Mode of Operation

For the Cisco WebEx Connect and Cisco Jabber desktop applications to operate in softphone mode, a Cisco Unified Client Services Framework device must be configured in Cisco Unified CM. The Cisco Unified Client Services Framework will then enable the Cisco Jabber and Cisco Unified Communications Integration for Cisco WebEx Connect applications to operate as a SIP-based single-line Cisco Unified IP Phone and will support the full registration and redundancy mechanisms of a Cisco Unified IP Phone.

Deskphone Control Mode of Operation

When the Cisco Jabber or Cisco WebEx Connect desktop application operates in deskphone control mode, the application uses CTI/JTAPI to control an associated Cisco Unified IP Phone. The Unified Client Services Framework uses the Cisco CallManager Cisco IP Phone Services (CCMCIP) service from Unified CM to provide a listing of valid Cisco Unified IP Phones to control.

The following design considerations should be taken into account when deploying Cisco Jabber and other desktop applications that use the Cisco Unified Client Services Framework:

- The administrator must determine how to install, deploy, and configure the Unified Client Services Framework desktop applications in their organization. Cisco recommends using a well-known installation package such as Altris to install the desktop application, and use Group Policies to configure the user registry settings for the required components such as TFTP, CTI Manager, CCMCIP, and LDAP server IP addresses and other pertinent information.
- The user ID and password configuration of the Cisco Unified Client Services Framework desktop application user must match the user ID and password of the user stored in the LDAP server to allow for seamless integration of the Unified Communications and back-end directory components.
- The directory number configuration on Cisco Unified CM and the telephoneNumber attribute in LDAP should be configured with a full E.164 number. A private enterprise dial plan can be used, but it might involve the need to use application dial rules and directory lookup rules.
- The deskphone mode for control of a Cisco Unified IP Phone uses CTI; therefore, when sizing a Unified CM deployment, you must also account for other applications that require CTI usage. For more information on CTI system sizing, refer to the section on Applications and CTI, page 29-30.

For additional information about the Cisco WebEx Messenger service (formerly Cisco WebEx Connect service), Cisco WebEx Connect, Cisco Jabber, and Cisco Unified Client Services Framework, see the chapter on Cisco Collaboration Clients and Applications, page 24-1.

For more information about Cisco Jabber for Windows, refer to the data sheets and product documentation at

http://www.cisco.com/en/US/products/ps12511/index.html

For more information about the Cisco Jabber for Mac, refer to the data sheets and product documentation at

http://www.cisco.com/en/US/products/ps11764/index.html

For more information about the Cisco WebEx Messenger service, Cisco WebEx Connect, and Cisco Unified Communications Integration, refer to the product information at

http://www.cisco.com/en/US/products/ps10528/index.html

General Deployment Considerations for Software-Based Endpoints

The following sections list important design considerations for deploying software-based endpoints.

For the latest information on the features and functions supported for software-based endpoints, refer to the product data sheets and documentation available at

http://www.cisco.com/en/US/products/ps6789/Products_Sub_Category_Home.html

Quality of Service

While some software-based client applications do mark their traffic in accordance with QoS marking best practices, many applications do not. Further, even when the application does properly mark traffic, the underlying operating system or hardware may not honor the markings. Given the general unpredictability and unreliability of traffic marking coming from desktop computers, as a general rule these traffic markings should not be trusted. This means that all traffic flows must be re-marked by the network based on protocol and/or port numbers, with real-time traffic flows being marked based on best practices. This includes re-marking of voice-only call media with DSC 46 or PHB EF, video call media (including voice) with DSCP 34 or PHB AF41, and call signaling with DSCP 24 or PHB CS3. These markings along with a properly configured network infrastructure ensure priority treatment for voice-only call media and dedicated bandwidth for video call media and call signaling. In addition to re-marking of software-based endpoint traffic, Cisco recommends using network-based policing and rate limiting to ensure that the software-based endpoint does not consume too much network bandwidth. This can occur when the desktop computer generates too much data traffic or when the endpoint application misbehaves and generates more voice and/or video media and signaling traffic than would be expected for a typical call. In cases where third-party software is used to fully control desktop computer network traffic marking, administrators may decide to trust desktop computer marking, in which case re-marking of packets would not be required. Network-based policing and rate limiting is still recommended to protect the overall network in case of a misbehaving endpoint.

Inter-VLAN Routing

Because software-based endpoints run on a desktop computer usually deployed on a data VLAN, when software-based endpoints are deployed on networks with voice and data VLAN separation, inter-VLAN routing should be configured and allowed so that voice traffic from these endpoints on the data VLAN can reach endpoints on the voice VLAN.

SRST and Unified CME as SRST

When deploying Cisco software-based endpoint desktop applications in branch locations separated from a centralized call processing platform by a low-speed or unreliable WAN link, it is important to consider local call processing redundancy. By using SRST or Unified CME as SRST on a Cisco IOS router in each branch location, basic IP telephony services can be maintained for software-based endpoints when connectivity to the centralized call processing platform is lost. However, the set of available user-facing features is much smaller when a desktop software-based endpoint is registered to SRST than when the application is registered to Unified CM.

Wireless Endpoints

Cisco wireless endpoints rely on an 802.11 wireless LAN (WLAN) infrastructure for network connectivity and to provide IP telephony functionality and features. This type of endpoint is ideal for mobile users that move around within a single enterprise location or between enterprise locations or environments where traditional wired phones are undesirable or problematic. Cisco offers the following voice and video over WLAN (VVoWLAN) IP phones:

- Cisco Unified Wireless IP Phones, including the Cisco Unified Wireless IP Phone 7925G and 7926G
- Cisco Unified IP Phone 9971

All are hardware-based phones with built-in radio antenna. The Cisco Unified Wireless IP Phones as well as the wirelessly attached Cisco Unified IP Phone 9971 enable 802.11b, 802.11g, or 802.11a connectivity to the network. The Cisco Unified Wireless IP Phones register and communicate with Cisco call processing platforms using SCCP signaling protocol, while the Cisco Unified IP Phone 9971 uses the SIP signaling protocol to register and communicate with Cisco call processing platforms.

For more information about the Cisco Unified Wireless IP Phones, refer to the data sheets and product documentation available at

http://www.cisco.com/en/US/products/hw/phones/ps379/index.html

For more information about the Cisco Unified IP Phone 9971, refer to the data sheets and product documentation available at

http://www.cisco.com/en/US/products/ps10453/index.html

General Deployment Considerations for Wireless Endpoints

The following sections list important design considerations for deploying wireless endpoints.

For the latest information on features and functions of wireless IP endpoints such as the Cisco Unified Wireless IP Phone 7925G, and for more information about deploying wireless IP endpoints, refer to the deployment guides at

http://www.cisco.com/en/US/products/hw/phones/ps379/products_implementation_design_guides_list.html

For the latest information on features and functions of the Cisco Unified IP Phone 9971 and for more information about deploying it wirelessly, refer to the deployment guide at

http://www.cisco.com/en/US/products/ps10453/products_implementation_design_guides_list.html

Network Radio Frequency Design and Site Survey

Before deploying wireless endpoints, you must ensure your WLAN radio frequency (RF) design minimizes same-channel interference while also providing sufficient radio signal levels and non-adjacent channel overlap so that acceptable voice and video quality can be maintained as the device moves from one location to another. In addition, you must perform a complete WLAN site survey to verify network RF design and to ensure that appropriate data rates and security mechanisms are in place. Your site survey should take into consideration which types of antennas will provide the best coverage, as well as where sources of RF interference might exist. Even when using third-party site survey tools, Cisco highly recommends that you verify the site survey using the wireless endpoint device itself because each endpoint or client radio can behave differently depending on antenna sensitivity and survey application limitations. Cisco Unified Wireless IP Phones and the Cisco Unified IP Phone 9971 provide a built-in site survey tool that enables easy verification of the surrounding WLAN network channels and signal strength. Cisco recommends relying on the 5 GHz WLAN band (802.11a/n) whenever possible for connecting wireless endpoints capable of generating voice and video traffic. 5 GHz WLANs provide better throughput and less interference for voice and video calls. Refer to the section on Wireless LAN Infrastructure, page 3-54, for more information about wireless network design.

Security: Authentication and Encryption

When deploying wireless endpoints, it is important to consider the security mechanisms used to control access to the network and to protect the network traffic. Cisco wireless endpoints support a wide range of authentication and encryption protocols including WPA, WPA2, EAP-FAST, PEAP, and so forth. Choose an authentication and encryption method that is supported by the WLAN infrastructure and the endpoint devices you deploy, and one that aligns with IT security policies. In addition, ensure that the authentication and encryption method chosen supports a fast rekeying method such as Cisco Centralized Key Management (CCKM) so that active voice and video calls can be maintained when the device is roaming from one location in the network to another.

Note In dual-band WLANs (those with both 2.4 GHz and 5 GHz bands), it is possible to roam between 802.11b/g and 802.11a with the same SSID, provided the client is capable of supporting both bands. However, with some devices this can cause gaps in the voice or video path. In order to avoid these gaps, use only one band for voice and video communications.

Wireless Call Capacity

When deploying wireless devices and enabling wireless device roaming within the enterprise WLAN, it is also important to consider the device connectivity and call capacity of the WLAN infrastructure. Oversubscription of the WLAN infrastructure in terms of number of devices or number of active calls will result in dropped wireless connections, poor voice and video quality, and delayed or failed call setup. The chances of oversubscribing a deployment of voice and video over WLAN are greatly minimized by deploying sufficient numbers of WLAN access points (APs) to handle required call capacities. AP call capacities are based on the number of simultaneous bidirectional streams that can be supported in a single channel cell area. The general rule for VVoWLAN call capacities is as follows:

- Maximum of 27 simultaneous VoWLAN bidirectional streams per 802.11g/n (2.4 GHz) channel cell with Bluetooth disabled or per 802.11a/n (5 GHz) channel and 24 Mbps or higher data rates enabled.
- Maximum of 8 simultaneous VVoWLAN bidirectional streams per 802.11 g/n (2.4 GHz) channel cell with Bluetooth disabled or per 802.11 a/n (5 GHz) channel cell assuming a video resolution of 720p (high-definition) and video bit rate of up to 1 Mbps.

These call capacity values are highly dependent upon the RF environment, the wireless handset features, and underlying WLAN system features. Actual capacities for a particular deployment could be less.

Note A single call between two wireless endpoints associated to the same AP is considered to be two simultaneous bidirectional streams.

The above capacities are based on voice activity detection (VAD) being disabled and a packetization sample size of 20 milliseconds (ms). VAD is a mechanism for conserving bandwidth by not sending RTP packets while no speech is occurring during the call. However, enabling or disabling VAD, also referred to as Silence Suppression, is sometimes a global configuration depending on the Cisco call control platforms. Thus, if VAD is enabled for wirelessly attached Cisco Unified IP Phones, then it may be enabled for all devices in the deployment. Cisco recommends leaving VAD (Silence Suppression) disabled to provide better overall voice quality.

At a sampling rate of 20 ms, a voice call will generate 50 packets per second (pps) in either direction. Cisco recommends setting the sample rate to 20 ms for almost all cases. By using a larger sample size (for example, 30 or 40 ms), you can increase the number of simultaneous calls per AP, but a larger end-to-end delay will result. In addition, the percentage of acceptable voice packet loss within a wireless environment decreases dramatically with a larger sample size because more of the conversation is missing when a packet is lost. For more information about voice sampling size, see the section on Bandwidth Provisioning, page 3-45.

Bluetooth Support

The Cisco Unified Wireless IP Phones 7925G, 7925G-EX, and 7926G, and the Cisco Unified IP Phone 9971 are Bluetooth-enabled devices. The Bluetooth radio or module within these wireless Cisco Unified IP Phones provides the ability to support Bluetooth headsets with the phones. Because Bluetooth devices use the same 2.4 GHz radio band as 802.11b/g devices, it is possible that Bluetooth and 802.11b/g devices can interfere with each other, thus resulting in connectivity issues.

While the Bluetooth and 802.11 WLAN radios co-exist in the Cisco Unified Wireless IP Phones and Cisco Unified IP Phone 9971, greatly reducing and avoiding radio interference between the Bluetooth and 802.11b/g radio, the Bluetooth radio in these wirelessly attached phones can cause interference for other 802.11b/g devices deployed in close proximity. Due to the potential for interference and disruption of 802.11b/g WLAN voice and video devices (which can result in poor voice and video quality, deregistration, and/or call setup delays), Cisco recommends deploying all WLAN voice and video devices on 802.11a, which uses the 5 GHz radio band. By deploying wireless phones on the 802.11a radio band, you can avoid interference caused by Bluetooth devices.

Note Using Bluetooth wireless headsets with the battery-powered Cisco Unified Wireless IP Phones will increase battery power consumption on your phone and will result in reduced battery life.

Quality of Service

When configuring network-level quality of service (QoS), Cisco wireless endpoints (including Cisco Unified Wireless IP Phones and the Cisco Unified IP Phone 9971) can be trusted and their packet markings honored. By default these endpoints mark the recommended and appropriate Layer 3 values for voice media and signaling (voice media as DSCP 46 or PHB EF; voice signaling as DSCP 24 or PHB CS3). Likewise, these devices mark appropriately at Layer 2 (voice media WMM User Priority (UP) of 6; call signaling WMM UP of 4). With these packet markings, end-to-end voice quality on the converged network will be acceptable.

SRST and Unified CME as SRST

When deploying wireless endpoints in branch locations separated from a centralized call processing platform by a low-speed or unreliable WAN link, it is important to consider local call processing redundancy. By deploying SRST or Unified CME as SRST on a Cisco IOS router in each branch location, basic IP telephony services can be maintained for wireless endpoints when connectivity to the centralized call processing platform is lost. However, the set of available user-facing features is much smaller when a wireless endpoint is registered to SRST than when it is registered to Unified CM.

Device Mobility

When wireless endpoints move between locations in a multi-site centralized call processing deployment, the Cisco Unified CM Device Mobility feature may be used to dynamically update the location of the device based on the IP address the device uses to register to Unified CM. This prevents issues with call routing, PSTN egress, and codec and media resource selection typically encountered when devices move between locations. For more information on Device Mobility, see the section on Device Mobility, page 25-14.

Mobile Endpoints

Cisco mobile endpoint devices and mobile endpoint client applications register and communicate with Unified CM for voice and video calling services. These devices and clients also enable additional features and services such as enterprise messaging, presence, and corporate directory integration by communicating with other back-end systems such as Cisco Unity Connection, Cisco IM and Presence, and LDAP directories. Cisco offers the following mobile endpoint devices and clients:

- Cisco Cius, page 18-15
- Cisco Jabber for Android and Apple iOS, page 18-16
- Cisco Jabber IM, page 18-16, for Android, BlackBerry and Apple iOS devices

Cisco Cius

Cisco Cius is an Android-based enterprise tablet that provides native voice and video calling over a WLAN or mobile data network when registered to Unified CM as a SIP device. In addition to enabling enterprise voice and video calling, Cius has native applications for XMPP-based enterprise instant messaging (IM) and presence, corporate directory access, and visual voicemail. For more information about Cisco Cius, refer to the data sheets and product documentation available at

http://www.cisco.com/en/US/products/ps11156/index.html

For more information about deploying Cisco Cius wirelessly, refer to the *Cisco Cius Deployment Guide* available at

http://www.cisco.com/en/US/products/ps11156/products_implementation_design_guides_list.html

Cisco Jabber for Android and Apple iOS

The Cisco Jabber mobile clients for Android and Apple iOS devices including the iPhone and iPad enable smartphones and tablets to make and receive enterprise calls using voice or voice and video over IP. The Cisco Jabber mobile client application running on the Android or Apple iOS device registers and communicates with Unified CM using the SIP signaling protocol. In some cases the device may register and communicate with the Cisco Video Communications System (VCS) instead. The Cisco Jabber mobile client also enables additional features such as corporate directory access, enterprise visual voicemail, and in some cases enterprise instant messaging and presence.

For more information about Cisco Jabber for Android, refer to the data sheet and product documentation at

http://www.cisco.com/en/US/products/ps11678/index.html

For more information about Cisco Jabber for iPhone, refer to the data sheet and product documentation at

http://www.cisco.com/en/US/products/ps11596/index.html

For more information about Cisco Jabber for iPad, refer to the data sheet and product documentation at

http://www.cisco.com/en/US/products/ps12430/index.html

Cisco Jabber IM

The Cisco Jabber IM client runs on specific BlackBerry and Android smartphones and on various Apple iOS devices including the iPhone, and it communicates via XMPP with on-premises Cisco IM and Presence services or off-premises cloud-based Cisco WebEx Messenger service.

Note Cisco Jabber for iPad provides native XMPP-based IM and presence capabilities.

For more information about Cisco Jabber IM for Android, refer to the data sheet and product documentation at

http://www.cisco.com/en/US/products/ps11678/index.html

For more information about Cisco Jabber IM for BlackBerry, refer to the data sheet and product documentation at

http://www.cisco.com/en/US/products/ps11763/index.html

For more information about Cisco Jabber IM for iPhone, refer to the data sheet and product documentation at

http://www.cisco.com/en/US/products/ps11596/index.html

Deployment Considerations for Mobile Endpoints and Clients

The following sections list important design considerations for deploying mobile endpoints and clients.

For additional design and deployment information about Cisco Cius and Cisco Jabber mobile clients, refer to the section on Cisco Mobile Clients and Devices, page 25-60.

Cisco Jabber Client Application Interaction

Although Cisco Jabber for Android and iPhone smartphones require separate applications for enterprise voice and enterprise XMPP-based IM and presence services, both mobile client applications can be installed on the same device (co-resident). While they are separate client applications, they are aware of each other and will cross-launch each other as required. For example, an IM conversation on the Cisco Jabber IM application can be escalated to a voice call, causing the Cisco Jabber client application to become active in order to handle the call.

WLAN Design

Because Cisco Jabber mobile clients and Cisco Cius are often attached to a WLAN, all of the previously mentioned WLAN deployment considerations apply to mobile clients and devices, including WLAN RF design and verification by site survey. In particular, Cisco recommends relying on the 5 GHz WLAN band (802.11a/n) whenever possible for connecting wireless endpoints capable of generating voice and video traffic. 5 GHz WLANs provide better throughput and less interference for voice and video calls. If the 2.4 GHz band is used for mobile clients and devices, Bluetooth should be avoided. Likewise, the WLAN channel cell voice-only and video call capacity numbers covered in the section on Wireless Call Capacity, page 18-13, should be considered when deploying these clients and devices.

Secure Remote Enterprise Attachment

If appropriately deployed, Cisco mobile endpoints and clients can also connect to the enterprise from remote locations by using public or private 802.11 Wi-Fi hot spots or over the mobile data network. In these scenarios the Cisco AnyConnect mobile VPN client can be used to connect the device or client to the enterprise with a secure SSL tunnel.

Quality of Service

Cisco mobile client applications and devices generally mark Layer 3 QoS packet values in accordance with Cisco collaboration QoS marking recommendations. This includes marking voice-only call media traffic with DSCP 46 or PHB EF, video call media (including voice) traffic with DSCP 34 or PHB AF41, and call signaling traffic with DSCP 24 or PHB CS3. Despite appropriate mobile client and device application Layer 3 packet marking, Layer 2 802.11 WLAN packet marking (User Priority, or UP) presents further challenges. Some devices such as Cisco Cius appropriately mark wireless Layer 2 802.11 User Priority (UP) values (voice-only call media UP 6, video call media UP 5, and call signaling UP 3). However, because Cisco mobile clients run on a variety of mobile devices, Layer 2 wireless QoS marking is inconsistent and therefore cannot be relied upon to provide appropriate treatment to traffic on the WLAN. In deployments with Cisco Unified Wireless LAN Controllers, enabling wireless SIP call admission control (CAC) might provide some relief for incorrect or nonexistent Layer 2 WLAN marking. SIP CAC utilizes media session snooping and ensures that downstream voice and video frames are prioritized and/or treated correctly. Even assuming appropriate mobile client application Layer 3 or even Layer 2 packet marking, mobile devices present many of the same challenges as desktop computers in terms of generating many different types of traffic, including both data and real-time traffic. Given this, mobile devices generally fall into the untrusted category of collaboration endpoints. For deployments where mobile client devices are not considered trusted endpoints, packet re-marking based on traffic type and port numbers is required to ensure that network priority queuing and dedicated bandwidth are applied to appropriate traffic. In addition to re-marking the mobile device traffic, Cisco recommends using network-based policing and rate limiting to ensure that the mobile client devices do not consume too much network bandwidth.

Note Mobile clients and devices may attach remotely to the enterprise using Cisco AnyConnect client over the mobile data network or public or private Wi-Fi hot spots. Because these connections traverse the Internet, there is no end-to-end QoS on the IP path and therefore all traffic is treated as best-effort. Voice and video quality cannot be guaranteed over these types of connections.

SRST and Unified CME as SRST

When deploying mobile endpoints and clients such as Cisco Cius or Cisco Jabber for iPhone in branch locations separated from a centralized call processing platform by a low-speed or unreliable WAN link, it is important to consider local call processing redundancy. By deploying SRST or Unified CME as SRST on a Cisco IOS router in each branch location, basic IP telephony services can be maintained for mobile endpoints when connectivity to the centralized call processing platform is lost. However, the set of available user-facing features is much smaller when a mobile device is registered to SRST than when the device is registered to Unified CM. Not all Cisco Jabber mobile clients support SRST; however, because most Cisco Jabber mobile clients run on smartphones with cellular voice radios, users may still be able to make call using the mobile provider network.

Video Endpoints

Cisco video endpoints provide IP video telephony features and functions similar to IP voice telephony, enabling users to make point to point and point to multi-point video calls. Cisco offers the following desktop video-capable endpoints:

- Cisco Unified IP Phone 9900 Series with the optional USB camera attachment.
- Cisco Unified IP Phones 8941 and 8945 with built-in camera
- Cisco TelePresence System EX Series
- Cisco Unified Client Services Framework (CSF) software-based desktop clients such as Cisco Jabber for Windows

For additional information on video telephony, see the chapter on IP Video Telephony, page 12-1.

Cisco Unified IP Phone 8900 and 9900 Series

The Cisco Unified IP Phones 8900 and 9900 Series are capable of transmitting video and receiving and displaying video natively on their screens. With the built-in camera (8941 and 8945) or optional, specially designed USB camera attachment (9900 Series), they can also transmit video. The screens on these phones can display a variety of video resolutions and frame rates. The video capabilities of these phones can be enabled and disabled or tuned as desired from the Cisco call control platform configuration pages. These devices register and communicate with Unified CM using either SCCP or

SIP signaling protocols in the case of the 8900 Series video-capable phones or via SIP only in the case of the 9900 Series phones. For more information about the Cisco Unified IP Phone 8900 and 9900 Series video capabilities, refer to the data sheets and product documentation available at

http://www.cisco.com/en/US/products/ps10453/index.html

Cisco TelePresence System EX Series

The Cisco TelePresence System EX Series video endpoints provide personal telepresence or video calling for the desktop and include the Cisco TelePresence System EX60 and EX90. The EX Series video endpoint models vary in screen size as well as viewing angle and video resolution, but both deliver near equivalent telephony features. The Cisco TelePresence System EX Series video endpoints register and communicate with Unified CM by means of the SIP signaling protocol.

Note The EX Series video endpoints do not support registration redundancy with Unified CM. If the primary Unified CM node to which the EX Series endpoint is registered becomes unreachable, the endpoint will not fail-over its registration to a secondary Unified CM node.

For more information about the Cisco TelePresence System EX Series video endpoints, refer to the data sheets and product documentation available at

http://www.cisco.com/en/US/products/ps11327/index.html

Cisco Unified Client Services Framework (CSF) Video

Some Cisco Unified CSF software-based desktop clients such as Cisco Jabber for Windows are able to send and receive video when running on a desktop computer with an integrated or USB-attached camera. These video-capable software-based endpoints register and communicate with Unified CM call control and operate as a SIP single-line voice and video enabled phone. These endpoints support the primary and backup registration redundancy mechanisms as provided by Unified CM. The Cisco Unified CSF software-based endpoint processes video on the computer where it is installed. The quality of the decoding and encoding depends on the availability of CPU and memory resources on that computer.

For more information about the video capabilities of Cisco Jabber for Windows, refer to the data sheet and product documentation available at

http://www.cisco.com/en/US/products/ps12511/index.html

General Deployment Considerations for Video Endpoints

The following sections list important design considerations for deploying video endpoints.

Quality of Service

When configuring network-level quality of service (QoS), Cisco video endpoints (including Cisco Unified IP Phone 8900 and 9900 Series and Cisco TelePresence System EX Series devices) generally mark traffic at Layer 3 according to Cisco general QoS guidelines related to voice and video packet marking (video media as DSCP 34 or PHB AF41; call signaling as DSCP 24 or PHB CS3) and therefore these devices can be trusted. Even when trusting the endpoint marking, Cisco recommends using network-based policing and rate limiting to ensure that the video endpoint does not consume too much network bandwidth. Software-based video-capable endpoints do present challenges when they do not or cannot mark traffic appropriately. In these situations, typical guidance is to re-mark media and signaling traffic within the network from best-effort to appropriate and recommended values (voice media as DSCP 46 or PHB EF; video media as DSCP 34 or PHB AF41; call signaling as DSCP 24 or PHB CS3) based on protocols and/or port numbers. However, in some cases software-based applications may send voice and video media on the same ports. This means that it would not be possible to provide differentiated packet marking for the voice stream of an audio-only call from the voice stream of a video call.

Note While some Cisco video-capable endpoints support Link Layer Discovery Protocol for Media Endpoint Devices (LLDP-MED), they do so only for VLAN and Power over Ethernet negotiation. Cisco video endpoints do not honor DSCP and CoS markings provided by LLDP-MED.

Inter-VLAN Routing

When deploying video endpoints on networks with voice and data VLAN separation, it is important to consider software-based video-capable endpoints as well as hardware-based video endpoints that need to access resources. Because software-based endpoints running on a desktop computer are primarily attached to the data VLAN, inter-VLAN routing should be configured and allowed so that voice traffic from these endpoints on the data VLAN can reach endpoints on the voice VLAN. Likewise, if hardware-based video endpoints such as the Cisco TelePresence System EX60 need access to network resources such as directory or management services deployed on the data VLAN, inter-VLAN routing must be allowed.

SRST and Unified CME as SRST

When deploying video endpoints in branch locations separated from a centralized call processing platform by a low-speed or unreliable WAN link, it is important to consider local call processing redundancy. By deploying SRST or Unified CME as SRST on a Cisco IOS router in each branch location, basic IP telephony services can be maintained for most video endpoints when connectivity to the centralized call processing platform is lost. However, the set of available user-facing features is much smaller when a video endpoint is registered to SRST than when the application is registered to Unified CM. Specifically, video endpoint devices registered to SRST will be capable of making and receiving only voice calls (audio-only). SRST is not supported with the Cisco TelePresence System EX Series video endpoints.

Cisco Virtualization Experience Clients

The Cisco Virtualization Experience Clients (VXC) are the integral collaboration components of the Cisco Virtualization Experience Infrastructure (VXI). The VXCs provide user access to data applications and services across various network environments, as well as user preferences and device form factors for a fully integrated voice, video, and virtual desktop environment.

Cisco offers the following VXC endpoints:

- Cisco Virtualization Experience Client 2000 Series
- Cisco Virtualization Experience Client 4000 Series
- Cisco Virtualization Experience Client 6000 Series

For additional information about Cisco Virtualization Experience Clients, see the section on Cisco Virtualization Experience Client Architecture, page 24-29.

Cisco Virtualization Experience Client 2000 Series

The Cisco Virtualization Experience Client 2000 Series provides simple devices with a small firmware footprint (also known as a "zero client") for virtual desktop access to a Citrix or VMware environment. The VXC 2112 and 2212 are specifically designed to work in a Citrix environment, while the VXC 2111 and 2211 work in a VMware environment. The VXC 2111 and 2112 integrated form factor devices are designed to replace the footstand on the Cisco Unified IP Phones 8961, 9951, or 9971 to provide a fully integrated voice, video, and virtual desktop environment. The VXC 2211 and 2212 standalone form factor devices are designed to operate as a simple virtual desktop environment, or they can be paired with any third-generation Cisco Unified IP Phone to provide a full voice, video, and virtual desktop environment.

For more information about the Cisco Virtualization Experience Client 2000 Series, refer to the data sheets and product documentation available at

http://www.cisco.com/en/US/products/ps11499/index.html

Cisco Virtualization Experience Client 4000 Series

The Cisco Virtualization Experience Client (VXC) 4000 Series software appliance, when used in conjunction with a repurposed PC, allows for secure access to a remote hosted virtual desktop while supporting rich media locally. Windows 7 and Windows XP are the only operating systems supported for the repurposed PC. The hosted virtual desktop is supported, using Citrix XenDesktop or VMware View, through a locally installed thick-client Citrix Receiver and VMware View Client, respectively.

For more information about the Cisco Virtualization Experience Client 4000 Series, refer to the data sheets and product documentation available at

http://www.cisco.com/en/US/products/ps11498/index.html

Cisco Virtualization Experience Client 6000 Series

The Cisco Virtualization Experience Client (VXC) 6000 Series thin client provides a fully integrated voice, video, and virtual desktop solution in a single device. The VXC 6215 is a Linux platform that can be used in Virtual Desktop Infrastructure (VDI) mode to support Citrix XenDesktop or VMware View, or it can be enabled for Unified Communications with a software appliance add-on that allows for full voice, video, and virtual desktop support of Citrix XenDesktop.

For more information about the Cisco Virtualization Experience Client 6000 Series, refer to the data sheets and product documentation available at

http://www.cisco.com/en/US/products/ps11976/index.html

Third-Party IP Phones

Some third-party IP phones and devices may be integrated with Cisco call control to provide basic IP telephony functionality, as described in this section.

Third-Party SIP IP Phones

Third-party phones have specific local features that are independent of the call control signaling protocol, such as features access buttons (fixed or variable). Basic SIP RFC support allows for certain desktop features to be the same as on Cisco Unified IP Phones and also allows for interoperability of certain features. However, these third-party SIP phones do not provide the full feature functionality of Cisco Unified IP Phones.

Cisco works with key third-party vendors who are part of the Cisco Developer Network and who are developing solutions that leverage Cisco Unified CM and Unified CME SIP capabilities. For example, Cisco worked with Research In Motion to integrate their BlackBerry Mobile Voice System (MVS) solution with Cisco call control platforms in order to enable Cisco Unified Communications and enterprise calling natively on BlackBerry smartphones. Another third-party vendor is Tenacity Operating, which provides a software-based endpoint called accessaphone ipTTY that enables terminal teletype (TTY) or text-based communications for IP telephony. This software-based endpoint can register and communicate with Cisco Unified CM as a third-party SIP phone.

For more information on Cisco's line-side SIP interoperability, refer to the Cisco Unified Communications Manager programming guides at

http://www.cisco.com/en/US/products/sw/voicesw/ps556/products_programming_reference_guides_list.html

For more information on the Cisco Developer Network and third-party development partners, refer to the information available on the Cisco Developer Community at

http://developer.cisco.com

High Availability for Unified Communications Endpoints

To stay in service even during failure of the Unified CM subscriber or other servers, Cisco Unified Communications endpoints are capable of being configured with multiple servers. For example, either through direct configuration or through DHCP during the boot-up phase, the endpoints can accept and process more than one TFTP server address. In case the primary TFTP server is down when the endpoint boots up, the endpoint can get its configuration files from the secondary TFTP server.

Each of the endpoints is also associated with a device pool. The device pool contains a Unified CM Group that has one or more Unified CM subscribers. A list of these subscribers is sent to the endpoints in their configuration files. The endpoints attempt to register with the first (the primary) subscriber in the list. If that Unified CM subscriber is unavailable, the endpoint attempts to register with the second subscriber in the list (the secondary), and so on. Once registered to a subscriber, an endpoint can fail-over to another subscriber in the priority list in the Unified CM Group if the current subscriber fails. When a higher-priority subscriber comes back up, the endpoint will re-register to it.

Note The EX Series video endpoints do not support registration redundancy as described above. If the primary Unified CM node is unreachable, the endpoint will not fail-over its registration to a secondary Unified CM node.

To protect against network failure for endpoints located across a WAN from the Unified CM cluster, a locally available Cisco Integrated Services Router (ISR) equipped with SRST or Unified CME acting as SRST may also be configured in the list of servers with which the endpoint may register. In case of a WAN failure, the endpoints register to the SRST router and provide uninterrupted telephony services (although the set of features they support in SRST mode might be smaller). Note that some endpoints might not support SRST.

Endpoints should be distributed uniformly across servers in the cluster to avoid overloading of any single server. For more information on redundancy methods between cluster subscribers, see the chapter on Call Processing, page 8-1.

Capacity Planning for Unified Communications Endpoints

Cisco call control platforms support the following high-level endpoint capacities:

- A Cisco Unified CM cluster supports a maximum of 40,000 SCCP or SIP endpoints.
- Cisco Business Edition supports a maximum of between 400 and 1,200 SCCP or SIP endpoints, depending on the version.
- Cisco Unified CM Express supports a maximum of 450 SCCP or SIP endpoints.

The above numbers are nominal maximum capacities. The maximum number of endpoints that the call control platform will actually support depends on all of the other functions that the platform is performing, the busy hour call attempts (BHCA) of the users, and so forth, and the actual capacity could be less than the nominal maximum capacity.

In addition to call control platform capacity, network capacity must be considered when it comes to 802.11 wireless attached devices such as the Cisco Unified Wireless IP Phone 7925G or an Android smartphone running Cisco Jabber for Android. See Wireless Call Capacity, page 18-13, for voice and video call capacities per 802.11 channel cell.

For more information on endpoint capacity with Cisco call control, including platform-specific endpoint capacities per node, see the chapter on Unified Communications Design and Deployment Sizing Considerations, page 29-1.

Design Considerations for Unified Communications Endpoints

The following list summarizes high-level design recommendations for deploying Cisco Unified Communications endpoints:

- Analog gateways are available both as standalone devices and as integrated interface modules on Cisco IOS multiservice routers, and both types can be used within the same deployment. Select the analog gateway or gateways that meet analog port density requirements across company locations. Ensure that appropriate port capacity is provided for all locations in order to accommodate the required analog devices.
- Enable the role of St**andard CTI Allow Control of Phones supporting Connected Xfer and conf** for the end-user configuration associated with the device in order to enable CTI monitoring and control of Cisco Unified IP Phone 8900 and 9900 Series endpoints. Only after this role has been enabled can CTI applications monitor or control these phones.

- To minimize endpoint firmware upgrade times over the WAN to remote branches, consider deploying a local TFTP server at the remote location and point endpoints located in that branch to this local TFTP server using the **load server** parameter. Alternatively, consider the use of the Peer File Sharing (PFS) feature when all or most of the devices at a particular remote location are the same phone model.
- Cisco Unified IP desk phones can be powered by power over Ethernet (PoE) when plugged into inline power-capable switches or when deployed with an inline power injector. Consider the use of inline power to reduce downtime and eliminate the need for an external power supply and wall power outlet.
- When deploying Cisco endpoints in branch locations separated from a centralized call processing platform by a low-speed or unreliable WAN link, it is important to consider local call processing redundancy. By using SRST or Unified CME as SRST on a Cisco IOS router in each branch location, basic IP telephony services can be maintained for the desk phones when connectivity to the centralized call processing platform is lost. However, the set of available user-facing features is much smaller when a device is registered to SRST than when the phone is registered to Unified CM.
- For deployments with network voice and data VLAN separation, ensure that inter-VLAN routing has been configured and allowed so that Cisco software-based endpoints that run on desktop computers usually connected to data VLANs can communicate with endpoints on the voice VLAN. This is also important for endpoints on the voice VLAN that may be dependent on data VLAN-based resources that provide services such as directory and management.
- A WLAN site survey must be conducted to ensure appropriate RF design and to identify and eliminate sources of interference prior to deploying wireless and mobile endpoints capable of generating real-time traffic on the wireless network. This is necessary to ensure acceptable voice and video quality for calls traversing the WLAN.
- Select a WLAN authentication and encryption method that not only adheres to company security policies but also enables fast rekeying or authentication so that audio and video calls are not interrupted when wireless endpoints move from one location to another.
- Cisco recommends relying on the 5 GHz WLAN band (802.11a/n) whenever possible for connecting wireless endpoints and mobile client devices capable of generating voice and/or video traffic. 5 GHz WLANs provide better throughput and less interference for voice and video calls. If the 2.4 GHz band is used for connecting wireless client devices and endpoints, Bluetooth should be avoided.
- Provide appropriate network and call control capacity to support the number of endpoints deployed. First, consider the endpoint registration and configuration capacities per call control platform (maximums between 40,000 endpoints per Unified CM cluster and 400 endpoints per Cisco Business Edition). Next, consider call capacities per wireless channel cell for wireless attached endpoints, and the maximum of 27 bidirectional voice-only streams or maximum of 8 simultaneous voice and video streams or calls per WLAN channel cell.
- Cisco TelePresence System EX Series video endpoints do not currently support registration redundancy with Unified CM. If the primary Unified CM node to which the EX series endpoint is registered becomes unreachable, the endpoint will not fail-over its registration to a secondary Unified CM node.

CHAPTER 19

Cisco Unified CM Applications

Revised: April 30, 2013; OL-27282-05

Cisco Unified Communications Manager (Unified CM) applications provide numerous operational and functional enhancements to basic IP telephony. External eXtensible Markup Language (XML) productivity applications or IP Phone Services can be run on the web server and/or client on most Cisco Unified IP Phones. For example, the IP phone on a user's desk can be used to get stock quotes, weather information, flight information, and other types of web-based information. In addition, custom IP phone service applications can be written that allow users to track inventory, bill customers for time, or control conference room environments (lights, video screen, temperature, and so forth). Unified CM also has a number of integrated applications that provide additional functionality, including:

- Cisco Extension Mobility (EM)

  The Extension Mobility feature enables mobile users to configure a Cisco Unified IP Phone as their own, on a temporary basis, by logging in to that phone.

- Cisco Unified Communications Manager Assistant (Unified CM Assistant)

  Unified CM Assistant is a Unified CM integrated application that enables assistants to handle one or more managers' incoming phone calls.

- Cisco WebDialer

  WebDialer is a click-to-call application for Unified CM that enables users to place calls easily from their PCs using any supported phone device.

In some cases these integrated applications also invoke IP Phone Services to provide additional functionality.

This chapter examines the following Unified CM applications:

What's New in This Chapter

Table 19-1 lists the topics that are new in this chapter or that have changed significantly from previous releases of this document.

Table 19-1 New or Changed Information Since the Previous Release of This Document

New or Revised Topic	Described in	Revision Date
Minor corrections and changes	Various sections	April 30, 2013
Cisco Unified Attendant Console	Attendant Consoles, page 19-43	September 28, 2012
Extension Mobility Cross Cluster (EMCC) with phones in secure mode	Support for Phones in Secure Mode, page 19-15	June 28, 2012
Other minor updates for Cisco Unified Communications System Release 9.0	Various sections throughout this chapter	June 28, 2012

IP Phone Services

Cisco Unified IP Phone Services are applications that utilize the web client and/or server and XML capabilities of the Cisco Unified IP Phone. The Cisco Unified IP Phone firmware contains a micro-browser that enables limited web browsing capability. These phone service applications provide the potential for value-added services and productivity enhancement by running directly on the user's desktop phone. For purposes of this chapter, the term *phone service* refers to an application that transmits and receives content to and from the Cisco Unified IP Phone.

This section examines the following design aspects of the IP Phone Services feature:

- IP Phone Services Architecture, page 19-2
- High Availability for IP Phone Services, page 19-6
- Capacity Planning for IP Phone Services, page 19-7
- Design Considerations for IP Phone Services, page 19-8

IP Phone Services Architecture

An IP Phone service can be initiated in several ways:

- User-initiated (pull)

  An IP Phone user presses the Services button, which sends an HTTP GET message to Unified CM for displaying a list of user-subscribed phone services. Figure 19-1 illustrates this functionality.

- Phone-initiated (pull)

  An idle time value can be set within the IP Phone firmware, as indicated by the URL Idle Time parameter. When this timeout value is exceeded, the IP Phone firmware itself initiates an HTTP GET to the idle URL location specified by the URL Idle parameter.

- Phone service-initiated (push)

  A phone service application can push content to the IP Phone by sending an HTTP POST message to the phone.

Note Unlike with the user-initiated and phone-initiated pull functionality, whereby the phone's web client is used to invoke phone services, the phone service-initiated push functionality invokes action on the phone by posting content (via an HTTP POST) to the phone's web server (not to its client).

Figure 19-1 shows a detailed illustration of the user-initiated IP Phone service operation. With Services Provisioning set to **External URL** or **Both** when a user presses the Services button, an HTTP GET message is sent from the IP Phone to the Unified CM getservicesmenu.jsp script by default (step 1). You can specify a different script by changing the Phone URL enterprise parameter. The getservicesmenu.jsp script returns the list of phone service URL locations to which the individual user has subscribed (step 2). The HTTP response returns this list to the IP Phone (step 3). Any further phone service menu options chosen by the user continue the HTTP messaging between the user and the web server containing the selected phone service application (step 4).

By default the Services Provisioning parameter is set to **Internal**. With this setting, the IP phone obtains the list of phone services from its configuration file instead of sending an HTTP GET message to Unified CM.

Note If the Service Provisioning enterprise parameter is set to Internal, steps 1 through 3 are bypassed and the operation of phone services begins with step 4.

Note The Cisco Unified IP Phone 7960 does not have the ability to parse the list of phone services from its configuration file, so it sends an HTTP GET to Unified CM to get that list, even if the Service Provisioning enterprise parameter is set to **Internal**.

Figure 19-1 User-Initiated IP Phone Service Architecture

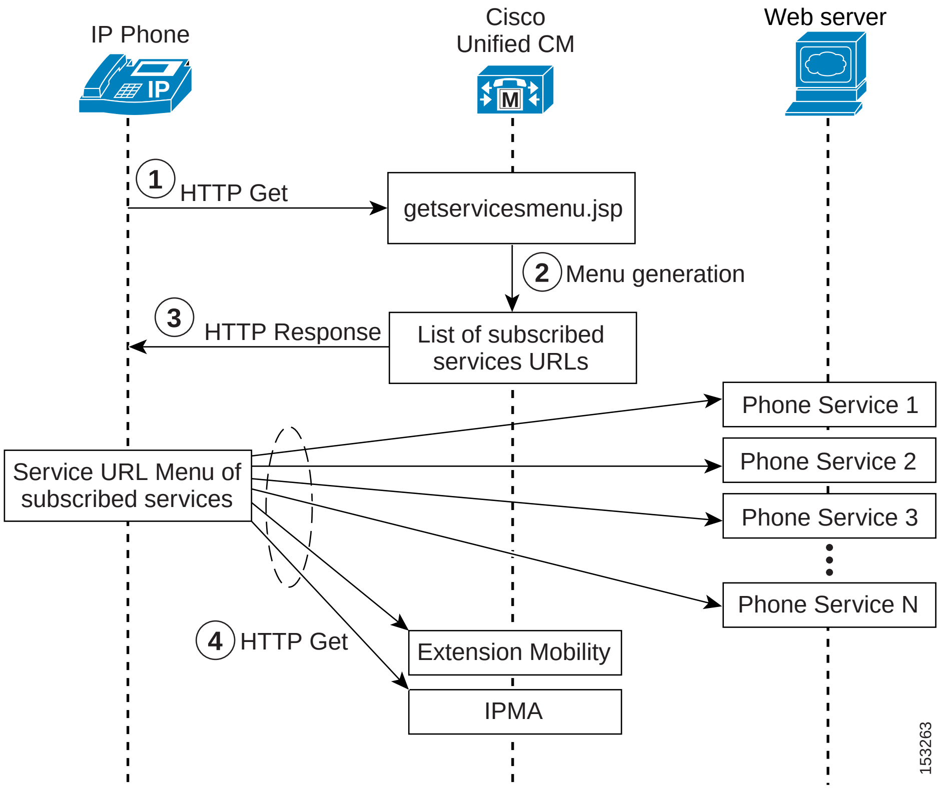
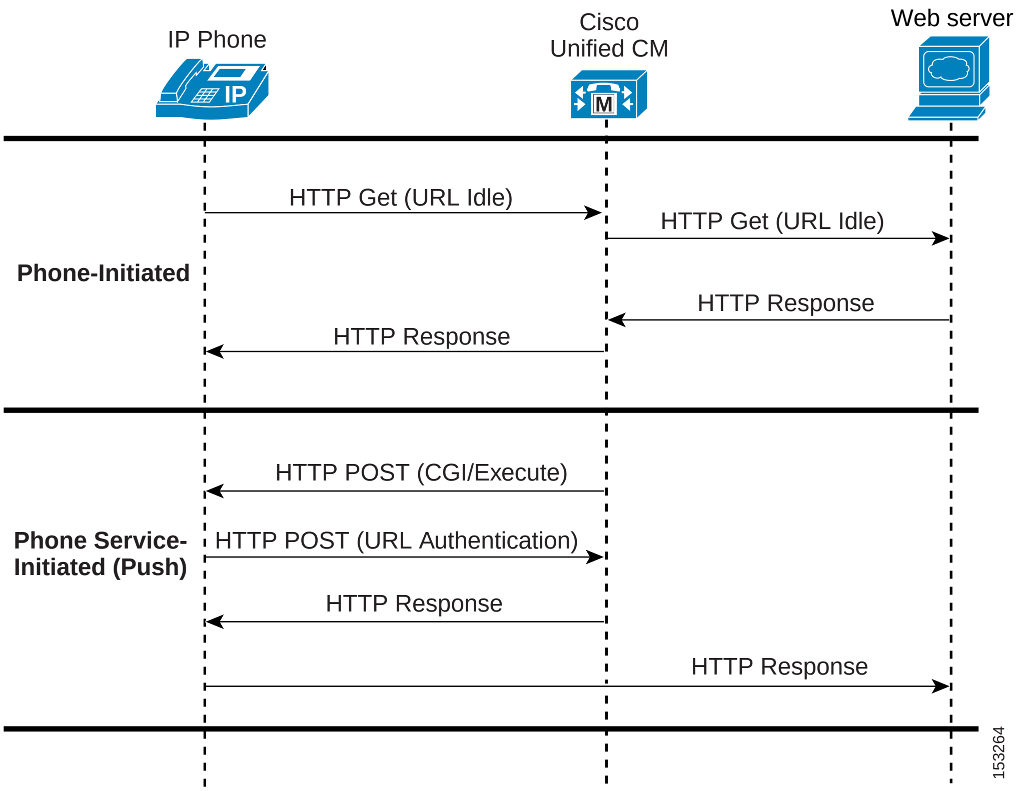


Figure 19-2 shows examples of both phone-initiated and phone service-initiated push functionality. In the phone-initiated example, the phone automatically sends an HTTP GET to the location specified under the URL Idle parameter when the URL Idle Time is reached. The HTTP GET is forwarded via Unified CM to the external web server. The web server sends back an HTTP Response, which is relayed by Unified CM back to the phone, and the phone displays the text and/or image on the screen.

In the phone service-initiated push example, the phone service on the external web server sends an HTTP POST with a Common Gateway Interface (CGI) or Execute call to the phone's web server. Before performing the CGI or Execute call, the phone authenticates the request using the proxy authentication service specified by the URL Authentication parameter. This proxy authentication service provides an interface between the phone and the Unified CM directory in order to validate requests made directly to the phone. If the request is authenticated, Unified CM forwards an HTTP Response to the phone. The phone's web server then performs the requested action, and the phone returns an HTTP response back to the external web server. If authentication fails, Unified CM forwards a negative HTTP Response, and the phone does not perform the requested CGI or Execute action but in turn forwards a negative HTTP Response to the external web server.

Figure 19-2 Phone-Initiated and Phone Service-Initiated IP Phone Service Architecture



In addition to XML Services, a new service can be created with a Service Category of Java MIDlet. When a Java MIDlet-type service is invoked, the configured Service URL contains the URL from which the MIDlet JAD file can be retrieved. When the application server receives the JAD file request, the server should return the appropriate JAR file for that device, which the phone's MIDlet-installer will download and process.

For more information on Java MIDlet support on Cisco IP Phones, refer to the Cisco IP Phone data sheets at http://www.cisco.com.

Note After a phone has downloaded its configuration file via TFTP, the phone parses the services configuration to determine whether or not the list of services has changed, and if so, it updates its local (persisted) services configuration. If any of the changed services were Java MIDlets (which are explicitly provisioned and stored on the phone), then the phone sequentially walks through the necessary install, upgrade, downgrade, and uninstall operations to comply with what was provisioned in the configuration file. If a MIDlet install fails, it will re-attempt the install the next time the phone checks its configuration file (during boot, reset, or restart).

The administrator has the added ability to specify the Service Type of configured services to be one of the following: IP Phone Services, Directories, or Messages. This gives the administrator the flexibility to control which button users must press on the IP phone to access new services. New services can optionally be configured as Enterprise Subscriptions, which forces them to appear automatically on all IP phones without the need to update subscriptions for each individual phone. In addition, services can be enabled or disabled without the need to delete the service from the Unified CM database.

Note Default services such as Missed Calls, Placed Calls, and Corporate Directory can also be disabled. This allows the administrator to create a custom service with a Service URL matching that of the corresponding default service, thus allowing phones to subscribe to these default services on an as-needed basis.

Unified CM provides the ability to configure a secure IP Phone Services URL using HTTPS in addition to a non-secure URL. Phones that support HTTPS will automatically use the secure URL. For more information about Trust Verification Services and security certificate handling for IP phones, along with a complete list of phones that support HTTPS, refer to the HTTPS information in the latest version of the *Cisco Unified Communications Manager Security Guide*, available at

http://cisco.com/en/US/products/sw/voicesw/ps556/prod_maintenance_guides_list.html

High Availability for IP Phone Services

To ensure reliable services for phone users, you must maintain a high level of system availability, with a seamless transition to redundant systems during a system failure.

With Services Provisioning set to Internal, the phone will receive its subscribed phone services from the phone's configuration file and store these (and their corresponding service URLs) in flash. This allows the phone to access the service URLs directly on a web server without first querying the Cisco CallManager IP Phone Service. With Services Provisioning set to Internal, the Corporate and Personal Directories default services also have an extra level of redundancy built into the phones. When these services are selected, the phone will attempt to send an HTTP message with the proper URL string to the Unified CM with which it is currently registered. Therefore, the Unified CM Group configuration of the phone's device pool provides redundancy for these services.

If Services Provisioning is set to External URL or both, while most of the back-end processing of a phone service occurs on a web server, the phones still depend upon Unified CM to inform them of the service URLs for their subscribed phone services. Given the architecture of IP phone service functionality and the message flows shown in Figure 19-1 and Figure 19-2, the following two main failure scenarios should be considered.

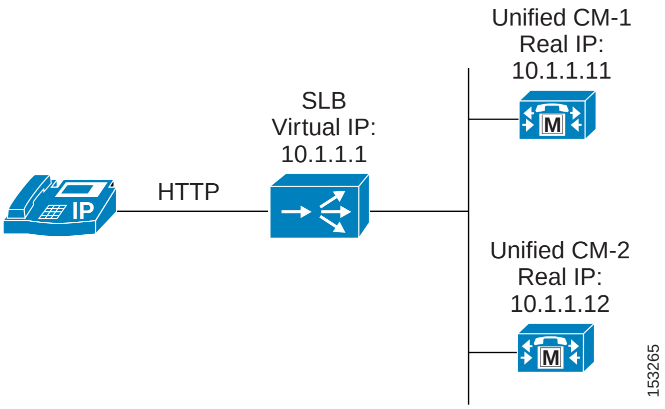
Failure Scenario 1: Server with Cisco CallManager Cisco IP Phone Services Fails

Redundancy in this case depends upon some type of server load balancing (SLB), as illustrated in Figure 19-3, where a virtual IP address (or DNS-resolvable hostname) is used to point to one or more Unified CM servers. This virtual IP address (or DNS-resolvable hostname) is used when configuring the URL Services parameter. The SLB device is configured with the real IP addresses of the Unified CM subscriber nodes. Thus, a Unified CM server failure does not prevent the IP Phone Services subscription list from being returned to the phone when the phone's Services button is pushed. In addition, phone services such as Extension Mobility and Unified CM Assistant that run on a Unified CM server are also potentially made redundant by this method. (See High Availability for Extension Mobility, page 19-16, and High Availability for Unified CM Assistant, page 19-24.)

Most SLB devices, such as the Cisco Application Control Engine (ACE), can be configured to monitor the status of multiple servers and automatically redirect requests during failure events. For more information on the Cisco Application Control Engine (ACE), refer to the documentation available at

http://www.cisco.com/en/US/products/ps5719/Products_Sub_Category_Home.html

Figure 19-3 Method for Providing Redundancy for Phone Services



Failure Scenario 2: External Web Server Hosting a Particular IP Phone Service Fails

In this scenario, the connection to the Unified CM server is preserved, but the link fails to the web server hosting the user-subscribed phone service. This is an easier scenario to provision for redundancy because the IP phone is still able to access the Unified CM server when the Services button is pressed. In this case, the IP phone is similar to any other HTTP client accessing a web server. As a result, you can again use some type of SLB functionality (similar to the one indicated in Figure 19-3) to redirect the HTTP request from the phone to one or more redundant web servers hosting the user-subscribed phone service.

Capacity Planning for IP Phone Services

Cisco Unified IP Phone Services act, for the most part, as an HTTP client. In most cases it uses Unified CM only as a redirect server to the location of the subscribed service. Because Unified CM acts as a redirect server to the phone service, there typically is minimal performance impact on Unified CM when a user initiates a phone service request by pressing the Services key, but a large number of requests (hundreds of requests per minute or more) could affect the server performance. To minimize the impact on the server performance, if an external URL does not need to be specified for the IP Phone Services, Cisco generally recommends leaving the Services Provisioning Enterprise Parameter set to **Internal**. If Services Provisioning has to be set to **External URL** or **Both**, or if you are using a large number phones that do not have the ability to retrieve the list of services from their configuration file (such as the Cisco Unified IP Phone 7960), carefully select the node that will provide the Cisco Unified IP Phone Services list. For example, consider using the Unified CM TFTP servers instead of the Unified CM publisher if the load on the publisher is already high, or consider using Unified CM subscribers that are not handling a lot of traffic.

Note In the case of Extension Mobility and Unified CM Assistant phone service, Unified CM acts as more than a redirect server, and additional performance impacts should be considered. See the sections on Extension Mobility, page 19-8, and Unified CM Assistant, page 19-20, for specific performance and scalability considerations for these applications.

Because the IP Phone is either an HTTP client or server, estimating the required bandwidth used by an IP Phone service is similar to estimating the bandwidth of an HTTP browser accessing the same text as HTTP content residing on a web hosting server.

Design Considerations for IP Phone Services

With the exception of the integrated Extension Mobility and Unified CM Assistant applications’ Phone Services, IP Phone services must reside on a separate off-cluster non-Unified CM web server. Running phone services other than Extension Mobility and Unified CM Assistant on the Unified CM server node is not supported.

Most Cisco IP phones support content with text and graphics. Some phones such as the Cisco Unified IP Phone 7911G support only text-based XML applications.

Extension Mobility

The Cisco Extension Mobility (EM) feature enables users to configure a Cisco Unified IP Phone as their own, on a temporary basis, by logging in to that phone. After a user logs in, the phone adopts the user's individual device profile information, including line numbers, speed dials, services links, and other user-specific properties of a phone. For example, when user X occupies a desk and logs in to the phone, that user's directory number(s), speed dials, and other properties appear on that phone; but when user Y uses the same desk at a different time, user Y's information appears. The EM feature dynamically configures a phone according to the authenticated user's device profile. The benefit of this application is that it allows users to be reached at their own extension on any phone within the Unified CM cluster, regardless of physical location, provided the phone supports EM.

This section examines the following design aspects of the Extension Mobility feature:

- Unified CM Services for Extension Mobility, page 19-8
- Extension Mobility Architecture, page 19-9
- Extension Mobility Security, page 19-14
- Extension Mobility Cross Cluster (EMCC), page 19-10
- High Availability for Extension Mobility, page 19-16
- Capacity Planning for Extension Mobility, page 19-18
- Design Considerations for Extension Mobility, page 19-19

Unified CM Services for Extension Mobility

The EM application relies on the Cisco Extension Mobility service, which is a feature service and which you must activate manually from the Serviceability page.

EM also relies on the Cisco Extension Mobility Application network service, which is activated automatically on all Unified CM nodes during installation.

The Cisco Extension Mobility Application service is a network service that provides an interface between the EM user phone and the Cisco Extension Mobility service. In addition, the Cisco Extension Mobility Application service subscribes to the change notification indications within the cluster and maintains a list of nodes in the cluster that have an active Cisco Extension Mobility service.

Extension Mobility Architecture

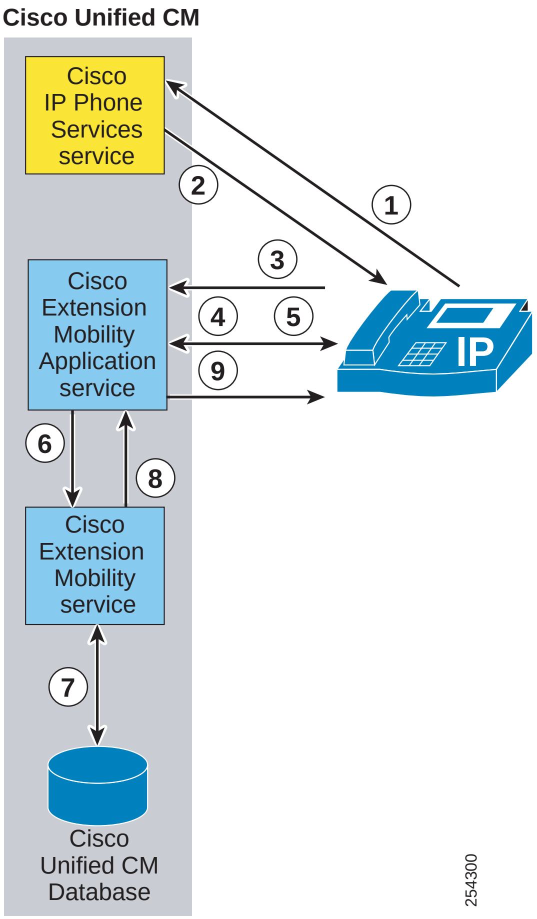
Figure 19-4 depicts the message flows and architecture of the EM application. When a phone user wants to access the EM application, the following sequence of events occurs:

1. When the user presses the Services button on the phone, this action generates a call to the URL specified under the URL Services parameter on the Enterprise Parameter configuration page (see step 1 in Figure 19-4).
2. An HTTP/XML call is generated to the IP Phone Services, which returns a list of all services to which the user's phone is subscribed (see step 2 in Figure 19-4).

Note With the Services Provisioning enterprise parameter set to Internal, steps 1 and 2 are bypassed. Alternatively, with Services Provisioning set to External URL or Both, a Service URL button can be configured for EM on a user's phone so that the user can press a line or speed-dial button to generate a direct call to the Cisco Extension Mobility Application service, also bypassing steps 1 and 2.

3. Next the user selects the Extension Mobility phone service listing. This selection in turn generates an HTTP call to the Cisco Extension Mobility Application service, which serves as the interface between the phone and the Cisco Extension Mobility service (see step 3 in Figure 19-4).
4. The Cisco Extension Mobility Application service then forwards an XML response back to the phone requesting user login credentials (userID and PIN) or, if the user is already logged in, a response asking if the user wants to log off the phone (see step 4 in Figure 19-4).
5. Assuming the user is attempting to log in, the user must use the phone's keypad to enter a valid userID and PIN. After the user presses the Submit softkey, a response containing the userID and PIN just entered is forwarded back to the Cisco Extension Mobility Application service (see step 5 in Figure 19-4).
6. The Cisco Extension Mobility Application service next forwards this login information to the Cisco Extension Mobility service, which interacts with the Unified CM database to verify the user’s credentials (see step 6 in Figure 19-4). The Cisco Extension Mobility Application service subscribes to cluster change notification, and it maintains a list of all nodes in the cluster with the Cisco Extension Mobility service activated. Therefore, in case the Cisco Extension Mobility service is not running on the same Unified CM node, the Cisco Extension Mobility Application service forwards the login information to other Unified CM nodes that are running the Cisco Extension Mobility service.
7. Upon successful verification of the user’s credentials, the Cisco Extension Mobility service also interacts with the Unified CM database to read and select the appropriate user device profile and to write needed changes to the phone configuration based on this device profile (see step 7 in Figure 19-4).
8. Once these changes have been made, the Cisco Extension Mobility service sends back a successful response to the Cisco Extension Mobility Application service (see step 8 in Figure 19-4).
9. The Cisco Extension Mobility Application service, in turn, sends a reset message to the phone, and the phone resets and accepts the new phone configuration (see step 9 in Figure 19-4).

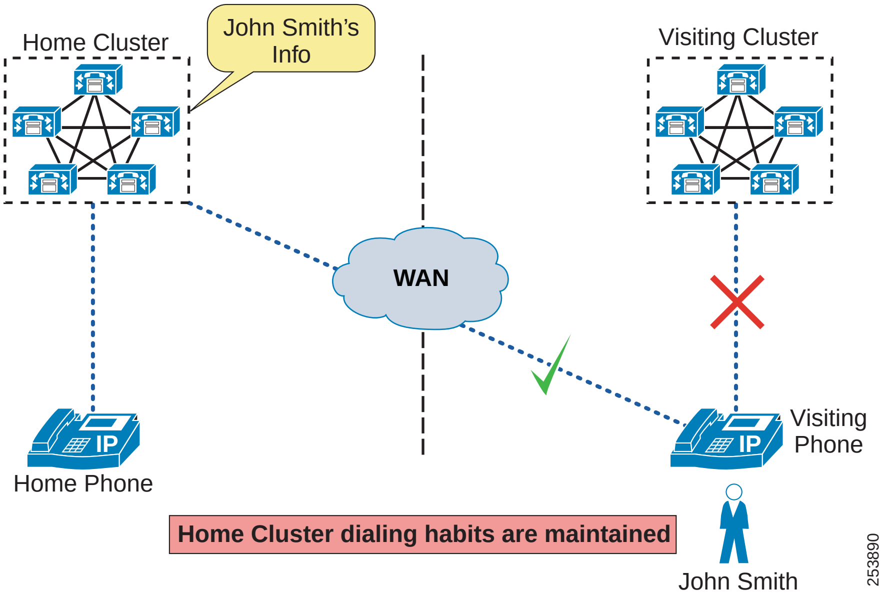
Figure 19-4 EM Application Architecture and Message Flow


Cisco Unified CM
Cisco IP Phone Services service
Cisco Extension Mobility Application service
Cisco Extension Mobility service
Cisco Unified CM Database
1
2
3
4
5
6
7
8
9
IP
254300

Extension Mobility Cross Cluster (EMCC)

Unified CM provides the ability to perform Extension Mobility logins between clusters within an enterprise with a new feature called Extension Mobility Cross Cluster (EMCC). It is important to understand the high-level architecture of EMCC. The EMCC feature employs the concepts of a home cluster and a visiting cluster, and these terms are defined from the perspective of the user performing the login. When a user travels to an office and attempts to log in to a phone, if the cluster to which this phone is registered does not contain the user's information in its database, then this cluster is considered a visiting cluster and the phone is hereinafter referred to as the visiting phone. Figure 19-5 illustrates the concept of home and visiting clusters.

Figure 19-5 EMCC Home Cluster and Visiting Cluster



The EM service in the visiting cluster attempts to locate the home cluster of the user by sending out queries to each of the EMCC remote clusters that have been configured in Unified CM. When the user's home cluster responds positively, this initiates communications between the EM services of both clusters to exchange information that essentially brings the device information into the home cluster database and allows the home cluster to build a configuration file for this visiting phone. This configuration file incorporates some device configuration from the visiting cluster, configuration parameters from the home cluster, and the user's device profile in the home cluster. Once the home cluster TFTP server has a configuration file for this visiting phone, a reset issued by the visiting cluster forces the visiting phone to download a small configuration from the visiting cluster, which further instructs it to download certificates and a full configuration from the home cluster. Ultimately, the visiting phone cross-registers with the home cluster. This means that all call control signaling occurs between a home cluster Unified CM subscriber and the visiting phone, and the user's home cluster dialing habits are maintained.

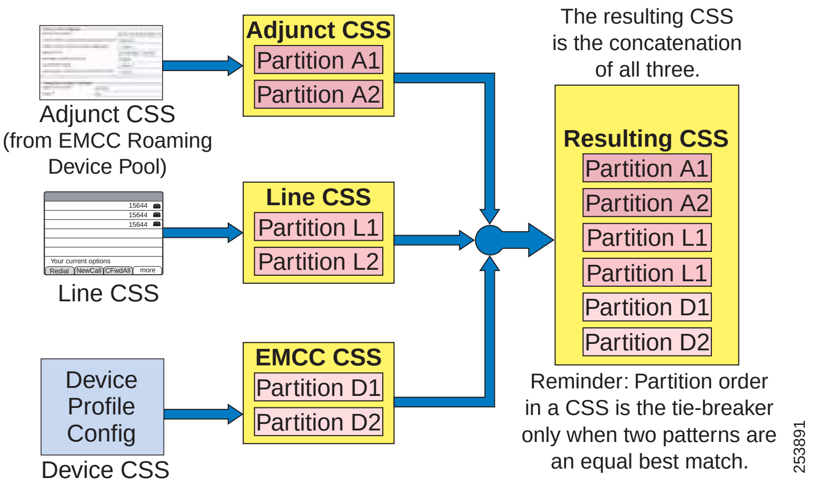
For a step-by-step description of the EMCC login process, refer to the Extension Mobility Cross Cluster information in the latest version of the *Cisco Unified Communications Manager Features and Services Guide*, available at

http://cisco.com/en/US/products/sw/voicesw/ps556/prod_maintenance_guides_list.html

Call Processing

EMCC call processing behavior is also critical to understand because it impacts dial plan design. When a user has logged into a phone in a visiting cluster, any digits dialed by the user are analyzed by the home cluster according to the visiting phone's assembled call search space (CSS), which is a concatenation of the Adjunct CSS in the home cluster's device pool for the visiting phone (referred to as the EMCC roaming device pool), the Line CSS configured on the directory number associated with the user's device profile, and the EMCC CSS configured on the user's device profile. Figure 19-6 illustrates the resulting CSS for an EMCC phone.

Figure 19-6 Resulting CSS for an EMCC Phone



The Adjunct Calling Search Space is a new call routing configuration parameter that is used by EMCC to intercept and route emergency numbers for users from a visiting cluster. The Adjunct CSS contains a partition with directory numbers such as 911, 112, or 999, that route the calls to the visiting cluster and allow the call to reach emergency services local to the physical phone's location. For more information on Adjunct Calling Search Spaces and the EMCC roaming device pool and how it is associated with a visiting phone, refer to the Extension Mobility Cross Cluster information in the latest version of the *Cisco Unified Communications Manager Features and Services Guide*, available at

http://cisco.com/en/US/products/sw/voicesw/ps556/prod_maintenance_guides_list.html

Note The EMCC roaming device pool associated with the EMCC feature is not related to the roaming device pool associated with the Device Mobility feature.

EMCC users must be aware that, when placing calls, they will be leveraging their home Unified CM routes and numbering plan. For example, if a user from Cluster A logs into a phone from Cluster B and wants to place a call to the directory number of a Cluster B phone located right next to it, the user would have to dial the appropriate pattern as if the user was placing the call from Cluster A to the phone in Cluster B. This implies that the home cluster may initiate an intercluster trunk call from Cluster A to Cluster B, but the media will flow locally between the visiting phone and the remote phone.

If the EMCC clusters have been deployed using +E.164 numbering, then the users should already be accustomed to dialing the full number of the target number and will not need to alter their dialing habits.

With PSTN routed calls, there are two different configurations that affect call processing behavior:

- Route patterns that do not use the Local Route Group (LRG) feature
- Route patterns that use the LRG feature

When an EMCC logged-in user dials a PSTN call, if the digit analysis matches a route pattern that ultimately leads to a voice gateway (either via the route list and route group construct or configured directly to a voice gateway), the call is offered out the gateway. If the Standard Local Route Group (Standard LRG) feature is not in use, and the call involves a voice gateway associated with the home cluster; therefore media will flow between the visiting phone (typically across a WAN) back to the voice gateway. When the route pattern leads to a route list configured to use Standard LRG, the behavior changes. (For more information about LRG, see Local Route Group, page 9-103.) When Unified CM logic must invoke a Standard LRG for an EMCC logged-in device, it recognizes the endpoint as an EMCC device and sends the PSTN call across a designated EMCC-specific SIP trunk to the visiting cluster to which this visiting phone is normally registered.

Note Only one SIP trunk with an EMCC trunk service type is required per cluster. There is no destination information configured on this trunk; that information is gathered dynamically when adding and updating an EMCC remote cluster.

When a call invite is received on the EMCC SIP trunk in the visiting cluster, the visiting cluster again performs digit analysis on the called number according to the CSS of the trunk (or alternatively, according to the CSS of the visiting phone's original device configuration), and routes the call accordingly. There is additional information included in a SIP invite across an EMCC SIP trunk, namely the device name of the visiting phone. This enables the visiting cluster to determine the configured device CSS of the visiting phone in the database (if required); and if the digit analysis results in matching a route pattern that ultimately points to the Standard LRG, the visiting cluster is able to determine the configured Standard LRG for this visiting phone. The Standard LRG in the visiting cluster will typically contain voice gateways associated with the visiting cluster, therefore the PSTN call is offered out a voice gateway local to the visiting phone.

The difference between LRG and non-LRG call processing behavior is critical when considering calls to emergency numbers. While the use of Local Route Groups (LRGs) is not required cluster-wide for an EMCC deployment, the EMCC logged-in phones must have access to an LRG in order to route emergency calls correctly. An LRG is required to correctly route an emergency call to a visiting cluster so that the call can be placed through an appropriate voice gateway local to the visiting phone. The Adjunct Calling Search Space in the roaming device pool configuration for an EMCC device enables an administrator to add emergency route patterns that will use an LRG for EMCC logged-in devices, but it will not affect emergency dialing for other devices in the home cluster. As discussed earlier, an EMCC logged-in phone will be associated with a device pool (by means of geolocations) that represents all phone devices from another cluster. The device pool’s Adjunct Calling Search Space allows for the visiting cluster's emergency route pattern to be configured so that only emergency calls for an EMCC logged-in phone will be sent through an LRG. So even if the home and visiting clusters use the same emergency route pattern, the EMCC logged-in phone's emergency call will route through the LRG to the visiting cluster. Once the call is received at the visiting cluster through the EMCC SIP trunk, the visiting cluster dial plan will be responsible for further processing of the call.

Note If any cluster supporting EMCC is also using Cisco Emergency Responder for emergency call processing, refer to the *Cisco Emergency Responder Administration Guide* for information on how to configure the dial plan to support the deployment, available at http://www.cisco.com/en/US/products/sw/voicesw/ps842/prod_maintenance_guides_list.html.

Note If Standard LRGs are already deployed for the emergency route pattern, and if the home and visiting clusters use the same emergency dial string, use of the Adjunct CSS is not required.

For detailed EMCC call processing examples and configuration, refer to the Extension Mobility Cross Cluster information in the latest version of the *Cisco Unified Communications Manager Features and Services Guide*, available at

http://cisco.com/en/US/products/sw/voicesw/ps556/prod_maintenance_guides_list.html

Media Resources

All media resources except for RSVP agents are allocated from the home cluster according to the media resource group list of the device pool assigned to the visiting phone. Conferencing, transcoding, and music on hold all function as normal, with the difference being that media is streaming between the visiting phone and media resources across (typically) a WAN separating the home and visiting clusters. When an EMCC logged-in user makes a call that requires use of an RSVP agent, the Unified CM EMCC logic is able to determine it is a visiting phone, and it sends a resource request across the EMCC SIP trunk to the remote cluster to which the visiting phone belongs. The device name of the visiting phone is included in this request, which enables the visiting cluster to verify the RSVP agent media resources that are normally assigned to this visiting phone and to allocate its use for the call. For more information on RSVP-based call admission control for EMCC, see Architecture and Considerations for Extension Mobility Cross Cluster, page 11-83.

Extension Mobility Security

Unified CM provides the ability to create an Extension Mobility secure service URL using HTTPS. This encrypts the entire EM login/logout exchange. Cisco recommends configuring a secure service URL for Extension Mobility. If there are phones deployed for EM that do not support HTTPS, a non-secure service URL must also be configured. When secure and non-secure service URLs exist for the service, phones that support HTTPS use the secure service URL by default. For a complete list of phones that support HTTPS, refer to the HTTPS information in the latest version of the *Cisco Unified Communications Manager Security Guide*, available at

http://cisco.com/en/US/products/sw/voicesw/ps556/prod_maintenance_guides_list.html

The EM feature provides an optional level of security for EM login and logout requests by validating the source IP address of the request. By default, EM does not perform this request validation; therefore, to enable EM security, the administrator must set the cluster-wide service parameter Validate IP Address to true.

For organizations that implement a web proxy to handle EM login and logout HTTP requests, the Allow Proxy service parameter must be set to true. A proxy server, while forwarding the HTTP request, will set the via-field of the HTTP header with its hostname. If there are multiple proxy servers between the device and Unified CM, and if the request is forwarded by all the servers, then the via-field in the HTTP header will have a comma-separated list of hostnames for each of the proxy servers in the forwarding path. The Allow Proxy service parameter, if set to true, will allow EM login and logouts received via a web proxy. In addition, if the proxied EM requests use the source IP address of the proxy server, this IP address must also be configured in the Trusted List of IPs service parameter.

With support for HTTPS and Security By Default starting in Unified CM 8.*x*, and with the introduction of secure phones support for EMCC in Unified CM 9.*x*, the intercluster interactions of EMCC require some extra steps to ensure that clusters can communicate with each other in a secure manner. In particular, all clusters that participate in EMCC must export their Tomcat (web) and TFTP certificates to a central sFTP server. Exporting the CAPF certificates is also required if phones used for EMCC will be in secure mode. These security certificates are all combined, and then each cluster must import the combined certificate into its cluster. It is important to remember that any time a new node that may participate in EMCC is added to the cluster, or if a certificate on any existing node is updated, the process of exporting, combining, and importing must be repeated. All of these steps have been streamlined via Unified CM Serviceability administration. For details on EMCC configuration, refer to the Extension Mobility Cross Cluster information in the latest version of the *Cisco Unified Communications Manager Features and Services Guide*, available at

http://cisco.com/en/US/products/sw/voicesw/ps556/prod_maintenance_guides_list.html

Support for Phones in Secure Mode

Starting with Cisco Unified CM 9.*x*, users can log in through EMCC using phones in secure mode — that is, phones with an authenticated or encrypted Device Security Profile. When a user logs in on a phone in secure mode, the configuration in the device security profile (such as the device security mode, TFTP encrypted option, and transport protocol) is transferred to the home cluster, allowing the phone to operate in the same secure mode as it was originally in the visiting cluster. For example, if the phone is configured with the encrypted device security mode in the visiting cluster and the user logs in through EMCC, the phone still operates in the encrypted device security mode with a secure TLS channel for signaling and sRTP for media. However, one condition is that the home cluster security mode must be configured as mixed mode. If the home cluster is configured as non-secure instead, the EMCC login will fail. If the phone is not in secure mode, the phone continues to operate in a non-secure mode after the EMCC login, regardless of whether the visiting cluster is in mixed mode or non-secure mode. Table 19-2 indicates this behavior.

Unified CM 8.*x* supports EMCC but not with phones in secure mode. For this reason, EMCC login attempts from a phone in secure mode registered to a visiting cluster running Unified CM 8.*x* will fail, regardless of whether the home cluster is running Unified CM 8.*x* or 9.*x*. Similarly, EMCC login attempts from a phone in secure mode to a home cluster running Unified CM 8.*x* will fail, regardless of whether the visiting cluster is running Unified CM 8.*x* or 9.*x*. Table 19-2 indicates this behavior.

Table 19-2 Phone Security Mode After EMCC Login

	Home Cluster Running Unified CM 8.*x*	**Home Cluster Running Unified CM 9.*x***	
Visiting Cluster	**Mixed Mode or Non-Secure Mode**	**Mixed Mode**	**Non-Secure Mode**
Phone in secure mode; visiting cluster running Unified CM 8.*x*	EMCC login fails	EMCC login fails	EMCC login fails
Phone in secure mode; visiting cluster running Unified CM 9.*x*	EMCC login fails	Secure mode	EMCC login fails
Phone in non-secure mode; visiting cluster running Unified CM 8.*x* or 9.*x* (Visiting cluster in mixed mode or non-secure mode)	Non-secure mode	Non-secure mode	Non-secure mode

Note As of Cisco Unified CM 9.0, the EMCC SIP trunk cannot be configured with a secure profile. Therefore, calls to the local PSTN do not use a secure channel for signaling. However, the media is encrypted if the phone and PSTN gateway are configured in a secure mode.

High Availability for Extension Mobility

According to the EM architecture illustrated in Figure 19-4, reads and writes to the Unified CM database are required. EM is a user-facing feature, and database writes pertaining to EM can be performed by subscriber nodes. Therefore, if the Unified CM publisher is unavailable, EM logins and logouts are still possible.

From a redundancy perspective, the following component levels of redundancy must be considered for full EM resiliency:

- Cisco CallManager Cisco IP Phone Services

  High availability for the CallManager Cisco IP Phone Services is obtained by using the Services Provisioning service parameter or by using an SLB device pointing to multiple Unified CM nodes running the Cisco CallManager Cisco IP Phone Services. For more details, see High Availability for IP Phone Services, page 19-6.

- Cisco Extension Mobility service

  High availability for the Cisco Extension Mobility service is obtained by activating the Cisco Extension Mobility service on multiple Unified CM nodes.

  Note While the Cisco Extension Mobility service can be activated on more than two nodes, a maximum of two nodes can actively handle login/logout requests at any given time. The other nodes running the Cisco Extension Mobility service should start handling login/logout requests only in case of failure.

Cisco recommends deploying a server load balancer device such as the Cisco Application Control Engine (ACE) to load-balance the requests across two Unified CM nodes and to provide redundancy. Without a server load balancer, load balancing would be uneven and the redundancy would be manual. For example, two EM IP Phone services could be configured on each phone. If one Unified CM node is not reachable, the end user would have to manually select the other EM IP Phone service to reach the other node.

Note While it is possible to provide redundancy for the EM IP Phone service by relying on end users to manually select an EM IP Phone service from a list of EM IP Phone services, achieving high availability in this manner can be problematic. Because there is no control over which EM IP Phone service a user might select from the phone services menu (or assigned feature keys), there is no way to ensure that the EM login/logout load is balanced between Unified CM nodes handling EM login/logout requests. Further, end user behavior when encountering delay in response of the EM service, which is typical in a failure scenario, will usually exacerbate the situation as users cancel EM service calls and select alternate EM IP Phone service. This can lead to added congestion and load on the network as well as on the remaining Unified CM node handling EM login/logout requests.

A deployment with two Unified CM nodes running the Cisco Extension Mobility service provides the highest capacity in terms of number of login/logout requests per minute. (See Capacity Planning for Extension Mobility, page 19-18, for details.) It also provides redundancy. However, in case of failure, the login/logout request capacity is reduced because there is only one node left. Therefore, to achieve the highest login/logout capacity and maintain this capacity in case of failure, the Cisco Extension Mobility service should be activated on additional Unified CM nodes. To load balance evenly across the active nodes and to ensure that only two nodes are handling login/logout requests at any given time, a server load balancer device such as the Cisco Application Control Engine (ACE) should be deployed.

The Cisco Application Control Engine has the capability to detect if a primary server is down and to start sending requests to backup servers in case of failure. For details on the Cisco Application Control Engine (ACE) configuration, refer to the documentation available at

http://www.cisco.com/en/US/products/ps5719/Products_Sub_Category_Home.html

Note Cisco does not recommend a redundancy design using DNS A or SRV records with multiple IP listings. With multiple IP addresses returned to a DNS request, the phones must wait for a timeout period before trying the next IP address in the list, and in most cases this results in unacceptable delays to the end user. In addition, this can result in more than two subscriber nodes with the Cisco Extension Mobility Application service enabled to handle login/logout requests, which is not supported.

With EMCC, remote clusters are administratively added via Unified CM web administration by specifying a single FQDN or IP address of a Unified CM subscriber node running the EM service in the remote cluster. The EM services between the two clusters provide information about the Unified CM version, an ordered list of EM Service nodes for EMCC EM Service communications, which EMCC SIP trunk services are enabled (PSTN Access and/or RSVP Agent) in the remote cluster, and an ordered list of up to three remote Unified CM nodes that handle EMCC SIP trunk operations for each EMCC service. EMCC EM service communications over HTTPS include locating users' home clusters, exchanging information during EMCC logins, and remote cluster updates. Upon an initial update, a remote cluster's Extension Mobility Application service is queried, which will return the first three EM Service nodes in its list. This ordered list determines which remote cluster EM Service nodes will be used for EMCC communications.

The remote cluster obtains the information regarding primary, secondary, and tertiary options for EMCC PSTN Access and RSVP Agent services from the Unified CM Group that is associated with the device pool of the assigned EMCC SIP trunk for those services. This ensures that, if the primary Unified CM subscriber handling the EMCC SIP trunk is offline, then the EMCC SIP trunk call will be handled by the secondary Unified CM subscriber, and so on.

Once a phone is logged in through EMCC, redundancy is provided for the phone in the form of the Unified CM Group configured in its assigned EMCC device pool. If the visiting phone is located in a remote site and there is a WAN outage in which both the visiting and home cluster are unreachable, then the SRST reference from the visiting cluster is maintained by the EMCC phone. Therefore, an EMCC logged-in phone will still be able to register with the appropriate SRST router in the site where it is located. The EMCC logged-in user's DID most likely will not be associated with the local gateway(s) at the SRST site, so incoming calls will still be routed based on the call forwarding rules on the user's home cluster. While in SRST mode, the user will also have to adapt to the visiting SRST site's configured dial habits during SRST failover registration. For additional examples of an EMCC logged-in phone's behavior during a networking failure, refer to the Cisco Extension Mobility Cross Cluster section in the *Cisco Unified Communications Manager Features and Services Guide*, available at

http://cisco.com/en/US/products/sw/voicesw/ps556/prod_maintenance_guides_list.html

Cisco also recommends configuring a default and backup Unified CM TFTP server to be used for visiting phones to download EMCC configuration files that will allow them to register with the home cluster. This is configured under EMCC Feature Configuration.

Capacity Planning for Extension Mobility

With a single Unified CM running the Cisco Extension Mobility application, the maximum cluster-wide capacity is 250 logins and/or logouts per minute with an MCS 7845-H2/I2 or MCS 7845-I3 server, or with a virtual machine using an equivalent OVA. Cisco Extension Mobility login and logout functionality can be distributed across a pair of subscriber nodes to increase login/logout cluster capacity. An SLB device can be used, or to manually distribute the EM load evenly between the two subscriber nodes, the phones should be divided into two groups, with one group of phones subscribed to an EM phone service pointing to one of the subscriber nodes and the other group of phones subscribed to a second EM phone service that is pointing to a second subscriber node. When the EM load is distributed in this way, evenly between two MCS 7845-H2/I2/I3 servers or two virtual machines using an equivalent OVA, the maximum cluster-wide capacity is 375 sequential logins and/or logouts per minute.

Note The Cisco Extension Mobility service can be activated on more than two nodes for redundancy purposes, but Cisco supports a maximum of two subscriber nodes actively handling logins/logouts at any given time.

Note Enabling EM Security does not diminish performance.

The EMCC login/logout process requires more processing resources than intracluster EM login/logout, therefore the maximum supported login/logout rates are lower. In the absence of any intracluster EM logins/logouts, Unified CM supports a maximum rate of 75 EMCC logins/logouts per minute with Cisco MCS 7845-H2/I2 and MCS 7845-I3 servers or the OVA equivalent. Most deployments will have a combination of intracluster and intercluster logins/logouts occurring. For this more common scenario, the mix of EMCC logins/logouts (whether acting as home cluster or visiting cluster) should be modeled for 40 per minute while the intracluster EM logins should modeled for 185 logins/logouts when using a single EM login server. The intracluster EM login rate can be increased to 280 login/logouts per minute when using MCS 7845-H2/I2 or MCS 7845-I3 servers or the OVA equivalent in dual EM service configuration.

For more details on the capacity limits, see the chapter on Unified Communications Design and Deployment Sizing Considerations, page 29-1.

EMCC logged-in devices (visiting phones) consume twice as many resources as any other endpoint in a cluster. The maximum supported number of EMCC logged-in devices is 2,500 per cluster, but this also decreases the theoretical maximum number of other devices per cluster from 30,000 to 25,000. Even if the number of other registered devices in the cluster is reduced, the maximum supported number of EMCC logged-in devices is still 2,500.

There is no technical limit to the number of EMCC remote clusters that can be added to a cluster; however, the full-mesh requirement will increase the load on the EM service as the number of remote clusters increases. For a high number of sites (more than 10), the EM CPU should be monitored by means of the Cisco Real-Time Monitoring Tool (RTMT).

Design Considerations for Extension Mobility

The following guidelines and restrictions apply with regard to the deployment and operation of EM within the Unified CM telephony environment:

- EM users should not move between locations or sites within a cluster when Automated Alternate Routing (AAR) and/or the Voice over PSTN (VoPSTN) deployment model are in use.

  EM functionality relies on the use of the IP network for routing calls. Call routing via the PSTN is more problematic because E.164 PSTN numbers are static and the PSTN is unable to account for movement of EM user directory numbers (DNs) from their home sites. AAR relies on the PSTN for call routing, as does the VoPSTN deployment model. In both cases, EM user movement between locations and sites is supported only if all sites the user is traversing are in the same AAR group. For additional information, see Extension Mobility, page 9-124.

- Restarting the Cisco Extension Mobility service or the node on which the service is running will affect auto-logout settings.

  If the Cisco Extension Mobility service is stopped or restarted, the system does not auto-logout users who are already logged in after the expiration of the maximum login interval. These phones will either have to be logged out manually or wait until the daily database clean-up process runs (typically at midnight).

WebDialer supports the use of phones logged in using Extension Mobility. For more information, please see WebDialer, page 19-34.

Design Considerations for Extension Mobility Cross Cluster (EMCC)

The following design considerations apply when deploying EMCC.

General Design Considerations

- EMCC requires that all users must be unique across all clusters in the enterprise. If LDAP synchronization is maintaining common users for multiple clusters, some type of filtering must be applied.
- Consider the network delay between clusters in combination with the features you plan to use. As the visiting phone is registered with the home cluster, features will work. However, depending on the network delay for a given deployment, all applications and features might not meet user requirements. Testing might be required to determine the usability of features for a given network.

  For example, EMCC supports dynamic CTI control of a visiting phone. But if an offhook is issued via an application and it takes 1 second before the phone goes offhook, this might be acceptable for an office worker but might not be acceptable for a call center agent.

- Phone load firmware is not enforced during the login process. Instead, the visiting cluster phone load information is maintained so that cross-registration does not result in new phone firmware downloads.
- If the home cluster locale is different than that of the visiting cluster, the phone will download the new locale from the visiting cluster TFTP server. If it is not available, then the phone will not change locales and instead will maintain the visiting cluster locale.
- DLUs are not consumed in the home cluster for the registered visiting phones.
- The total number of EMCC logins is controlled by the total number of EMCC inserted devices in the Bulk Administration Tool (BAT).

- EMCC supports only RSVP-based call admission control. Unified CM locations-based call admission control is not supported.
- Except for RSVP agents, all other media resources are allocated from the home cluster according to the media resource group list associated with the EMCC roaming device pool.
- Audio and video codecs are determined by the EMCC region settings. These settings override normal region configuration for EMCC registered phones. All EMCC region parameters must be configured with the same values in all clusters. If they are different, RSVP Agent for that cluster will be disabled by the remote cluster update operation.
- For the EMCC roaming device pool to be assigned correctly, EMCC-capable phones must have a geo-location configured via device configuration or a device pool.

Call Processing Design Considerations

- Incoming calls for a user's directory number will always be received on a home cluster voice gateway, therefore RTP media will flow between the visiting phone and the home gateway for incoming calls.
- Calls sent across the EMCC SIP trunk will have gone through digit manipulation in the home cluster. The called number may require manipulation to match visiting cluster route patterns.
- Verify configured codec capabilities of H.323 and SIP gateways in the home cluster. For example, if home cluster gateways are configured to accept only G.711 calls and the EMCC region bandwidth is set to 8 kbps (G.729), a transcoder is required to complete the call. Alternatively, the H.323 or SIP gateway dial peers may be configured to allow for G.729 in addition to G.711.
- Design considerations must be made regarding the calling party for EMCC emergency calls. Depending on dial plan configurations, the calling party number leaving the visiting cluster gateway may be the user's DID that is normally associated with the home cluster. This would require transforming the calling number incoming on the EMCC SIP trunk, on route patterns, or egressing on the visiting gateways.
- When EMCC is deployed with Cisco Emergency Responder, Emergency Responder should be deployed in all clusters handled by a single Emergency Responder cluster. If the visiting cluster is deployed with Emergency Responder and the home cluster is not, Emergency Responder will not be able to identify the visiting phone when the call arrives back to the visiting cluster.

Unified CM Assistant

Cisco Unified Communications Manager Assistant (Unified CM Assistant) is a Unified CM integrated application that enables assistants to handle incoming calls on behalf of one or more managers. With the use of the Unified CM Assistant Console desktop application or the Unified CM Assistant Console phone service on the assistant phone, assistants can quickly determine a manager's status and determine what to do with a call. Assistants can manipulate calls using their phone's softkeys and service menus or via the PC interface with either keyboard shortcuts, drop-down menus, or by dragging and dropping calls to the managers’ proxy lines.

This section examines the following design aspects of the Unified CM Assistant feature:

- Unified CM Assistant Architecture, page 19-21
- High Availability for Unified CM Assistant, page 19-24
- Capacity Planning for Unified CM Assistant, page 19-27
- Design Considerations for Unified CM Assistant, page 19-29
- Unified CM Assistant Console, page 19-33

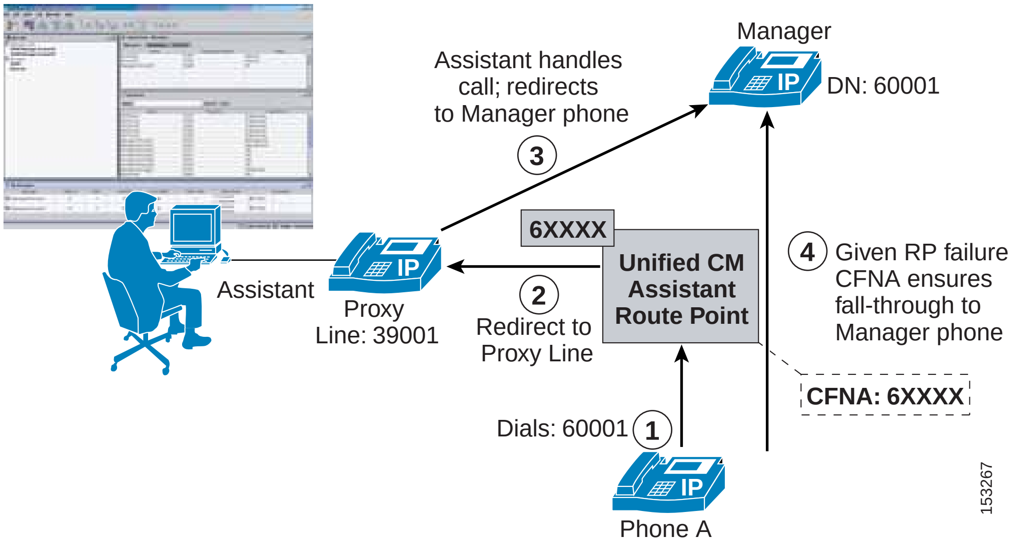
Unified CM Assistant Architecture

The Unified CM Assistant application can operate in two modes: proxy line mode and shared line mode. The operation and functionality of each mode is different, and each has specific advantages and disadvantages. Both modes can be configured within a single cluster. However, mixing modes on the same assistant is not allowed. A single assistant providing support for one or more managers can support those managers in either shared line mode or proxy line mode.

Unified CM Assistant Proxy Line Mode

Figure 19-7 illustrates a simple call flow with Unified CM Assistant in proxy line mode. In this example, Phone A calls the Manager phone with directory number (DN) 60001 (step 1). The CTI/Unified CM Assistant Route Point (RP) intercepts this call based on a configured DN of 6XXXX. Next, based on the Manager DN, the call is redirected by the route point to the Manager's proxy line (DN: 39001) on the Assistant's phone (step 2). The Assistant can then answer or handle the call and, if appropriate, redirect the call to the Manager's phone (step 3). In the event of Unified CM Assistant application failure or if the Unified CM Assistant RP fails, a fall-through mechanism exists via the Call Forward No Answer (CFNA) 6XXXX configuration of the RP, so that calls to the Manager's DN will fall-through directly to the Manager's phone (step 4).

Figure 19-7 Unified CM Assistant Proxy Line Mode

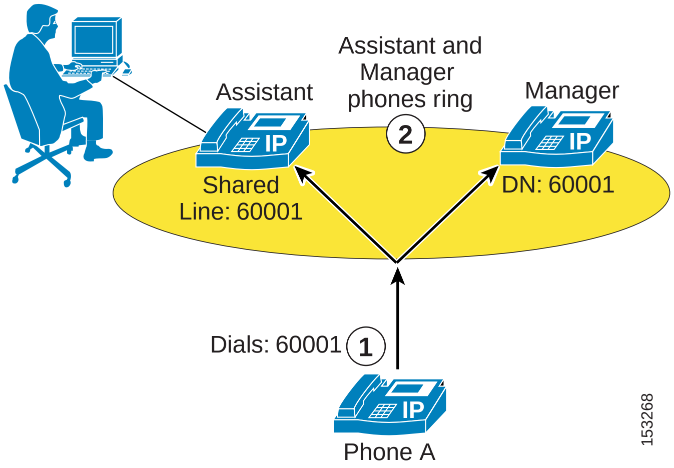

Manager
DN: 60001
Assistant handles call; redirects to Manager phone
3
6XXXX
Unified CM Assistant Route Point
4
Given RP failure, CFNA ensures fall-through to Manager phone
Assistant
Proxy Line: 39001
2
Redirect to Proxy Line
CFNA: 6XXXX
Dials: 60001
1
Phone A
153267

Note The CFNA fall-through mechanism illustrated in Figure 19-7 requires configuration of the same summarized digit-string as the Unified CM Assistant RP directory number in both the Forward No Answer Internal and Forward No Answer External fields under the Unified CM Assistant RP directory number configuration page. In addition, the calling search space (CSS) field for each of these call forward parameters should be configured with the calling search space containing the partition with which the Manager phone DNs are configured, so that the Manager phone DNs can be reached if the Unified CM Assistant RP or Unified CM Assistant application fails.

Unified CM Assistant Share Lined Mode

Figure 19-8 illustrates a simple call flow with Unified CM Assistant in shared line mode. In this example, Phone A calls the Manager phone with directory number (DN) 60001, which is a shared line on the Assistant phone (step 1). The call will ring at both the Assistant and Manager phones unless the Manager has invoked the Do Not Disturb (DND) feature, in which case the Assistant's phone will be the only phone that rings audibly (step 2).

Figure 19-8 Unified CM Assistant Shared Line Mode



In Unified CM Assistant shared line mode, the Unified CM Assistant RP is not needed or required for intercepting calls to the Manager phone. However, the Do Not Disturb (DND) feature on the Manager phone and the Unified CM Assistant Console desktop application still depend on the Cisco IP Manager Assistant (IPMA) and Cisco CTIManager services. Furthermore, in Unified CM Assistant shared line mode, features such as call filtering, call intercept, assistant selection, and Assistant Watch are not available.

Unified CM Assistant Architecture

The architecture of the Unified CM Assistant application is as important to understand as its functionality. Figure 19-9 depicts the message flows and architecture of Unified CM Assistant. When Unified CM Assistant has been configured for Unified CM Assistant Manager and Assistant users, the following sequence of interactions and events can occur:

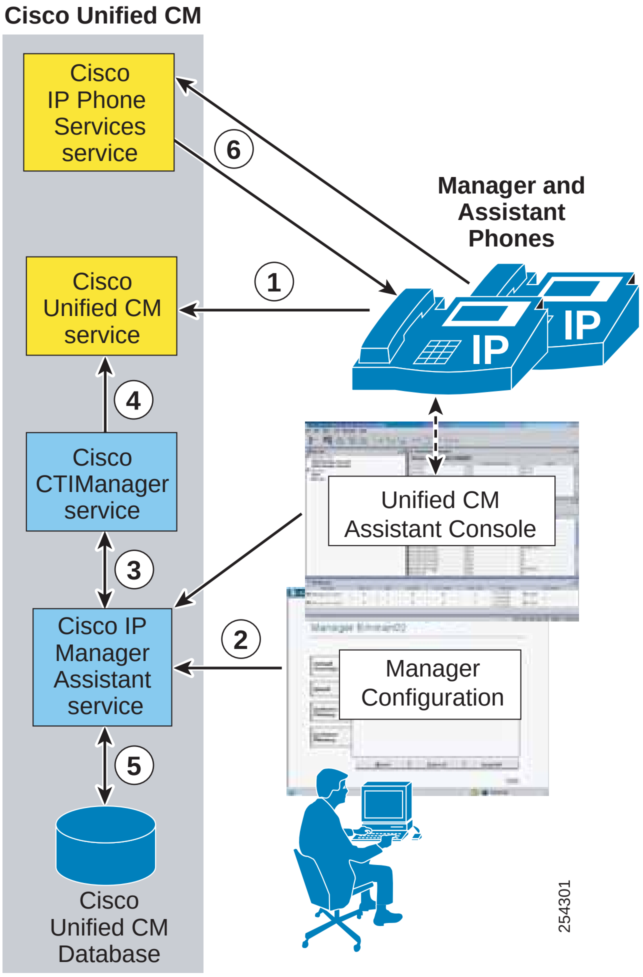
1. Manager and Assistant phones register with the Cisco CallManager Service, and the phone’s keypad and softkeys are used to handle call flows (see step 1 in Figure 19-9).
2. Both the Unified CM Assistant Console desktop application and the Manager Configuration web-based application communicate and interface with the Cisco IP Manager Assistant service (see step 2 in Figure 19-9).
3. The Cisco IP Manager Assistant service in turn interacts with the CTIManager service for exchanging line monitoring and phone control information (see step 3 in Figure 19-9).
4. The CTIManager service passes Unified CM Assistant phone control information to the Cisco CallManager service and also controls the Unified CM Assistant RP (see step 4 in Figure 19-9).

5. In parallel, the Cisco IP Manager Assistant service reads and writes Unified CM Assistant application information to and from the Unified CM database (see step 5 in Figure 19-9).
6. The Manager may choose to invoke the Unified CM Assistant phone service by pushing the Services button, thus generating a call to the IP Phone Services service that will return a list of all services (including the Unified CM Assistant phone service) to which the phone is subscribed (see step 6 in Figure 19-9).

The Unified CM Assistant phone service is controlled by the Cisco IP Manager Assistant service, and configuration changes made by the Manager using the phone are handled and propagated via the Cisco IP Manager Assistant service.

Note With the Services Provisioning enterprise parameter set to Internal, steps 1 and 2 are bypassed. Alternatively, with Services Provisioning set to External URL or Both, a Service URL button can be configured for the Unified CM Assistant phone service on a user's phone so that the user can press a line or speed-dial button to generate a direct call to the Cisco IP Manager Assistant service, also bypassing steps 1 and 2.

Figure 19-9 Unified CM Assistant Architecture


Cisco Unified CM
Cisco IP Phone Services service
Cisco Unified CM service
Cisco CTIManager service
Cisco IP Manager Assistant service
Cisco Unified CM Database
Manager and Assistant Phones
Unified CM Assistant Console
Manager Configuration
1
2
3
4
5
6
254301

Note While Figure 19-9 shows the IP Phone Services, Cisco CallManager, CTIManager, and Cisco IP Manager Assistant services all running on the same node, this configuration is not a requirement. These services can be distributed between multiple nodes in the cluster but have been shown on the same node here for ease of explanation.

High Availability for Unified CM Assistant

Unified CM Assistant application redundancy can be provided at two levels:

- Redundancy at the component and service level

  At this level, redundancy must be considered with regard to Unified CM Assistant service or server redundancy and CTIManager service redundancy. Likewise, the lack of publisher redundancy and the impact of this component failing should also be considered.

- Redundancy at the device and reachability level

  At this level, redundancy should be considered as it relates to Assistant and Manager phones, the Unified CM Assistant route point, and the Unified CM Assistant Console desktop application and phone service, as well as redundancy in terms of Assistant and Manager reachability.

Service and Component Redundancy

As shown in Figure 19-9, Unified CM Assistant functionality is primarily dependent on the Cisco IP Manager Assistant (IPMA) service and the Cisco CTIManager service. In both cases, redundancy is automatically built-in using a primary and backup mechanism. Up to three pairs of active and backup Unified CM Assistant servers (nodes running the Cisco IP Manager Assistant service) can be defined, for a total of six Unified CM Assistant servers within a single cluster. Active and backup Unified CM Assistant server pairs are configured using the Cisco IPMA Server IP Address, Pool 2 Cisco IPMA Server IP Address, and Pool 3 Cisco IPMA Server IP Address service parameters. With the configuration of these parameters, the required Cisco IP Manager Assistant service is made redundant. Given a failure of any of the primary Unified CM Assistant servers, the backup or standby Unified CM Assistant servers are able to handle Unified CM Assistant service requests. For each pair of Unified CM Assistant servers, only one Unified CM Assistant server can be active and handling request at a given time, while the other Unified CM Assistant server will be in a standby state and will not handle requests unless the active server fails.

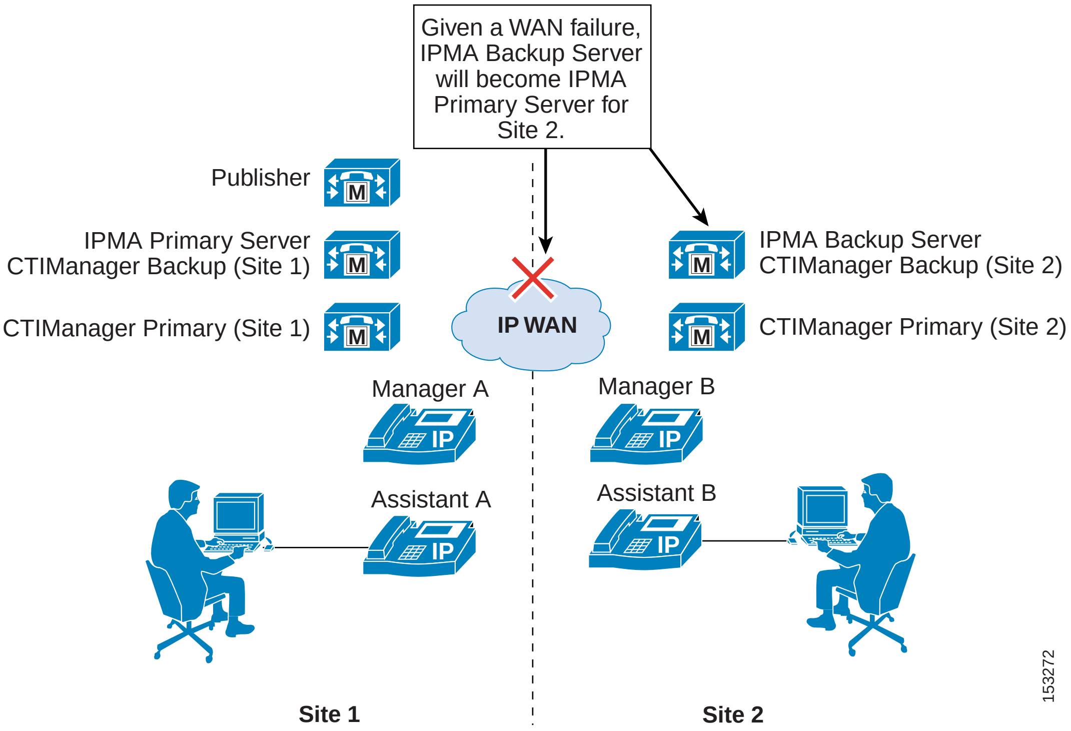
In addition, two CTIManager servers or services can be defined for each Unified CM Assistant server using the CTIManager (Primary) IP Address and CTIManager (Backup) IP Address service parameters. By configuring these parameters, you can make the CTIManager service redundant. Thus, given a failure of a primary CTIManager, CTIManager services can still be provided by the backup CTIManager. If all Cisco IP Manager Assistant and CTIManager services on cluster nodes fail, the Unified CM Assistant route point, Unified CM Assistant Console desktop application and phone service, and in turn the Unified CM Assistant application as a whole will fail. However as noted previously, given a failure of the Unified CM Assistant application, the CFNA fall-through mechanism will continue to work, allowing calls to a Manager to be routed directly to the Manager's phone.

Note If configured in Unified CM Assistant shared-line mode, a complete failure of Cisco IP Manager Assistant and CTIManager service will not keep the Assistant from continuing to handle calls on behalf of the Manager because the phones will continue to shared a line. However, the Unified CM Assistant Console desktop application and phone service and the DND feature will not be available.

Figure 19-10 shows an example redundancy configuration for Unified CM Assistant and CTIManager primary and backup servers in a two-site deployment with clustering over the WAN. In order to provide maximum redundancy, a node at Site 1 is configured as the primary Unified CM Assistant server and a node at Site 2 is configured as the backup Unified CM Assistant server. In the event of a WAN failure, the backup Unified CM Assistant server at Site 2 will become a primary Unified CM Assistant server because the existing primary Unified CM Assistant server will be unreachable from Site 2. In this way, Unified CM Assistant servers can be made redundant in the clustering-over-the-WAN environment given a WAN failure. Furthermore, with a primary and backup CTIManager configured at both Site 1 and Site 2, CTIManager is made redundant given a WAN failure, and additional redundancy is provided for a CTIManager failure at each site.

Note The redundancy scenario depicted in Figure 19-10 shows a special circumstance. During normal operation it is not possible to have any pair of Unified CM Assistant servers active at the same time. If an active and backup pair of Unified CM Assistant servers can communicate over the network, then one server will be in backup mode and cannot handle requests.

Figure 19-10 Unified CM Assistant Redundancy with Two-Site Clustering over the WAN


Given a WAN failure, IPMA Backup Server will become IPMA Primary Server for Site 2.
Publisher
M
IPMA Primary Server
CTIManager Backup (Site 1)
M
CTIManager Primary (Site 1)
M
IP WAN
IPMA Backup Server
CTIManager Backup (Site 2)
M
CTIManager Primary (Site 2)
M
Manager A
IP
Manager B
IP
Assistant A
IP
Assistant B
IP
Site 1
Site 2
153272

As previously mentioned, the publisher is a single point of failure when it comes to writing Unified CM Assistant information to the Unified CM database. Given a publisher failure, all aspects of the Unified CM Assistant application will continue to work; however, no changes to the Unified CM Assistant application configuration can be made. Configuration changes via the Unified CM Assistant Console desktop application, the Manager configuration web-based application, the phone softkeys, or the Unified CM Assistant phone service, will not be possible until the publisher is restored. This condition includes enabling or disabling features such as Do Not Disturb, DivertAll, Assistant Watch, and call filtering, as well as changing call filter and assistant selection configuration.

Device and Reachability Redundancy

Redundancy for Unified CM Assistant at the devices level relies on a number of mechanisms. First and foremost, manager and assistant phones as well as the Unified CM Assistant RP rely on the built-in redundancy provided by a combination of the device pool and Unified CM group configuration for device registration.

In addition, some devices rely on component services for additional redundancy and functionality. For example, the Unified CM Assistant RP also relies on CTIManager for call control functionality and therefore must rely on the primary and back CTIManager mechanism described in the previous section.

The Unified CM Assistant Console desktop application also relies on the component services for redundancy and functionality. The Assistant Console desktop application supports automatic failover from the primary to the backup Unified CM Assistant server (and vice versa) in order to continue to handle incoming calls for managers. The amount of time this automatic failover will take can be controlled using the Cisco IPMA Assistant Console Heartbeat Interval and the Cisco IPMA Assistant Console Request Timeout service parameters. Although the heartbeat or keep-alive frequency can be configured so that failures of the Unified CM Assistant server are detected by the desktop application more quickly, be careful not to affect the network adversely by sending keep-alives too frequently. This consideration is especially important if there are a large number of Assistant Console desktop applications in use.

The Unified CM Assistant Console phone service, unlike the Unified CM Assistant Console desktop application, requires manual intervention for redundancy given the failure of the primary Unified CM Assistant server. If the primary Unified CM Assistant server goes down, assistants using the phone console will not see an indication of this condition. However, the assistant phone will receive a "Host not found Exception" message upon trying to use a softkey. In order to continue using the phone console with the backup Unified CM Assistant server, the user must manually select the secondary Unified CM Assistant phone service from the IP Services menu and log in again.

There are several other failover mechanisms which ensure that Manager and Assistant reachability are redundant. First, calls sent to a Manager's Assistant via the Unified CM Assistant application (in proxy line mode) can be forwarded to the Manager's next available Assistant if the call is not answered after a configured amount of time. If the next Assistant does not answer the call after the configured amount of time, the call can again be forwarded to the Manager's next available Assistant, and so on. The mechanism is configured using the Cisco IPMA RNA Forward Calls and Cisco IPMA RNA Timeout service parameters. Second, as mentioned previously, if all Cisco IP Manager Assistant and CTI services on cluster nodes fail, the Unified CM Assistant RP will become unavailable. However, based on the CFNA configuration of the Unified CM Assistant RP, calls to all Manager DNs will fall-through directly to the Manager phones so that Manager reachability is sufficiently redundant.

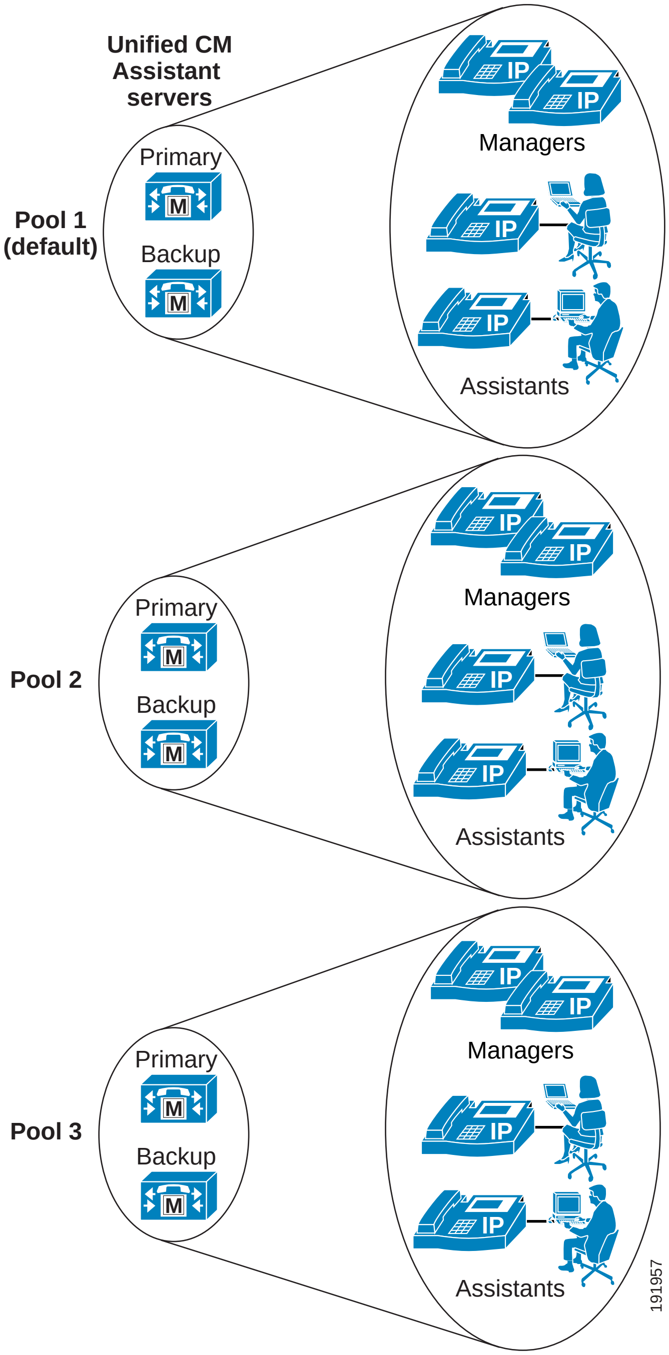
Capacity Planning for Unified CM Assistant

The Cisco Unified CM Assistant application supports the following capacities:

- A maximum of 10 Assistants can be configured per Manager.
- A maximum of 33 Managers can be configured for a single Assistant (if each Manager has one Unified CM Assistant-controlled line).
- A maximum of 3500 Assistants and 3500 Managers (7000 total users) can be configured per cluster using the Cisco MCS 7845 server or OVA equivalent.
- A maximum of three pairs of primary and backup Unified CM Assistant servers can be deployed per cluster if the Enable Multiple Active Mode advanced service parameter is set to True and a second and third pool of Unified CM Assistant servers are configured.

In order to achieve the maximum Unified CM Assistant user capacity of 3500 Managers and 3500 Assistants (7000 users total), multiple Unified CM Assistant server pools must be defined. As illustrated in Figure 19-11, up to three pools can be configured. Each pool consists of a primary and backup Unified CM Assistant server and a group of Managers and Assistants. Pool 1's Unified CM Assistant servers are configured with the Cisco IPMA Server (Primary/Backup) IP Address service parameters, Pool 2's servers are configured with the Pool2: Cisco IPMA Server (Primary/Backup) IP Address advanced service parameters, and Pool 3's servers are configured with the Pool3: Cisco IPMA Server (Primary/Backup) IP Address advanced service parameters.

Figure 19-11 Multiple Active Mode with Unified CM Assistant Server Pools



The Cisco Unified CM Assistant application interacts with the CTIManager for line monitoring and phone control. Each line (including Intercom lines) on a Unified CM Assistant or Manager phone requires a CTI line from the CTIManager. In addition, each Unified CM Assistant route point requires a CTI line instance from the CTIManager. When you configure Unified CM Assistant, the number of required CTI lines or connections must be considered with regard to the overall cluster limit for CTI lines or connections. (For more information on CTI connection limits per cluster, see Capacity Planning for CTI, page 8-34.) If additional CTI lines are required for other applications, they can limit the capacity of Unified CM Assistant.

Design Considerations for Unified CM Assistant

Unified CM Assistant has the following limitations with regard to overlapping and shared extensions, which you should keep in mind when planning directory number provisioning:

- With Unified CM Assistant in proxy line mode, the proxy line number(s) on the assistant phone should be unique, even across different partitions.
- With Unified CM Assistant in proxy line mode, two Managers cannot have the same Unified CM Assistant controlled line number (DN), even across different partitions.

When enabling Multiple Active Mode and using more than one Unified CM Assistant server pool, ensure that the appropriate server pool (1 to 3) is selected in the Assistant Pool field under the end user Manager Configuration page so that Managers and Assistants are evenly distributed between the Unified CM Assistant server pools. A Manager's associated Assistant will automatically be assigned to the pool where their Manager is configured.

Unified CM Assistant supports a non-secure or secure connection (Transport Layer Security) to the CTI Manager.

Unified CM Assistant Extension Mobility Considerations

Unified CM Assistant Managers can use Extension Mobility (EM) to log in to their phones in both proxy-line and shared-lined modes. However, the Manager must be configured as a Mobile Manager under the Cisco Unified CM Assistant Manager configuration page of the End-user Directory. When using EM in conjunction with Unified CM Assistant, users should not be able to log in to more than one phone using EM. This behavior can be enabled/disabled via the EM service parameter Multiple Login Behavior. If multiple EM logins by the same user are required within the cluster, Unified CM Assistant Managers who use EM should be instructed not to log in to multiple phones. Allowing a manager to log in to two different phones with EM violates the previously stated restriction that, in proxy line mode, two Managers cannot have the same Unified CM Assistant controlled line number (DN), even across different partitions.

Note Unified CM Assistants cannot use EM to log in to their phones because there is no concept of a Mobile Assistant.

Unified CM Assistant Dial Plan Considerations

Dial plan configuration is extremely important for Unified CM Assistant configured in proxy line mode. To ensure that calls to Manager DNs are intercepted by the Unified CM Assistant RP and redirected to the Assistant phone, calling search spaces and partitions must be configured in such a way that Manager DNs are unreachable from all devices except the Unified CM Assistant RP and the Manager's proxy line on the Assistant phone.

Figure 19-12 shows an example of a proxy line mode Unified CM Assistant dial plan with the minimum requirements for calling search spaces, partitions, and the configuration of various types of devices within these dial plan components. Three partitions are required for proxy line mode, and for the example in Figure 19-12 they are as follows:

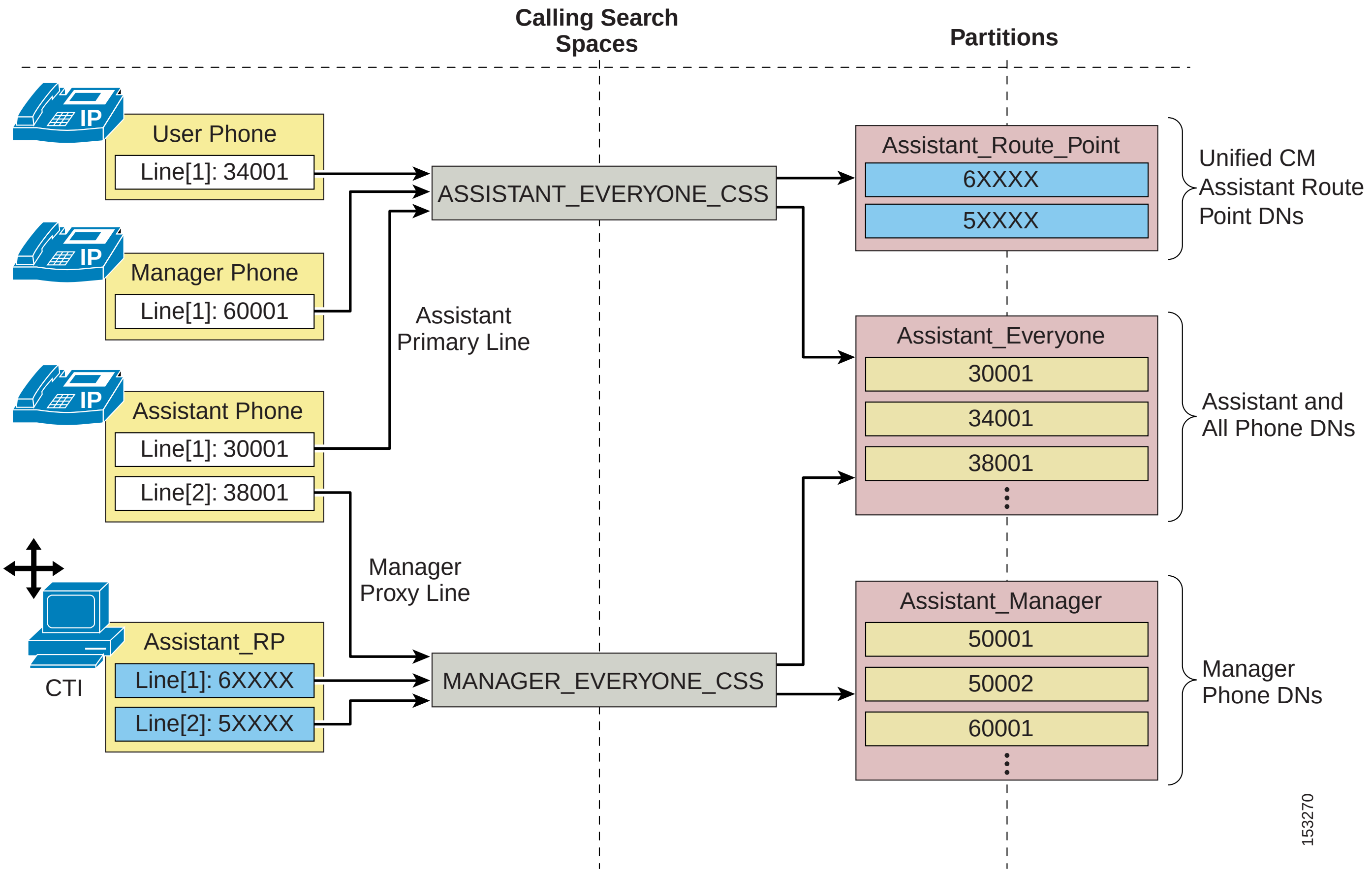
- Assistant_Route_Point partition, containing all the Unified CM Assistant RP DNs
- Assistant_Everyone partition, containing all the Assistant and other user phone DNs
- Assistant_Manager partition, containing all the Manager phone DNs

In addition, two calling search spaces are required, and for the example in Figure 19-12 they are as follows:

- ASSISTANT_EVERYONE_CSS calling search space, containing both the Assistant_Route_Point and Assistant_Everyone partitions.
- MANAGER_EVERYONE_CSS calling search space, containing both the Assistant_Manager and Assistant_Everyone partitions.

That is the extent of the dial plan for this example. However, it is also important to properly configure the various phone and Unified CM Assistant RP DNs or lines with the appropriate calling search spaces so that call routing works as required. In this case all user, Assistant primary (or personal), and Manager phone lines would be configured with the ASSISTANT_EVERYONE_CSS calling search space so that all of these lines can reach all the DNs in the Assistant_Everyone and Assistant_Route_Point partitions. Intercom lines and any other lines configured on devices within the telephony network would be configured with this same calling search space. All Manager proxy lines and all Assistant_RP lines are configured with the MANAGER_EVERYONE_CSS calling search space so that all of these lines can reach the Manager DNs in the Assistant_Manager partition as well as all the DNs belonging to the Assistant_Everyone partition. In this way, the dial plan ensures that only the Assistant_RP lines and the Manager proxy lines on the Assistant phones are capable of reaching the Manager phone DNs directly.

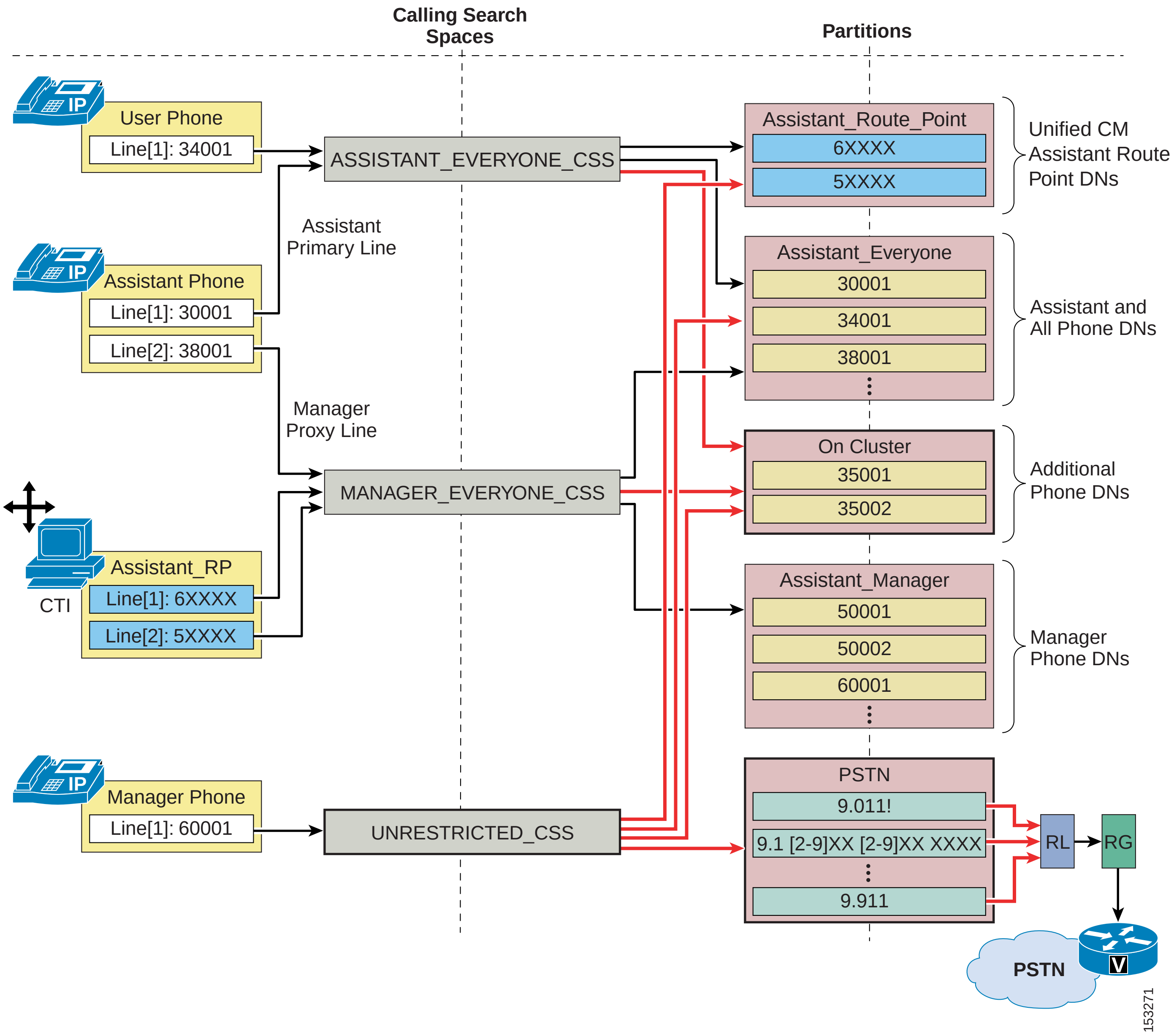
Figure 19-12 Unified CM Assistant Proxy Line Mode Dial Plan Example



The example in Figure 19-12 shows the minimum dial plan requirements for Unified CM Assistant in proxy line mode. However, most real-world telephony networks will have additional or existing dial plan requirements that must be integrated with the Unified CM Assistant calling search spaces and partitions. Figure 19-13 illustrates such an integration dial plan. In this example, the previously discussed dial plan must now handle two additional partitions and an additional calling search space. The On Cluster partition has been added in Figure 19-13, and it contains some additional phone DNs. The On Cluster partition has been added to both of the existing Unified CM Assistant calling search spaces (ASSISTANT_EVERYONE_CSS and MANAGER_EVERYONE_CSS) so that existing devices can reach these added DNs. The UNRESTRICTED_CSS calling search space has also been added to the existing dial plan. This calling search space is configured with the Assistant_Route_Point, Assistant_Everyone, and the recently added On Cluster partitions. In addition, a second new partition called PSTN has been added, and it contains a set of route patterns used for routing calls to the PSTN via the common route list (RL), route group (RG), and voice gateway mechanism. This PSTN partition is configured as part of the UNRESTRICTED_CSS calling search space.

Phone and device line calling search space configurations may be adjusted to incorporate the newly added partitions and calling search spaces, provided the Assistant_RP and Assistant phone Manager proxy lines remain assigned to the MANAGER_EVERYONE_CSS calling search space. In this example, the Manager phone line has been moved from the originally configured ASSISTANT_EVERYONE_CSS calling search space to the new UNRESTRICTED_CSS because it is likely that a Manager would be given unrestricted access to the PSTN.

Figure 19-13 Unified CM Assistant Proxy Line Mode Dial Plan Integration Example



As Figure 19-13 illustrates, integrating additional partitions and calling search spaces into a new or existing Unified CM Assistant dial plan is feasible, but care must be taken to ensure that the underlying proxy line mode mechanism remains intact.

For Unified CM Assistant shared line mode, no special dial plan provisioning is required. Manager and Assistant phones can be configured with calling search spaces and partitions like any other phones in the network because there are no Unified CM Assistant RPs or proxy lines to be concerned about. The only requirement with regard to shared line mode is that the Manager and Assistant DNs must be in the same partition so that shared line functionality is possible.

Unified CM Assistant Console

The Unified CM Assistant Console desktop application or the Unified CM Assistant Console phone service is required in order for assistants to handle calls on a manager's behalf. The desktop application provides assistants with a graphical interface for handling calls, while the phone service provides a menu-driven interface for handling calls. Both the desktop application and the IP phone service allow the assistant to configure the Manager phone and environment and monitor line status and availability. In addition, the desktop application provides other functions such as click-to-call speed dialing and directory entries, which can also be performed on the assistant phone using the traditional softkey and menu approach.

Unified CM Assistant Console Installation

The Unified CM Assistant Console desktop application can be installed from the following URL:

https://<*Server_IP-Address*>:8443/plugins/CiscoUnifiedCallManagerAssistantConsole.exe

(where <*Server_IP-Address*> is the IP address of any node in the cluster)

The Unified CM Assistant Console phone service does not require any installation. To enable the Assistant's phone as a console, subscribe the phone to the Unified CM Assistant phone service. (This is the same service to which Manager phones must also be subscribed.)

Unified CM Assistant Desktop Console QoS

After installation, and in order to handle calls on a Manager's behalf, the Assistant must log on to the application by providing userID and password (as configured in the End-user directory on Unified CM) and will have to toggle status to "online" by clicking the Go Online icon or menu item. Once the user is logged in and online, the desktop application communicates with the Unified CM Assistant server at TCP port 2912. The application chooses an ephemeral TCP port when sourcing traffic. Because the Unified CM Assistant server on Unified CM interfaces with the desktop application for call control (generation and handling of call flows), traffic sourced from Unified CM on TCP port 2912 is QoS-marked by Unified CM as Differentiated Services Code Point (DSCP) of 24 or Per Hop Behavior (PHB) of CS3. In this way, Unified CM Assistant phone control traffic can be queued throughout the network like all other call signaling traffic.

In order to ensure symmetrical marking and queuing, the Unified CM Assistant Console application traffic destined for Unified CM TCP port 2912 should also be marked as DSCP 24 (PHB CS3) to ensure this traffic is placed in the appropriate call signaling queues along the network path toward Unified CM and the Unified CM Assistant server. The Unified CM Assistant Console application marks all traffic as best-effort. This means that you will have to apply an access control list (ACL) at the switch port level (or somewhere along the network path, preferably as close to the console PC as possible) to remark traffic sent by the application PC destined for Unified CM on TCP port 2912 from DSCP 0 (PHB Best Effort) to DSCP 24 (PHB CS3).

Unified CM Assistant Console Directory Window

The directory window within the Assistant Console desktop application enables an assistant to search for end-users in the Unified CM Directory. Search strings entered into the Name field of the directory window are sent to the Unified CM Assistant server, and searches are generated directly against the Unified CM database. Responses to search queries are then sent back to the desktop application by the Unified CM Assistant server.

While the additional traffic generated by directory searches within the desktop application is nominal, this traffic can be problematic in centralized call processing deployments when one or more Unified CM Assistant console applications are running at remote sites. A directory search resulting in a single entry generates approximately one (1) kilobit of traffic from the Unified CM Assistant server to the desktop application. Fortunately, a maximum of 25 entries can be retrieved per search, meaning that a maximum of approximately 25 kilobits of traffic can be generated for each search made by the desktop application. However, if directory searches are made by multiple Unified CM Assistant Console desktop applications across low-speed WAN links from the Unified CM Assistant server, the potential for congestion, delay, and queuing is increased. In addition, directory retrieval traffic is sourced from Unified CM on TCP port 2912, like all other Unified CM Assistant traffic to the desktop. This means that directory retrieval traffic is also marked with DSCP 24 (PHB CS3) and therefore is queued like call signaling traffic. As a result, directory retrieval could potentially congest, overrun, or delay call control traffic.

Note If a directory search generates more than 25 entries, the assistant is warned via a dialog box with the message: “Your search returned more than 25 entries. Please refine your search.”

Given the potential for network congestion, Cisco recommends that administrators encourage Unified CM Assistant Console users to do the following:

- Limit their use of the directory window search function.
- To reduce the number of entries returned, enter as much information as possible in the Name field and avoid wild-card or blank searches when using the feature.

These recommendations are especially important if either of the following conditions is true:

- There are many Unified CM Assistant Assistants within the cluster.
- There are many assistants separated from the Unified CM and/or Unified CM Assistant servers by low-speed WAN links.

Unified CM Assistant Phone Console QoS

In order to handle calls on a Manager's behalf using the Unified CM Assistant Phone Console phone service, the Assistant must log on to the service by providing a userID and PIN (as configured in the End-user directory on Unified CM). Once the user is logged in, the phone console service communicates with Unified CM using HTTPS and SCCP. Call control traffic for Unified CM Assistant call generation and call handling is sent between the phone and Unified CM using SCCP. By default this traffic is marked as Differentiated Services Code Point (DSCP) of 24 or Per Hop Behavior (PHB) of CS3, thus ensuring it is queued throughout the network as call signaling traffic, therefore no additional QoS configuration or marking is required.

WebDialer

WebDialer is a click-to-call application for Unified CM that enables users to place calls easily from their PCs using any supported phone device. There is no requirement for administrators to manage CTI links or build JTAPI or TAPI applications because Cisco WebDialer provides a simplified web application and HTTP or Simple Objects Access Protocol (SOAP) interface for those who want to provide their own user interface and authentication mechanisms. Alternatively, the **Click to Call** Cisco Unified Communications Widget makes use of the SOAP interface and is currently available for download (login authentication required) at

http://tools.cisco.com/support/downloads/go/Redirect.x?mdfid=278875240

This section examines the following design aspects of the WebDialer feature:

- WebDialer Architecture, page 19-35
- High Availability for WebDialer, page 19-40
- Capacity Planning for WebDialer, page 19-41
- Design Considerations for WebDialer, page 19-42

WebDialer Architecture

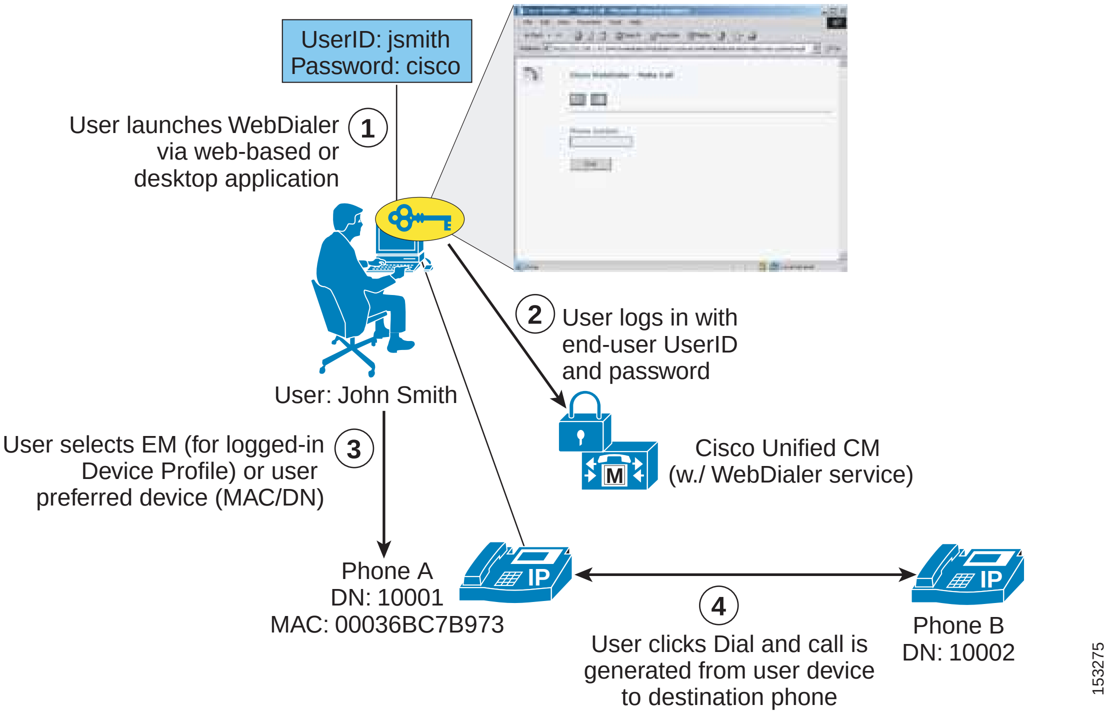
The WebDialer application contains two servlets: the WebDialer servlet and the Redirector servlet. Both servlets are enabled when the Cisco WebDialer Web service is activated on a subscriber server. While related, they each serve different functions and can be configured to run simultaneously.

WebDialer Servlet

Figure 19-14 illustrates a simple WebDialer example. In this example, user John Smith launches WebDialer from a web-based or desktop application such as the Click to Call Cisco Unified Communications Widget (step 1). WebDialer responds with a request for login credentials. The user must respond with a valid userID and password as configured in the Unified CM end-user directory. In this case, John Smith submits userID = jsmith and password = cisco (step 2). Next, based on this login, WebDialer responds with the Cisco WebDialer Preferences configuration page, and the user must indicate either “User preferred device” or “Use Extension Mobility” (assuming the user has an EM device profile). In this case, user John Smith selects “User preferred device” and selects the appropriate MAC address (SEP00036BC7B973) and directory number (10001) for his phone from drop-down menus on the configuration page (step 3). Finally, the user is presented with a screen requesting the phone number to be called (this value may already be indicated) and must click Dial. In this case, John Smith enters 10002 and, after clicking Dial, a call is automatically generated from his phone to Phone B at number 10002 (step 4).

Note If the user has previously logged in to the WebDialer application and a web browser and server cookie are still active, the user will not be prompted to log in again during subsequent requests. The user will be prompted to log in again when the cookie has been cleared at the browser or by a restart of the WebDialer server. Alternatively, the user web browser cookie can be set to expire automatically after a certain number of hours as configured by the User Session Expiry WebDialer service parameter.

Figure 19-14 WebDialer Servlet Operation


UserID: jsmith
Password: cisco
1
User launches WebDialer via web-based or desktop application
2
User logs in with end-user UserID and password
User: John Smith
Cisco Unified CM (w./ WebDialer service)
3
User selects EM (for logged-in Device Profile) or user preferred device (MAC/DN)
Phone A
DN: 10001
MAC: 00036BC7B973
4
User clicks Dial and call is generated from user device to destination phone
Phone B
DN: 10002
153275

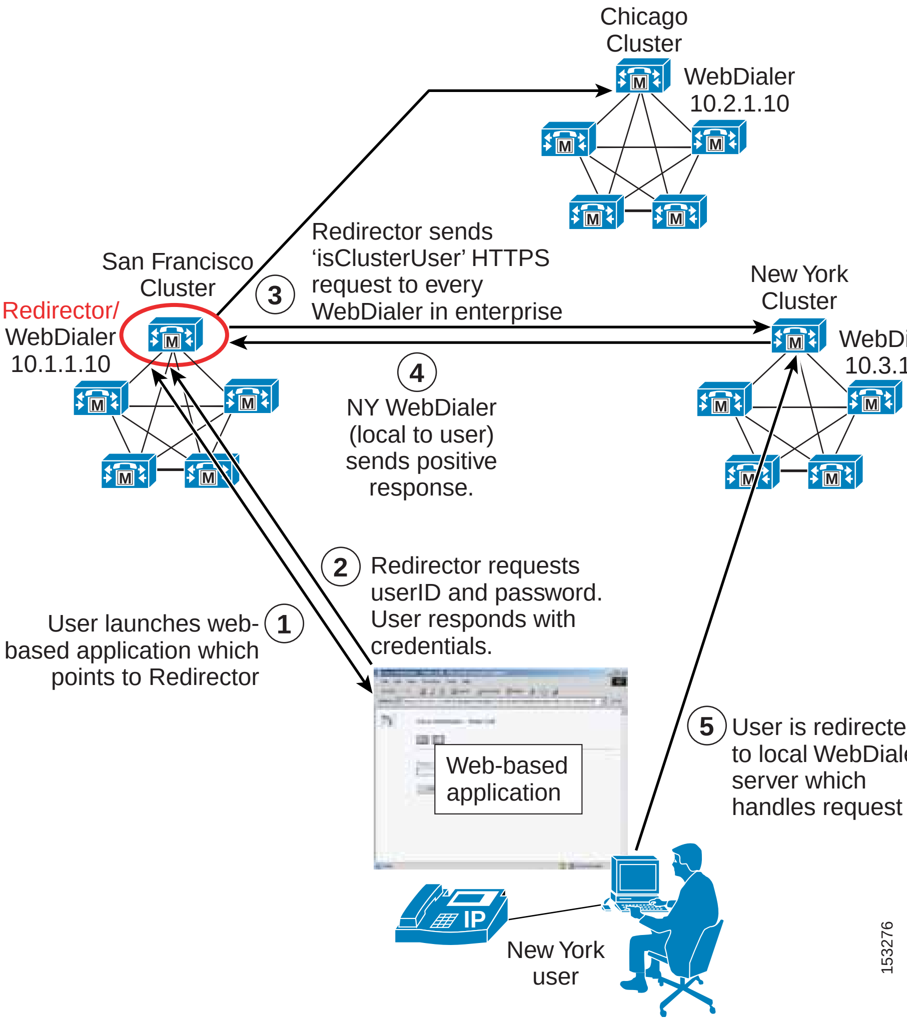
Redirector Servlet

The Redirector servlet provides WebDialer functionality in a multi-cluster or distributed call processing environment. This functionality allows the use of a single enterprise-wide web-based WebDialer application between all Unified CM clusters. Figure 19-15 illustrates the basic operation of the Redirector servlet as part of the WebDialer application. In this example, the enterprise has three Unified CM clusters: New York, Chicago, and San Francisco. All three clusters have been configured with a single WebDialer application. The San Francisco cluster has been designated as the Redirector.

Note If the user has previously logged in to the WebDialer application and a web browser and server cookie are still active, the user will not be prompted to log in again during subsequent requests. Alternatively, the user web browser cookie can be set to expire automatically after a certain number of hours as configured by the User Session Expiry WebDialer service parameter.

The Redirector then broadcasts an isClusterUser HTTPS request to every WebDialer in the enterprise simultaneously (as configured in the List of WebDialers service parameter). In this example, the requests go to the Chicago and New York WebDialer servers (see step 3 in Figure 19-15). Because the New York user is local to the New York cluster, the New York WebDialer responds with a positive response (see step 4 in Figure 19-15). Finally, the New York user is redirected to their local WebDialer server, which will handle the application request (see step 5 in Figure 19-15). The user is not notified of the redirect; however, the URL in the browser address bar will be changed as the user is redirected from the Redirector to the local WebDialer server). In this example, only one Redirector is deployed; but in order to provide redundancy for the Redirector, configure the Redirector on multiple clusters, as discussed in the section on Service and Component Redundancy, page 19-41.

Figure 19-15 Redirector Servlet Operation



Note Because the Redirector application is an enterprise-wide application that requires user authentication against the Unified CM Database, Cisco highly recommends that all end-user userIDs be unique across all Unified CM clusters. If they are not, then it is possible that more than one positive response to the isClusterUser request could be received by the Redirector application. If this happens, the user will be asked by the Redirector application to select their local WebDialer server manually. The user will then have to know which server is their local server. If the wrong server is chosen, the WebDialer request will fail.

WebDialer Architecture

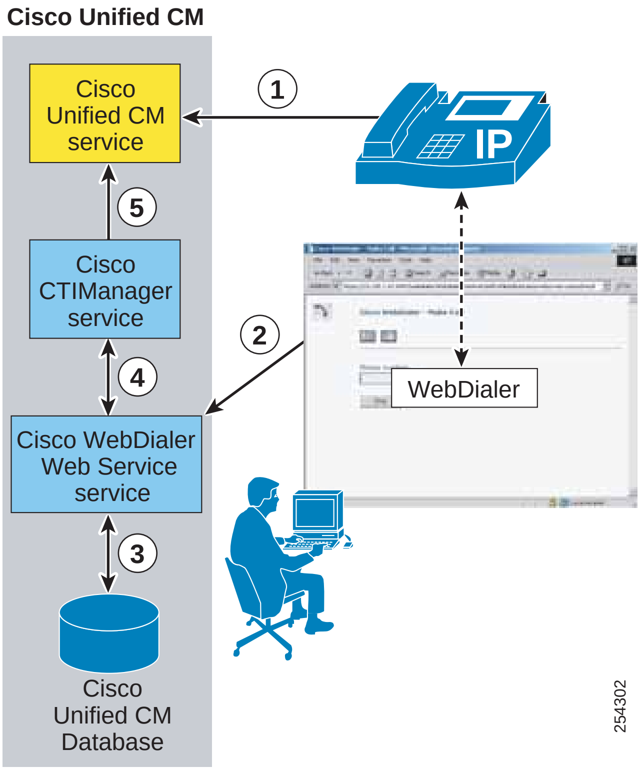
The architecture of the WebDialer application is as important to understand as its functionality. Figure 19-16 depicts the message flows and architecture of WebDialer. The following sequence of interactions and events can occur:

1. WebDialer user phones register and make and receive calls via the Cisco CallManager service (see step 1 in Figure 19-16).
2. The WebDialer application on the user's PC communicates with the Cisco WebDialer Web Service (see step 2 in Figure 19-16) via one of the following interfaces:
   - HTML over HTTPS

     This interface is used by web-based applications based on the HTTPS protocol. This is the only interface that provides access to the WebDialer and Redirector servlets.
   - Simple Object Access Protocol (SOAP) over HTTPS

     This interface is used by desktop applications based on the SOAP interface.
3. The WebDialer Web service reads user and phone information from the Unified CM Database (see step 3 in Figure 19-16).
4. The WebDialer Web service in turn interacts with the CTIManager service for exchanging line and phone control information (see step 4 in Figure 19-16).
5. The CTIManager service passes WebDialer phone control information to the Cisco CallManager service (see step 5 in Figure 19-16).

Figure 19-16 WebDialer Architecture



Note Although Figure 19-16 shows the Cisco CallManager, CTIManager, and WebDialer Web Service services all running on the same node, this configuration is not a requirement. These services can be distributed among multiple nodes in the cluster, but they are shown on the same node here for ease of explanation.

WebDialer URLs

The WebDialer application can be accessed from web-based applications via the HTML-over-HTTPS interface using the following URLs:

- WebDialer servlet

  https://<*Server-IP_Addr*>:8443/webdialer/Webdialer?destination=<*Number_to_dial*>

  (where <*Server_IP-Address*> is the IP address of any node in the cluster running the Cisco WebDialer Web Service service, and where <*Number_to_dial*> is the number that the WebDialer user wishes to dial)

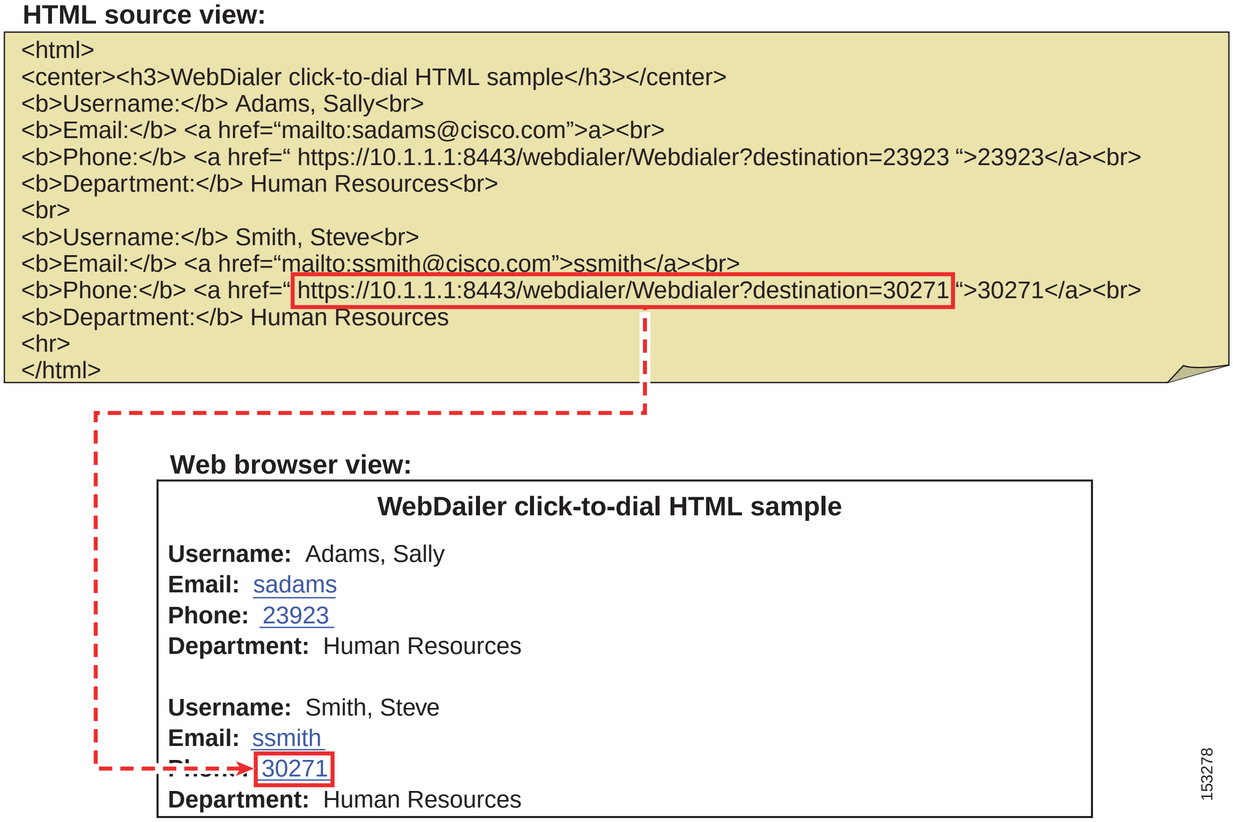
- Redirector servlet

  https://<*Server-IP_Addr*>:8443/webdialer/Redirector?destination=<*Number_to_dial*>

  (where <*Server_IP-Address*> is the IP address of any node in the enterprise running the Cisco WebDialer Web Service service, and where <*Number_to_dial*> is the number that the WebDialer user wishes to dial)

Figure 19-17 gives an example of HTML source code used in a click-to-call web-based application calling the Cisco WebDialer application. In this example, the URL https://10.1.1.1:8443/webdialer/Webdialer?destination=30271 in the HTML source view corresponds to the "Phone: 30721" link for user Steve Smith within the web browser view. A user clicking on this link would launch the WebDialer application and, after logging in and clicking Dial, would generate a call from the user's phone to Steve Smith's phone. The same code could be used for a click-to-call application using the Redirector function by changing the URL to https://10.1.1.1:8443/webdialer/Redirector?destination=30271.

Figure 19-17 WebDialer URL HTML Example



```
<html>
<center><h3>WebDialer click-to-dial HTML sample</h3></center>
<b>Username:</b> Adams, Sally<br>
<b>Email:</b> <a href=“mailto:sadams@cisco.com”>a><br>
<b>Phone:</b> <a href=“ https://10.1.1.1:8443/webdialer/Webdialer?destination=23923 “>23923</a><br>
<b>Department:</b> Human Resources<br>
<br>
<b>Username:</b> Smith, Steve<br>
<b>Email:</b> <a href=“mailto:ssmith@cisco.com”>ssmith</a><br>
<b>Phone:</b> <a href=“ https://10.1.1.1:8443/webdialer/Webdialer?destination=30271 “>30271</a><br>
<b>Department:</b> Human Resources
<hr>
</html>
```

WebDailer click-to-dial HTML sample

Username: Adams, Sally
Email: sadams
Phone: 23923
Department: Human Resources

Username: Smith, Steve
Email: ssmith
Phone: 30271
Department: Human Resources

For information and examples of SOAP-over-HTTPS source code to be used in click-to-call desktop applications, refer to the WebDialer API Programming information in the *Cisco Unified Communications Manager Developers Guide*, available at

http://www.cisco.com/en/US/products/sw/voicesw/ps556/products_programming_reference_guides_list.html

High Availability for WebDialer

WebDialer application redundancy can be provided at two levels:

- Redundancy at the component and service level

  At this level, redundancy must be considered with regard to WebDialer and CTIManager service redundancy. Likewise, the lack of publisher redundancy and the impact of this component failing should also be considered.

- Redundancy at the device and reachability level

  At this level, redundancy should be considered as it relates to user phones and the WebDialer user interface.

Service and Component Redundancy

As shown in Figure 19-16, WebDialer functionality is primarily dependent on the Cisco WebDialer Web Service and the Cisco CTIManager services. The WebDialer service can be enabled on multiple nodes within the cluster. Reachability to those multiple nodes is described in the section on Device and Reachability Redundancy, page 19-41. In the case of CTIManager, redundancy is automatically built-in using a primary and backup mechanism. Two CTIManager servers or services can be defined within the cluster using the Primary Cisco CTIManager and the Backup Cisco CTIManager service parameters. By configuring these parameters, you can make the CTIManager service redundant. Thus, if the primary CTIManager fails, CTIManager services can still be provided by the backup CTIManager. If the WebDialer server to which the web-based (or desktop) application is pointing fails and the primary and backup CTIManager services on cluster nodes also fail, the WebDialer application will fail. The WebDialer service is not dependant upon the Unified CM publisher

Device and Reachability Redundancy

Redundancy for WebDialer at the device level relies on a number of mechanisms. First and foremost, user phones rely on the built-in redundancy provided by a combination of the device pool and Unified CM group configuration for device registration.

The WebDialer service can run on multiple Unified CM subscribers in the same cluster to provide redundancy, however many applications might not be equipped to handle more than one IP address. Cisco recommends using a Server Load Balancer (SLB) to mask the presence of multiple WebDialer servers in the enterprise. SLB functionality provides a virtual IP address or DNS-resolvable hostname that front-ends the real IP addresses of the WebDialer servers. Most SLB devices, such as the Cisco Application Control Engine (ACE) or the Cisco IOS SLB feature, can be configured to monitor the status of multiple WebDialer servers and automatically redirect requests during failure events. The SLB feature can also be configured to load-balance WebDialer requests when additional click-to-call capacity is required. As an alternative, DNS Service (SRV) records can also be used to provide redundancy.

Similarly in a multi-cluster environment, if a single Redirector servlet is supporting multiple WebDialers, it could be a single point of failure. To avoid this single point of failure, configure Redirector servlets for each cluster and use a Server Load Balancer (SLB) to provide a virtual IP address or DNS-resolvable hostname that front-ends the real IP addresses of the Redirector servers.

In enterprise deployments, link cost might also be an important consideration. The Cisco ACE Global Site Selector (GSS) appliance builds upon the capabilities of the SLB feature by adding link cost and location to the load-balancing algorithm, among other features. For more information on ACE and GSS, refer to the product documentation available at http://www.cisco.com.

Capacity Planning for WebDialer

The WebDialer and Redirector services can run on one or more subscriber nodes within a Unified CM cluster, and they support the following capacities:

- Each WebDialer service can handle up to 4 call requests per second per node.
- Each Redirector service can handle up to 8 call requests per second.

The following general formula can be used to determine the number of WebDialer calls per second (cps):

(Number of WebDialer users) ∗ ((Average BHCA) / (3600 seconds/hour))

When performing this calculation, it is important to estimate properly the number of BHCA per user that will be initiated specifically from using the WebDialer service. The following example illustrates the use of these WebDialer design calculations for a sample organization.

Example 19-1 Calculating WebDialer Calls per Second

Company XYZ wishes to enable click-to-call applications using the WebDialer service, and their preliminary traffic analysis resulted in the following information:

- 10,000 users will be enabled for click-to-call functionality.
- Each user averages 6 BHCA.
- 50% of all calls are dialed outbound, and 50% are received inbound.
- Projections estimate 30% of all outbound calls will be initiated using the WebDialer service.

Note These values are just examples used to illustrate a WebDialer deployment sizing exercise. User dialing characteristics vary widely from organization to organization.

10,000 users each with 6 BHCA equates to a total of 60,000 BHCA. However, WebDialer deployment sizing calculations must account for placed calls only. Given the initial information for this sizing example, we know that 50% of the total BHCA are placed or outbound calls. This results in a total of 30,000 placed BHCA for all the users enabled for click-to-call using WebDialer.

Of these placed calls, the percentage that will be initiated using the WebDialer service will vary from organization to organization. For the organization in this example, several click-to-call applications are made available to the users, and it is projected that 30% of all placed calls will be initiated using WebDialer.

(30,000 placed BHCA) ∗ 0.30 = 9,000 placed BHCA using WebDialer

To determine the number of WebDialer servers required to support a load of 9,000 BHCA, we convert this value to the average call attempts per second required to sustain this busy hour:

(9,000 call attempts / hour) ∗ (hour/3600 seconds) = 2.5 cps

Each WebDialer service can support up to 4 cps, therefore one node should be configured to run the WebDialer service in this example. In order to maintain WebDialer capacity during a server failure, additional backup WebDialer servers should be deployed to provide redundancy.

Keep in mind that the Cisco WebDialer application interacts with the CTIManager for phone control. When enabled, each WebDialer service opens a single persistent CTI connection to the CTIManager. In addition, each WebDialer individual MakeCall (or EndCall) request generates a temporary CTI connection. The number of CTI connections required to handle WebDialer call rates also applies against the CTI connection limits per cluster. (For more information on CTI connection limits per cluster, see Capacity Planning for CTI, page 8-34.)

Design Considerations for WebDialer

The following guidelines and restrictions apply with regard to deployment and operation of WebDialer within the Unified CM telephony environment:

- The administrator should ensure that all WebDialer users are associated with a phone or device profile in the Unified CM end-user directory.
  - If the user selects "Use permanent device" under the Cisco WebDialer Preferences screen with no phone association, then the following message is received when the Dial button is pressed:

    “No supported device configured for user”

  - If the user selects Use Extension Mobility under the Cisco WebDialer Preferences screen with no device profile association (or the user is not logged in using a profile), then the following message is received when the Dial button is pressed:

    "Call to <*dialed_ number*> failed: User not logged in on any device"

- An application interfaces with the WebDialer and Redirector servlets through HTTPS.
- When using Client Matter Codes (CMC) or Forced Authorization Codes (FAC), WebDialer users must enter the proper code at the tone by using the phone's keypad. Failure to enter the appropriate code at the tone will result in call failure signaled by a reorder tone.
- Cisco WebDialer is available on any Cisco endpoints that support Cisco Computer Telephony Integration (CTI). For a list of Cisco endpoints that support Cisco Computer Telephony Integration (CTI), refer to the *Cisco CTI Supported Device Matrix*, available at

  http://developer.cisco.com/web/jtapi/wikidocs/-/wiki/Main/Cisco+CTI+Supported+Device+Matrix

Attendant Consoles

Attendant console integrations enable a receptionist to answer and transfer or dispatch calls within an organization from a desktop application designed specifically for this purpose. Attendant consoles allow for access to the corporate directory and, in some cases, monitoring of line state for specific users. The Cisco Unified Communications portfolio provides the following types of Cisco Unified Attendant Consoles:

- Cisco Unified Attendant Console Department Edition
- Cisco Unified Attendant Console Business Edition
- Cisco Unified Attendant Console Enterprise Edition
- Cisco Unified Attendant Console Premium Edition

Cisco Unified Attendant Consoles have a client attendant console application that installs on an attendant's Windows PC. It also requires an attendant console server application installed on a separate physical server than Unified CM. The attendant console application communicates with the attendant console server application, and the attendant console server application communicates with Unified CM securely through CTI and AXL over Secure Socket Layer (SSL). Multiple attendant consoles can connect to a single attendant console server. The Department, Business, Enterprise, and Premium Editions differ in their limits to various capabilities such as the number of supported operator clients and the number of supported directory entries.

This section examines the following design aspects of the attendant consoles:

- Attendant Console Architecture, page 19-44
- High Availability for Attendant Consoles, page 19-45
- Capacity Planning for Attendant Consoles, page 19-46
- Design Considerations for Attendant Consoles, page 19-46

Attendant Console Architecture

Figure 19-18 illustrates the high-level architecture of a Cisco Unified Attendant Console integration. Understanding the functionality and operation of the solution enhances the understanding of the architecture itself. The following steps (denoted in Figure 19-18) detail the events involved for a typical call into an attendant console.

1. A call comes into Unified CM, and the called number matches the directory number configured on a CTI route point.
2. The CTI route point is CTI-controlled by the attendant console server application and is associated with a Queue Direct Dial In (DDI) configured on the server.
3. The attendant console server application immediately redirects the call internally to one of its Computer Telephony (CT) Gateway Devices. As part of this process, the attendant console server application sends a CTI redirect message to the CTI Manager service to redirect the call to a CTI port.

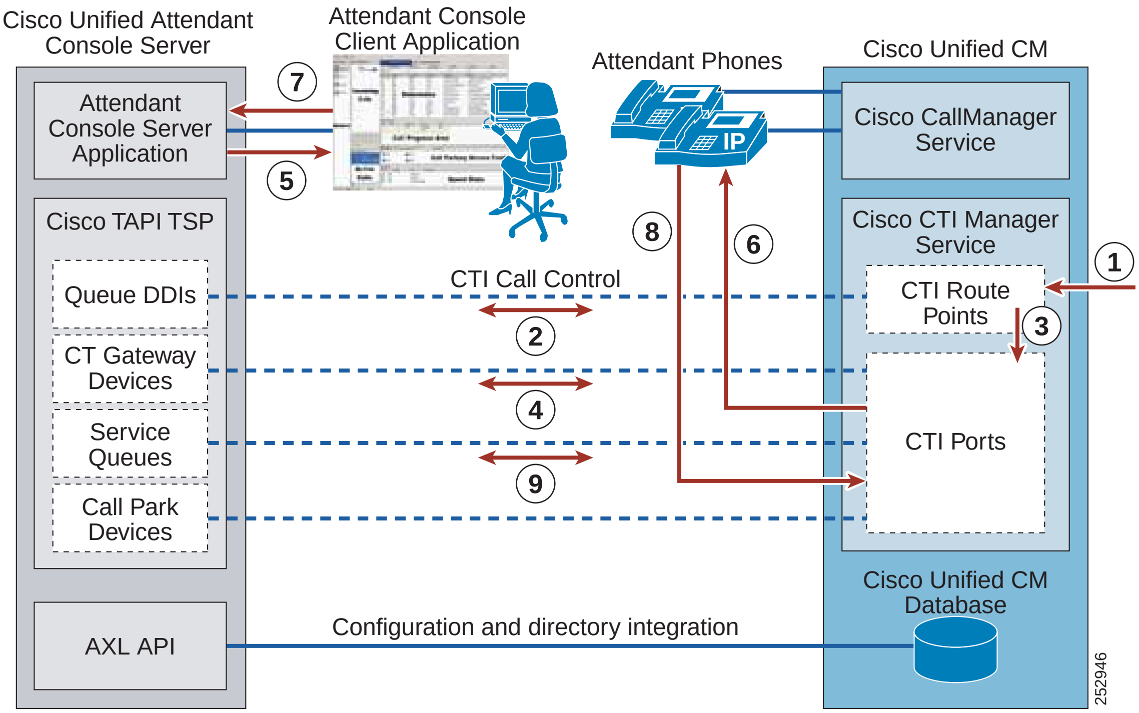
Note A CTI redirect message does not result in a connected call; the call is not answered and there is no media connection.

4. The attendant console server application now associates the call with the CT Gateway Device and controls the call on the CTI port.
5. At this point, the call is presented to the attendant console client applications in the system that are associated with the Queue DDI.
6. Once an attendant chooses to answer the call through the attendant console client application, another CTI redirect message is sent to the CTI Manager service, which moves the call from the CTI port to the answering attendant's physical phone. The call is automatically connected on the attendant's phone, either to the handset or the headset, depending on the phone configuration. The region and location settings of the attendant's phone and the initiating gateway or phone dictate the codec used for media.
7. When a transfer to another extension is required, the attendant initiates the transfer through the attendant console client application, which communicates the transfer to the attendant console server application.
8. The attendant console server application internally associates the call with a Service Queue and sends a CTI redirect message to the CTI Manager service. This redirects the call from the attendant's phone to a CTI port controlled by the attendant console server application.

Note A call transfer may also be initiated from the attendant's phone; however, this would remove the attendant console server application from the call flow, and enhanced functionality (such as the transfer recall feature) would no longer be possible.

9. At this stage, the Service Queue actually answers the call (there is a short connect) before issuing the transfer, therefore the Cisco Media driver installed on the attendant console server application is invoked. The region and location settings of this CTI port and the call-initiating gateway or phone dictate the codec used for media. The configured Music on Hold (MoH) audio sources of the CTI port also affect the MoH heard by the caller. Transfers are performed in this manner so that the attendant console client application still maintains control of the call if there is no answer. Once the call is received by the final party, the attendant console server application is removed from the call flow.

Figure 19-18 Architecture for Cisco Unified Department, Business, and Enterprise Attendant Consoles


Cisco Unified Attendant Console Server
Attendant Console Client Application
Attendant Phones
Cisco Unified CM
Attendant Console Server Application
7
5
Cisco CallManager Service
IP
Cisco TAPI TSP
8
6
Cisco CTI Manager Service
1
Queue DDIs
CTI Call Control
CTI Route Points
2
3
CT Gateway Devices
4
Service Queues
CTI Ports
9
Call Park Devices
Cisco Unified CM Database
AXL API
Configuration and directory integration
252946

The attendant console server application's call park function does not use the inherent call park feature of Unified CM. Instead, it uses its own call park facility using Call Park Devices. Call Park Devices work very much like the Service Queues as outlined in steps 7 to 9 of Figure 19-18. Similar to transfers, Call Park Devices allow the attendant console server application to maintain control of the call for the duration of the parked call.

High Availability for Attendant Consoles

Cisco Unified Attendant Console Premium Edition can be installed in a resilient configuration with two Cisco Unified Attendant Console servers:

- Publisher — The primary server used by the clients. If this server fails, all attendant operators are switched to the subscriber server. Once the publisher is running again, the operators are prompted to reconnect to the publisher.
- Subscriber — Used if the publisher stops running for any reason.

The Cisco Unified Attendant Console Department, Business, and Enterprise Editions are deployed with a single Cisco Unified Attendant Console server.

You should consider providing redundancy on both sides of the integration for both CTI and AXL communication.

Regarding CTI, the attendant console server application uses the Cisco TAPI Telephony Service Provider (TSP) plug-in (downloaded from Unified CM) to communicate with the CTI Manager service. Cisco TSP allows for the configuration of a primary and backup CTI Manager service. Cisco recommends enabling the CTI Manager service on at least two Unified CM subscriber nodes in the cluster to gain resilience in case the primary CTI Manager service goes offline. In the event of an attendant console server failure, resilience can be achieved by configuring a Call Forward Unregistered (CFU) and Call Forward CTI failure destination on all of the CTI route points associated with Queue DDIs. If the attendant console server application is offline, calls will automatically follow the Call Forward setting. For example, with the Cisco Unified Attendant Console Premium Edition, calls can be forwarded to the Cisco Unified Attendant Console subscriber server. With other Cisco Unified Attendant Console Editions, the destination could be a Hunt Pilot number or a Directory Number (DN) associated with a single IP phone.

AXL communication is enabled by activating the Cisco AXL Web Service on a Unified CM node. Multiple Unified CM nodes can have the Cisco AXL Web Service enabled, but the attendant console server application has only a single entry for Unified CM connectivity. In the event of a failure, an administrator could update this entry to a backup Unified CM node running the Cisco AXL Web Service. Cisco Unified Attendant Console Premium Edition has AXL resiliency.

The Unified CM also has a series of CTI route points and CTI ports configured for integration with Cisco Unified Attendant Console. These devices have a device pool and therefore are assigned a Unified CM group, which specifies a prioritized list of the Unified CM call processing nodes responsible for maintaining registration. When the primary Unified CM in the Unified CM group is offline, the CTI route points and CTI ports have the ability to register with a secondary Unified CM node, thus allowing for high availability of the CTI route points and ports themselves.

Capacity Planning for Attendant Consoles

For a comparison of the various Cisco Unified Attendant Console Department, Business, Enterprise, and Premium Editions and their respective capacities, refer to the *Cisco Unified Attendant Consoles Business/Department/Enterprise/Premium Edition Design Guide*, available at

http://www.cisco.com/en/US/products/ps7282/products_implementation_design_guides_list.html

To size a Unified CM cluster properly, your Cisco Partner or Cisco Systems Engineer should use the Cisco Unified Communications Sizing Tool (http://tools.cisco.com/cucst) to validate all designs that incorporate a large number of CTI resources and high call volumes, because there are many interdependent variables that can affect Unified CM cluster scalability. The Sizing Tool can accurately determine the number of servers or clusters required to meet your Attendant Console design criteria.

For performance and capacity information about the various Cisco Unified Attendant Console Editions, refer to the product documentation available at

http://www.cisco.com/en/US/products/ps7282/tsd_products_support_series_home.html

Design Considerations for Attendant Consoles

The following design guidelines and restrictions apply with regard to the deployment and operation of Cisco Unified Attendant Console within the Unified CM telephony environment.

- The following general design guidance applies to the attendant console server application components:
  - Queue DDI

    One unique Queue DDI is required for each unique incoming directory number in the system that should be routed specifically to the attendant consoles.
  - CT Gateway Device

  Every incoming call into a Queue DDI is immediately redirected to a CT Gateway Device. Design the system so that the number of CT Gateway Devices can handle the maximum expected number of incoming calls at any given time.

  - Service Queue

    Each time an attendant transfers a call or places a call on hold, a Service Queue is required. The system should be designed so that there are enough Service Queues to sustain the maximum number of calls that all attendants in the system are in the process of transferring or putting on hold at any given time. A general guideline is to provide 3 or 4 Service Queues per attendant, but some scenarios might require more.

  - Call Park Device

    Each time an attendant invokes the Call Park feature through the attendant console client application, a Call Park Device is required. This feature does not use the inherent Call Park capability of Unified CM. Design the system so that there are sufficient Call Park Devices to handle the maximum number of calls parked by all attendants in the system at any given time.

- Every Queue DDI, CT Gateway Device, Service Queue, and Call Park Device configured in the attendant console server application creates a CTI route point or CTI port in Unified CM. The number of CTI connections required to handle the Unified Department, Business, or Enterprise Attendant Console integration also counts toward the CTI connection limits per cluster. (For more information on CTI connection limits per cluster, see Capacity Planning for CTI, page 8-34.)

- The attendant console server application provides busy lamp field (BLF) monitoring of end-user devices, but it is important to note that this does not use the same facility in Unified CM that provides BLF speed dial capability. Instead, the attendant console server application communicates through CTI with Unified CM to obtain line state information on monitored devices. Once the attendant console server application monitors an end-user device, it continues monitoring this device through CTI until the number of devices monitored for BLF reaches a certain level (2,000). Once this limit is reached, the BLF plug-in begins to drop devices from the list of monitored devices in order to add newly requested devices to the list, thus ensuring that the number of devices monitored by the attendant console server through CTI does not exceed the limit (2,000). These devices monitored through CTI also count toward the CTI limits in Unified CM.

- The attendant console server application provides Busy Lamp Field (BLF) monitoring of end-user devices, but it is important to note that this does not use the same facility in Unified CM that provides BLF speed dial capability. Instead, the attendant console server application communicates through CTI with Unified CM to obtain line state information on monitored devices.

- With respect to Quality of Service (QoS), the attendant console server application, the attendant console client application, and the Cisco TSP all send their traffic marked as Best Effort (DSCP=0). If this traffic traverses a WAN or a link that is typically congested, packets must be marked to receive preferential treatment through the network. For a complete list of the TCP port numbers associated with these applications, refer to the Unified Department, Business, or Enterprise Attendant Console design guide, available with appropriate login authentication at

  http://www.cisco.com/go/ac

- Cisco TSP is not aware of partitions. Therefore, if the same directory number (DN) exists in multiple partitions, the monitored device might not be the correct DN.

- Cisco Unified Attendant Console can also integrate with the Cisco IM and Presence Service through the SIP SIMPLE protocol. For more information about this type of integration, refer to the appropriate Cisco Unified Attendant Console administration guide, available at

  http://www.cisco.com/en/US/products/ps7282/prod_maintenance_guides_list.html

- For design guidance on Cisco Unified Attendant Consoles, refer to the documentation available at

  http://www.cisco.com/en/US/products/ps7282/products_implementation_design_guides_list.html

PART 4

Unified Communications Applications and Services

CHAPTER 20

Overview of Cisco Unified Communications Applications and Services

Revised: February 29, 2012; OL-27282-05

Once the network, call routing, and call control infrastructure has been put in place for your Cisco Unified Communications System, additional applications and services can be added or layered on top of that infrastructure.   There are numerous applications and services that can be deployed on an existing Cisco Unified Communications infrastructure, and the following applications and services are typically deployed:

- Voice messaging — Provides voicemail services and message waiting indication.
- Rich media conferencing — Provides audio and video conferencing as well as web-based application and document sharing.
- Presence services — Provide user availability tracking across user devices and clients.
- Mobility services — Provide enterprise-level unified communications features and functionality to users outside the enterprise.
- Contact center — Provides call handling, queuing, and monitoring for large call volumes.
- Collaboration client services — Provide integration to unified communications services and leveraging of various applications.

The chapters in this part of the SRND cover the applications and services mentioned above. Each chapter provides an introduction to the application or service, followed by discussions surrounding architecture, high availability, capacity planning, and design considerations. The chapters focus on design-related aspects of the applications and services rather than product-specific support and configuration information, which is covered in the related product documentation.

This part of the SRND includes the following chapters:

- Cisco Voice Messaging, page 21-1

  This chapter examines voice messaging, a common and prevalent application within most unified communications deployments, which allows callers to send messages and subscribers of the system to retrieve messages. The chapter examines messaging deployment models, voice messaging features and functionality, voicemail networking, and design and deployment best practices for voice messaging applications.

- Cisco Collaborative Conferencing, page 22-1

  This chapter explores rich media conferencing, which allows users of the unified communications system to schedule, manage, and attend audio, video, and/or web collaboration conferences. The chapter considers various aspects of rich media conferencing, including components, deployment models, video capabilities, H.323 and SIP call control integrations, capacity and redundancy, and various solution recommendations and design best practices.

- Cisco IM and Presence, page 23-1

  This chapter discusses presence services, an increasingly critical piece of most unified communications deployments due to the productivity improvements that can be realized from user availability-based applications. This chapter defines presence and explores the various presence components and features, protocols, deployment models, redundancy, capacity, and general design guidelines.

- Cisco Collaboration Clients and Applications, page 24-1

  This chapter covers collaboration clients and applications, which are quickly closing the gap between traditional hardware-based phones and feature-rich PC-based clients. This chapter explores the various collaboration clients, their features, and the various integration methods, as well as integrations with various third-party collaboration applications.

- Mobile Unified Communications, page 25-1

  This chapter looks at mobility applications, which are becoming extremely important given the growth of mobile work forces and the blurring of enterprise boundaries for unified communications features and services, resulting in an increased demand for mobility applications and services. This chapter discusses mobility solution architectures, functionality, and design and deployment implications.

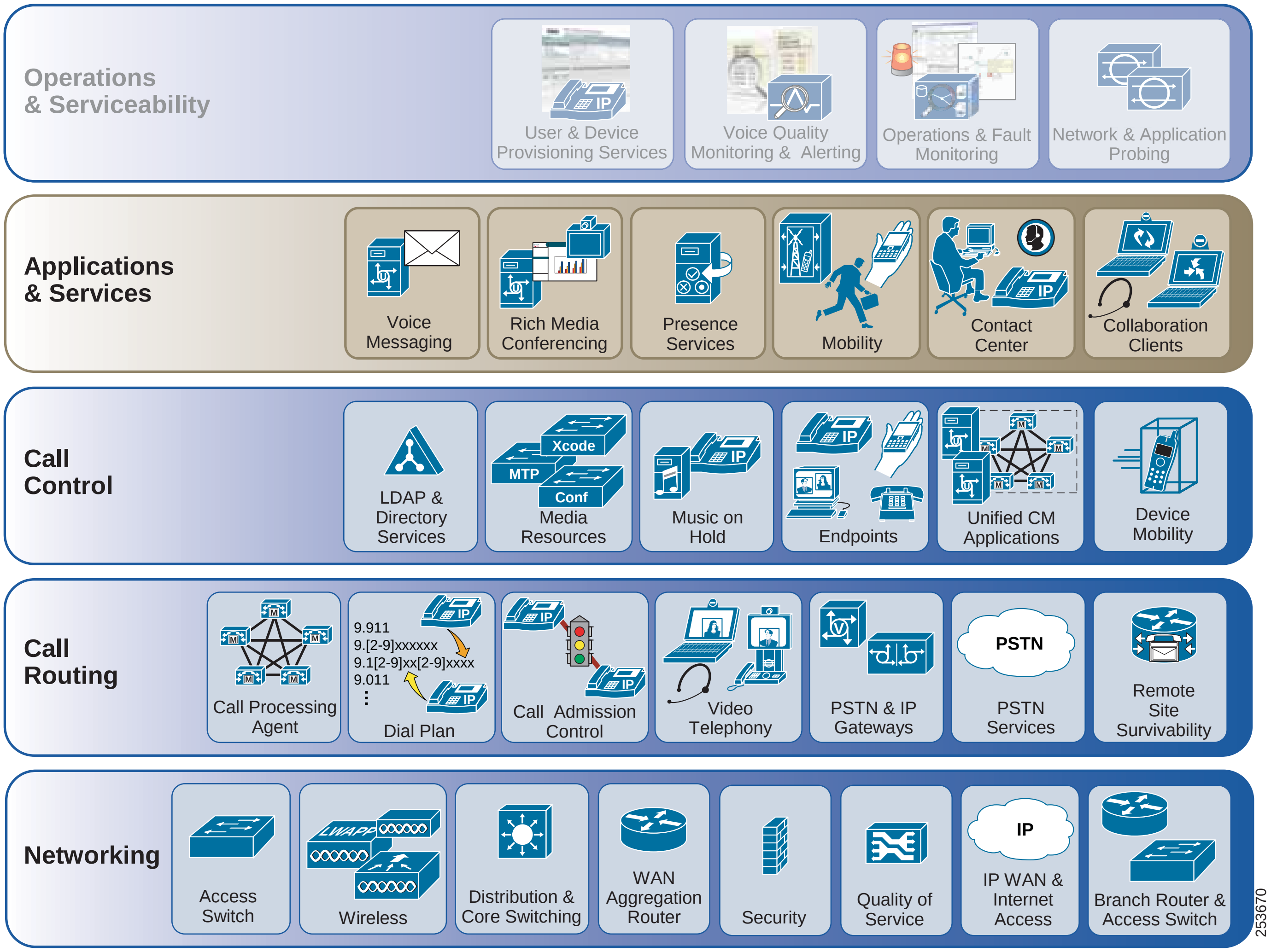
- Cisco Unified Contact Center, page 26-1

  This chapter covers contact center solutions, an important and integral part of large unified communications deployments requiring high-volume call center applications. This chapter examines call center solution architectures, functionality, and design and deployment implications.

Architecture

As with other network and application technology systems, unified communications applications and services must be layered on top of the underlying network and system infrastructures. Figure 20-1 shows the logical location of unified communications applications and services in the overall Cisco Unified Communications System architecture.

Figure 20-1 Cisco Unified Communications Applications and Services Architecture



Unified communications applications and services such as voice messaging, rich media conferencing, presence, mobility, contact center, and collaboration clients rely on the underlying unified communications call routing and call control infrastructure and network infrastructure for everything from network connectivity to basic unified communications functions such as call control, supplementary services, dial plan, call admission control, and gateway services. For example, voice messaging and rich media conferencing applications leverage the network infrastructure for reaching users in campus sites, in branch sites, and on the Internet. Further, these same applications depend on the unified communications voice and video endpoints, call routing, PSTN connectivity, and media resources provided by the call routing and control infrastructure. In addition to relying on these infrastructure layers and basic unified communications services, applications and services are also often dependent upon each other for full functionality.

High Availability

As with network, call routing, and call control infrastructures, critical unified communications applications and services should be made highly available to ensure that required features and functionality remain available if failures occur in the network or applications. It is important to understand the various types of failures that can occur and the design considerations around those failures. In some cases, the failure of a single server or feature can impact multiple services because many unified communications applications are dependent on other applications or services. For example, while the various application service components of a contact center deployment might be functioning properly, the loss of all call control servers would effectively render the contact center unusable because the deployment is dependent upon the call control servers to route calls to the call center applications.

For applications and services such as voice messaging and rich media conferencing, high availability considerations include temporary loss of functionality due to network connectivity or application server failures resulting in the inability of callers to leave messages, of users to retrieve messages, and of users to schedule or attend conferences. In addition, failover considerations for callers and users of voice messaging and rich media conferencing applications include scenarios in which portions of the functionality can be handled by a redundant resource that allows end users to continue to access services in the event of certain failures.

High availability considerations are also a concern for services such as presence and mobility. Interrupted network connectivity or server failures will typically result in reduced functionality or, in some case, complete loss of functionality. For presence services, this can mean that some or all devices and clients will be unable to send or receive presence or availability updates. For mobility services, high availability considerations include the potential for loss of specific functionality such as two-stage dialing or dial-via-office, or reduced functionality for features such as single number reach (resulting in situations where only the enterprise phone rings or only the mobile phone rings). Further, in some failure scenarios, enterprise phones and mobile clients might have to reregister, reconnect and/or re-authenticate before full functionality is available again.

For contact center deployments, there are numerous servers and components for which high availability must be considered. Typically, an isolated single-server or single-component failure can be handled without loss of features or functionality as long as the server or component has been made redundant. In other situations, loss of multiple servers or components will typically result in loss of some features or functionality. However, in scenarios where there is complete loss of a particular component such as all call control servers, more catastrophic loss of features or functionality is possible.

When considering collaboration clients and applications, high availability is certainly important. Not only can specific collaboration features or functions become unavailable in failure scenarios, but in some cases presence-capable clients might be unable to connect to the network for even basic functionality such as registration and making or receiving calls. In other cases, clients or devices might have to reconnect and re-authenticate in order to return to service.

Capacity Planning

Network, call routing, and call control infrastructures must be designed and deployed with an understanding of the capacity and scalability of the individual components and the overall system. Similarly, deployments of unified communications applications and services must also be designed with attention to capacity and scalability considerations. When deploying various unified communications applications, not only is it important to consider the scalability of the applications themselves, but you must also consider the scalability of the underlying infrastructures. Certainly the network infrastructure must have available bandwidth and be capable of handling the additional traffic load the applications will create. Likewise, the call routing and control infrastructure must be capable of handling user and device configuration and registration as well as application integration loads surrounding protocols and connections. For example, with applications and services such as mobility, presence, and contact center, there are capacity implications for each of these individual applications in terms of users, devices, and features, but just as important is the scalability of the underlying infrastructure to handle connections and protocols such as Computer Telephony Integration (CTI). While a mobility, presence, or contact center application may be able to support many CTI connections, the underlying call control and routing infrastructure might not have available capacity to handle the added CTI load of these application and services.

For applications and services such as voice messaging and rich media conferencing, capacity planning considerations include things like number of mailboxes or users, mailbox size, audio and video ports, and MCU sessions. In most cases additional capacity can be added by increasing the number of application servers and MCUs or by upgrading server or MCU hardware with higher-scale models, assuming the underlying network and call routing and control infrastructures are capable of handling the additional load.

Capacity planning considerations are also a concern for services such as presence and mobility. Scalability must be contemplated not only for things like numbers of configured and supported users and devices, but also for the number of integrations and connections between those applications and others. The volume of two-stage dialing and dial-via-office calls is of particular concern for mobility applications from the perspective of both the call control capacity and the PSTN gateway capacity. With presence services, on the other hand, critical scalability concerns include frequency of presence status changes and the propagation of these changes to the network, as well as text or instant message volumes. Typically, additional application servers or hardware upgrades will result in increased capacity for these applications and services, but the underlying call routing and control infrastructures must be capable of handling any increases in load.

Contact center deployments are no different than other applications and services in terms of scalability concerns. Certainly the number of agents and agent devices handling calls is important in terms of user and device configuration and registration. However, the major concerns in terms capacity for contact center deployments are the high number of busy hour call attempts (BHCA) common in contact centers and the number of CTI integrations to the call control and routing infrastructure.

When considering collaboration clients and application capacity planning, device registration and configuration are the most important scalability concerns. However, certainly there are other scalability implications in terms of the back-end applications and services such as presence and messaging. Further, when deploying or integrating various clients with third-party applications and infrastructures, you must also consider the supported capacities for those third-party deployments.

For a complete discussion of system sizing, capacity planning, and deployment considerations related to sizing, refer to the chapter on Unified Communications Design and Deployment Sizing Considerations, page 29-1.

CHAPTER 21

Cisco Voice Messaging

Revised: April 30, 2013; OL-27282-05

This chapter describes the voice messaging solutions available in the Cisco Unified Communications System. It includes the Cisco voice messaging products Cisco Unity Connection and Cisco Unity Express, and it covers the design guidelines and best practices for deploying these products together with Cisco Unified Communications Manager (Unified CM). This chapter also covers aspects of integration with third-party voicemail systems using industry standard protocols.

Although this guide focuses on the messaging deployment scenarios with regard to Unified CM, Cisco Unified Communications Manager Express (Unified CME) is also noted where applicable, especially when used with Survivable Remote Site Telephony (SRST) fallback support in a centralized Unified CM deployment.

This chapter covers the following topics:

The chapter begins with a short description of each of the products in the Cisco messaging solutions portfolio and provides a simple overview of where each product fits in an enterprise Unified Communications solution. Next, messaging deployment models form the basis of discussion for voicemail integrations, which start with a definition of the various messaging deployment models and then explain how each of the messaging deployment models fits into the various Unified CM call processing deployment models. Cisco Unity Connection is discussed in this section, while Cisco Unity Express has a dedicated section for its supported deployment models. Key design guidelines are covered for interoperability available within the Cisco Voice Messaging product portfolio. Virtualization, a new concept, is covered along with the important design factors to be considered while designing the virtual system. Many system-level design considerations and best practices, including transcoding and various integrations with Cisco Unified Communications Manager, are explained in this section. In addition, this chapter provides details on third-party voicemail integration for supported industry-standard protocols.

This chapter presents a high-level design discussion and is focused on how the voice messaging products fit into the Unified Communications System with Unified CM. For detailed design guidelines for each product as well as interoperability information for third-party messaging and telephony systems, refer to the Cisco Unity Connection design guides, available at

http://www.cisco.com/en/US/products/ps6509/products_implementation_design_guides_list.html

What's New in This Chapter

Table 21-1 lists the topics that are new in this chapter or that have changed significantly from previous releases of this document.

Table 21-1 New or Changed Information Since the Previous Release of This Document

New or Revised Topic	Described in:	Revision Date
A few minor updates	Various sections	April 30, 2013
Various updates for Cisco Unified Communications System Release 9.0	Numerous sections throughout this chapter	June 28, 2012

Voice Messaging Portfolio

The Cisco Unified Communications messaging portfolio consists of two main messaging products: Cisco Unity Connection and Cisco Unity Express. Each product fits different requirements yet each one contains overlapping features and scalability with regard to the others. They also have the ability to interwork with one another using Voice Mail Networking to achieve voicemail interoperability as well as higher scalability, as discussed later in this chapter.

When considering these products, it helps to think of the messaging types that the products apply to in order to understand the messaging options they include and to determine which options could fit your deployment requirements. The following definitions help describe these messaging types:

- *Voicemail-only* refers to a telephony voicemail integration where there is no access to the voicemail via any messaging client.
- *Integrated messaging* refers to voicemail with telephony access as well as voicemail-only access via a messaging client.
- *Unified messaging* refers to voicemail with telephony access as well as voicemail, email, and fax access via a messaging client.

Table 21-2 shows which Cisco products support these types of messaging.

Table 21-2 Supported Messaging Environments per Product

Messaging Type	Cisco Unity Connection	Cisco Unity Express
Voicemail-only	Yes	Yes
Integrated messaging	Yes	Yes
Unified messaging	Yes	No

Note For further details on Unified Messaging with Cisco Unity Connection, see Single Inbox with Cisco Unity Connection, page 21-40.

Based on the above messaging types and definitions, the two messaging product options are:

- Cisco Unity Connection

  This option combines unified/integrated messaging, voice recognition, and call transfer rules into an easy-to-manage system for medium-sized businesses with up to 20,000 users, or it can network up to 10 nodes in a digital network system. (Additionally, if required, a maximum of two digital networks can be joined to support a maximum of 20 nodes.) Cisco Unity Connection can support up to 100,000 users or contacts in a digital network. For organizations with up to 500 users, Cisco Unity Connection is available as a single-server solution with Cisco Business Edition 3000 and 5000. For more information on Cisco Business Edition, see Design Considerations for Call Processing, page 8-28.

- Cisco Unity Express

  This option provides cost-effective voice and integrated messaging, automated attendant, and interactive voice response (IVR) capabilities in certain Cisco Integrated Services Routers for small and medium-sized businesses and enterprise branch offices with up to 500 users.

For a complete comparison of product feature, refer to the *Cisco Messaging Products: Feature Comparison*, available at http://www.cisco.com/en/US/products/sw/voicesw/ps2237/products_data_sheets_list.html.

Table 21-3 gives a brief product comparison with regard to scalability.

Table 21-3 Scalability of Voice Messaging Solutions

	Users supported on a single server (or failover or clustered deployment)			**Maximum number of users supported in a digital networking solution**	
Solutions	**500**	**15,000**	**20,000**	**100,000**	**250,000**
Cisco Unity Express	Y	N	N	Y	Y
Cisco Business Edition	Y	N	N	N	N
Cisco Unity Connection (Unified/Integrated Messaging)	Y	Y	Y	Y	N

Note In addition to providing an increase in the maximum number of supported users, Digital Networking provides server discovery and directory synchronization functionality.

This chapter focuses on the design aspects of integrating Cisco Unity Connection and Cisco Unity Express with Cisco Unified Communications Manager (Unified CM). Cisco Unified CM provides functionality for Session Initiation Protocol (SIP) trunks, which support integration directly to Cisco Unity Connection without the need for a SIP proxy server.

For information on earlier releases of Cisco Unity Connection, Unity Express, and Unified CM or Unified CM Express, refer to the appropriate online product documentation available at http://www.cisco.com.

As mentioned, the design topics covered in this chapter apply to voicemail-only, unified messaging, and integrated messaging configurations. Additionally, this chapter discusses design aspects of deploying Cisco Unity Connection with Microsoft Exchange (2003, 2007, or 2010). Cisco Unity Connection and Unity Express have no dependencies on an external message store.

For additional design information about Cisco Unity Connection, including integrations with other non-Cisco messaging systems, refer to the *Design Guide for Cisco Unity Connection*, available at http://www.cisco.com.

For additional design information about Cisco Unity Express, including integrations with other non-Cisco messaging systems, refer to the applicable product documentation, available at http://www.cisco.com.

Messaging Deployment Models

This section summarizes the various messaging deployment models for Cisco Unity Connection and Cisco Unity Express. For a complete discussion of the deployment models and design considerations specific to Cisco Unity Connection and the various messaging components, refer to the *Design Guide for Cisco Unity Connection*, available at http://www.cisco.com. For Cisco Unity Express, refer to the applicable product documentation available at http://www.cisco.com.

Cisco Unity Connection supports three primary messaging deployment models:

- Single-site messaging
- Multisite deployment with centralized messaging
- Multisite deployment with distributed messaging

Cisco Unity Express also supports three primary messaging deployment models:

- Single-site messaging
- Multisite deployment with distributed messaging
- Multisite deployment with distributed messaging with Cisco Unified CME

Note The Cisco Unity Express supports centralized voice messaging for up to 10 Unified CMEs. For more information, refer to the Cisco Unified Communications Manager Express documentation on http://www.cisco.com.

Although the call processing deployment models for Cisco Unified CM and Unified CME are independent of the messaging deployment models for Cisco Unity Connection and Unity Express, each has implications toward the other that must be considered.

Cisco Unity Connection messaging redundancy is available in an active/active configuration. For more information, refer to the *Design Guide for Cisco Unity Connection* available on http://www.cisco.com.

All messaging deployment models support voicemail, integrated messaging, and unified messaging installations.

Single-Site Messaging

In this model, the messaging systems and messaging infrastructure components are all located at the same site, on the same highly available LAN. The site can be either a single site or a campus site interconnected via high-speed metropolitan area networks (MANs). All clients of the messaging system are also located at the single (or campus) site. The key distinguishing feature of this model is that there are no remote clients.

Centralized Messaging

In this model, similar to the single-site model, all the messaging system and messaging infrastructure components are located at the same site. The site can be one physical site or a campus site interconnected via high-speed MANs. However, unlike the single-site model, centralized messaging clients can be located both locally and remotely.

Distributed Messaging

A distributed messaging model consists of multiple single-site messaging systems distributed with a common messaging backbone. There can be multiple locations, each with its own messaging system and messaging infrastructure components. All client access is local to each messaging system, and the messaging systems share a messaging backbone that spans all locations. Message delivery from the distributed messaging systems occurs via the messaging backbone through a full-mesh or hub-and-spoke type of message routing infrastructure. No messaging infrastructure components should be separated by a WAN from the messaging system they service.

Distributed messaging is essentially multiple, single-site messaging models with a common messaging backbone. The exception to this rule is the PBX-IP Media Gateway (PIMG) and T1-IP Media Gateway (TIMG) integrations. PIMG and TIMG integrations are not discussed in this design document. For further information regarding PIMG or TIMG, refer to the Cisco Unity Connection integration guides available on http://www.cisco.com.

The distributed messaging model has the same design criteria as centralized messaging with regard to local and remote GUI clients, TRaP, and message downloads.

Messaging and Unified CM Deployment Model Combinations

This section discusses the design considerations for integrating the various messaging deployment models with the Unified CM call processing deployment models. Table 21-4 lists the various combinations of messaging and call processing deployment models supported by Cisco Unity Connection and Unity Express.

Table 21-4 Supported Combinations of Messaging and Unified CM Call Processing Deployment Models

Model Type	Cisco Unity Connection	Cisco Unity Express
Single-site messaging and single-site call processing	Yes	Yes
Centralized messaging and centralized call processing	Yes	No[1]
Distributed messaging and centralized call processing	Yes	Yes
Centralized messaging and distributed call processing	Yes	No[1]
Distributed messaging and distributed call processing	Yes	Yes
Centralized messaging with cluster over the WAN	Yes	No
Distributed messaging with cluster over the WAN	Yes	Yes

[1] Support for centralized voicemail messaging with Unified CME is available starting with Cisco Unity Express 3.2; however, this is not applicable to Unified CM call processing deployment models.

This section covers the following topics:

- Cisco Unity Connection messaging and Unified CM deployment models
- Cisco Unity Express deployment models

Each topic defines a messaging and Unified CM deployment model combination and then highlights each Cisco voicemail messaging product applicable to that model as well as the design considerations for that model combination. Not all combinations are discussed for each product. Some examples are provided, with best practices and design considerations for each product. The intention is to provide an understanding of the base messaging deployment models and the interaction with Unified CM without detailing all possibilities.

Refer to the *Design Guide for Cisco Unity Connection*, available at http://www.cisco.com, for further details on site classification and a detailed analysis of supported combinations of messaging and call processing deployment models.

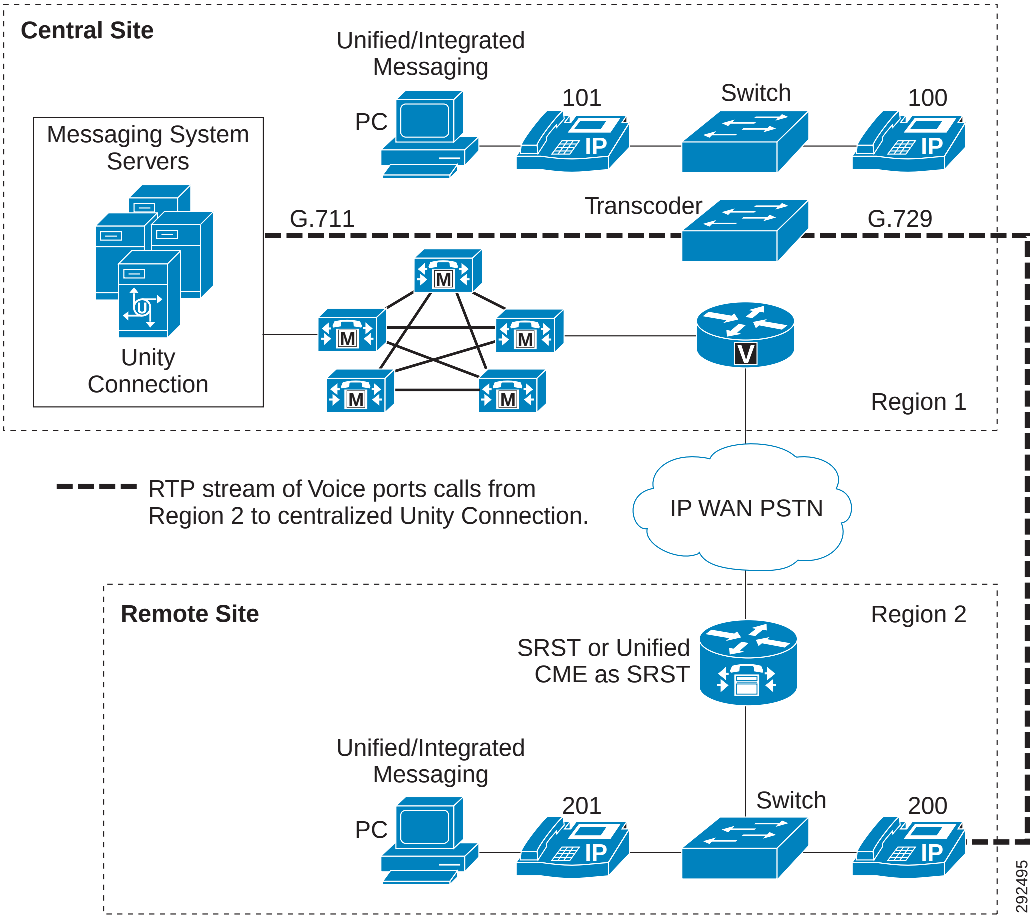
Cisco Unity Connection Messaging and Unified CM Deployment Models

This section discusses some of the various combinations of messaging and call processing deployment models for Cisco Unity Connection.

Centralized Messaging and Centralized Call Processing

In centralized messaging, the voice messaging server is located in the same site as the Unified CM cluster. With centralized call processing, subscribers may be located either remotely and/or locally to the cluster and messaging server(s). (See Figure 21-1.) When remote users access resources at the central site (such as voice ports, IP phones, or PSTN gateways, as in Tail-End Hop-Off (TEHO)), these calls are transparent to gatekeeper call admission control. Therefore, regions and locations must be configured in Unified CM for call admission control. (See Managing Bandwidth, page 21-28.) When making inter-region calls to IP phones or MGCP gateways, IP phones automatically select the inter-region codec that has been configured.

Figure 21-1 Centralized Messaging with Centralized Call Processing


Central Site
Unified/Integrated Messaging
PC
101
Switch
100
Messaging System Servers
G.711
Transcoder
G.729
M
V
Unity Connection
Region 1
RTP stream of Voice ports calls from Region 2 to centralized Unity Connection.
IP WAN PSTN
Remote Site
Region 2
SRST or Unified CME as SRST
Unified/Integrated Messaging
PC
201
Switch
200
IP
292495

In Figure 21-1, regions 1 and 2 are configured to use G.711 for intra-region calls and G.729 for inter-region calls.

As Figure 21-1 shows, when a call is made from extension 200 to the voicemail ports in Region 1, the inter-region G.729 codec is used at the endpoint but the RTP stream is transcoded to use G.711 on the voice ports. Unified CM transcoding resources must be located at the same site as the voicemail system.

Impact of Non-Delivery of RDNIS on Voicemail Calls Routed by AAR

In centralized messaging environments, automated alternate routing (AAR), a Unified CM feature, can route calls over the PSTN to the messaging store at the central site when the WAN is oversubscribed. However, when calls are rerouted over the PSTN, Redirected Dialed Number Information Service (RDNIS) can be affected. Incorrect RDNIS information can impact voicemail calls that are rerouted over the PSTN by AAR when Cisco Unity Connection is remote from its messaging clients. If the RDNIS information is not correct, the call will not reach the voicemail box of the dialed user but will instead receive the auto-attendant prompt, and the caller might be asked to re-enter the extension number of the party they wish to reach. This behavior is primarily an issue when the telephone carrier is unable to ensure RDNIS across the network. There are numerous reasons why the carrier might not be able to ensure that RDNIS is properly sent. Check with your carrier to determine if they provide guaranteed RDNIS deliver end-to-end for your circuits. The alternative to using AAR for oversubscribed WANs is simply to let callers hear reorder tone in an oversubscribed condition.

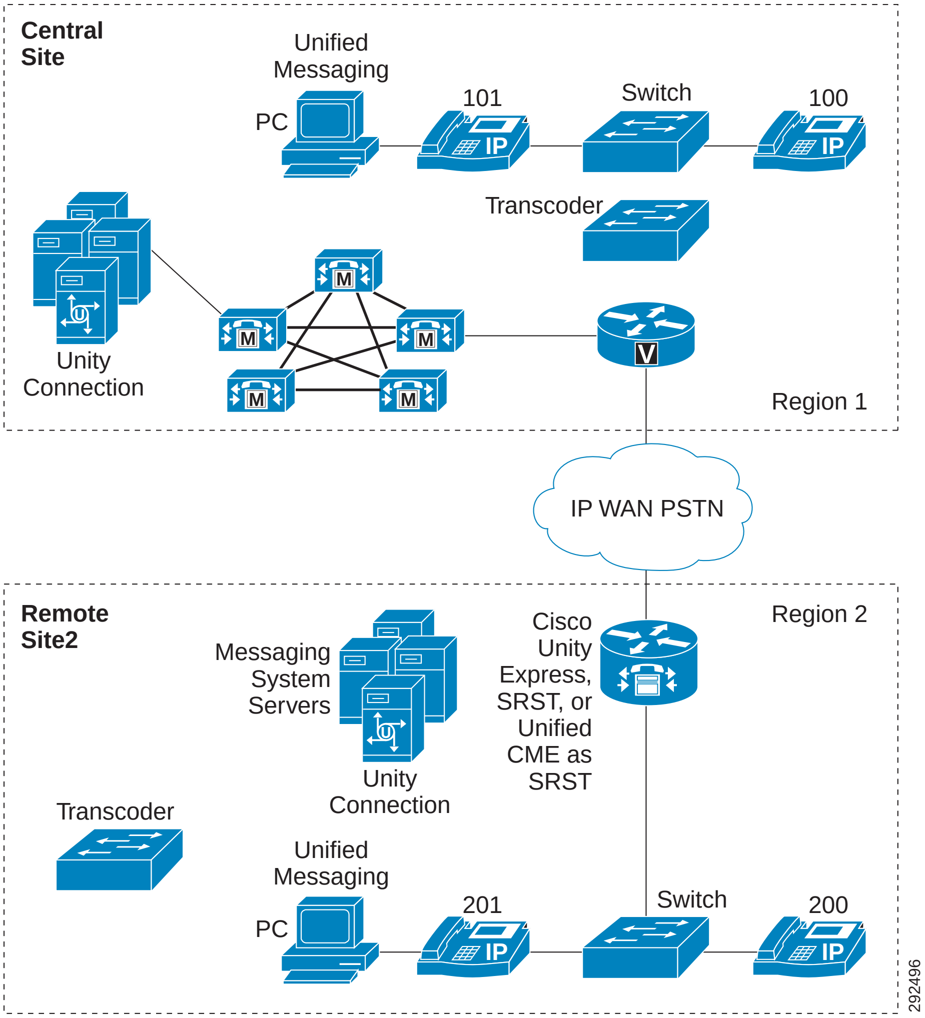
Distributed Messaging with Centralized Call Processing

Distributed messaging means that there are multiple messaging systems distributed within the telephony environment, and each messaging system services only local messaging clients. This model differs from centralized messaging, where clients are both local and remote from the messaging system.

Figure 21-2 illustrates the distributed messaging model with centralized call processing. As with other multisite call processing models, the use of regions and locations is required to manage WAN bandwidth.

Note that Cisco Unified Communications Manager Express (Unified CME) in SRST mode is used for call processing backup of both IP phones and Cisco Unity Connection voicemail ports. Deployed at the remote site (for example, Region 2 in Figure 21-2), this fallback support provides backup call processing in the event that the phones lose connectivity with Unified CM, such as during a WAN failure, while simultaneously providing users at the remote site with access to the local Cisco Unity Connection server as well as MWI support during WAN failure. For further details on Unified CME in SRST mode, refer to the Unified CME product documentation on http://www.cisco.com.

Figure 21-2 Distributed Messaging with Centralized Call Processing



For the configuration in Figure 21-2, transcoder resources must be local to each Cisco Unity Connection message system site. Regions 1 and 2 are configured to use G.711 for intra-region calls and G.729 for inter-region calls.

Voice messaging ports for both Cisco Unity Connection servers must be assigned the appropriate region and location by means of calling search spaces and device pools configured on the Unified CM server. In addition, to associate telephony users with a specific group of voicemail ports, you must configure Unified CM voicemail profiles. For details on configuring calling search spaces, device pools, and voicemail profiles, refer to the applicable version of the *Cisco Unified Communications Manager Administration Guide*, available at http://www.cisco.com.

Cisco Unity Connection supports digital networking, allowing multiple Cisco Unity Connection systems to be networked together. Up to 10 nodes (single or active/active pair) can be connected together in a digital network system. Additionally, if required, two digital networks can be joined to support a maximum of 20 nodes. This provides support for up to 100,000 entities in the directory. Cisco Unity Connection can integrate with a corporate directory such as Microsoft Active Directory to synchronize users and use digital networking simultaneously. In this configuration, each Cisco Unity Connection server or server pair will be able to synchronize up to 20,000 users from the corporate directory. Refer to the *Design Guide for Cisco Unity Connection*, available at http://www.cisco.com, for further information regarding digital networking or directory integration in Cisco Unity Connection.

Cisco Unity Connection with Unified CME in SRST Mode

Unified CME in SRST mode offers the possibility for Cisco Unity Connection servers located in remote sites and registered with a Unified CM at the central site to fall-back to Unified CME in the remote location. When the WAN link is down and the phones fail-over to Unified CME in SRST mode, Cisco Unity Connection voicemail ports can also fail-over to Unified CME in SRST mode to provide the remote site users with access to their voicemail with MWI during the WAN outage.

This scenario requires the following:

- Cisco Unified CME 4.0 or later
- Cisco Unity Connection 2.*x* or later

Note MWI has to be resynchronized from the Cisco Unity Connection server whenever a failover happens from Unified CM to Unified CME in SRST mode, or vice versa.

Combined Messaging Deployment Models

It is possible to combine messaging models in the same deployment, provided that the deployment adheres to all the guidelines listed in the preceding sections. Figure 21-3 shows a user environment in which both centralized and distributed messaging are employed simultaneously.

Figure 21-3 Combined Deployment Models

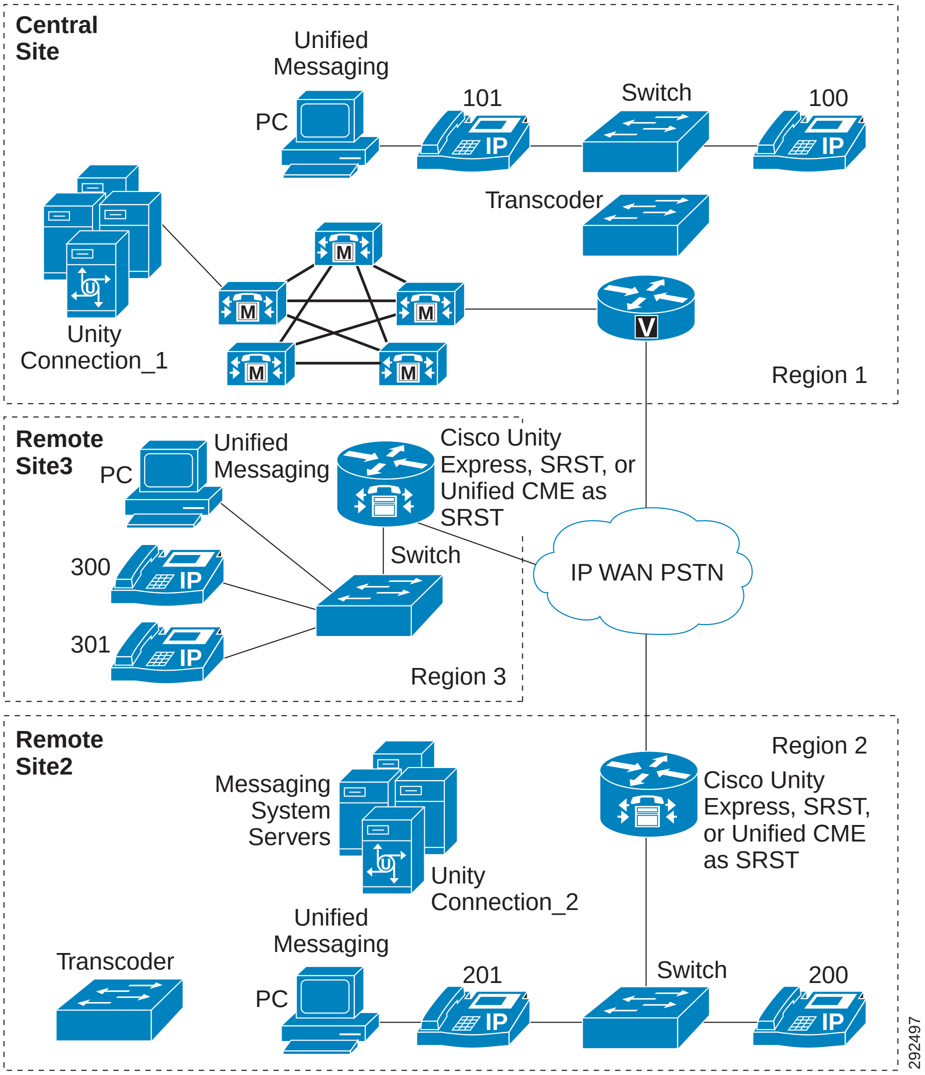


Figure 21-3 shows the combination of two messaging models. Regions 1 and 3 use centralized messaging with centralized call processing, while Region 2 uses distributed messaging with centralized call processing. All regions are configured to use G.711 for intra-region calls and G.729 for inter-region calls.

In Figure 21-3, centralized messaging and centralized call signaling are used between the Central Site and Site3. The messaging system at the Central Site provides messaging services for clients at both the Central Site and Site3. Site2 uses the distributed messaging model with centralized call processing. The messaging system (Unity Connection 2) located at Site2 provides messaging services for only those users located within Site2. In this deployment, both models adhere to their respective design guidelines as presented in this chapter. Transcoding resources are located locally to each messaging system site, and they support clients who access messaging services from a remote site (relative to the messaging system), as in the case of a Site2 user leaving a message for a Central Site user.

In addition, Cisco Unified Communications Manager Express (Unified CME) in SRST mode is used for call processing backup of both IP phones and Cisco Unity Connection voicemail ports. Deployed at the remote site (for example, Region 2 in Figure 21-3), this fallback support provides backup call processing in the event that the phones lose connectivity with Unified CM, such as during a WAN failure, while simultaneously providing users at the remote site with access to the local Cisco Unity Connection server as well as MWI support during WAN failure. For further details on Unified CME in SRST mode, refer to the product documentation on http://www.cisco.com.

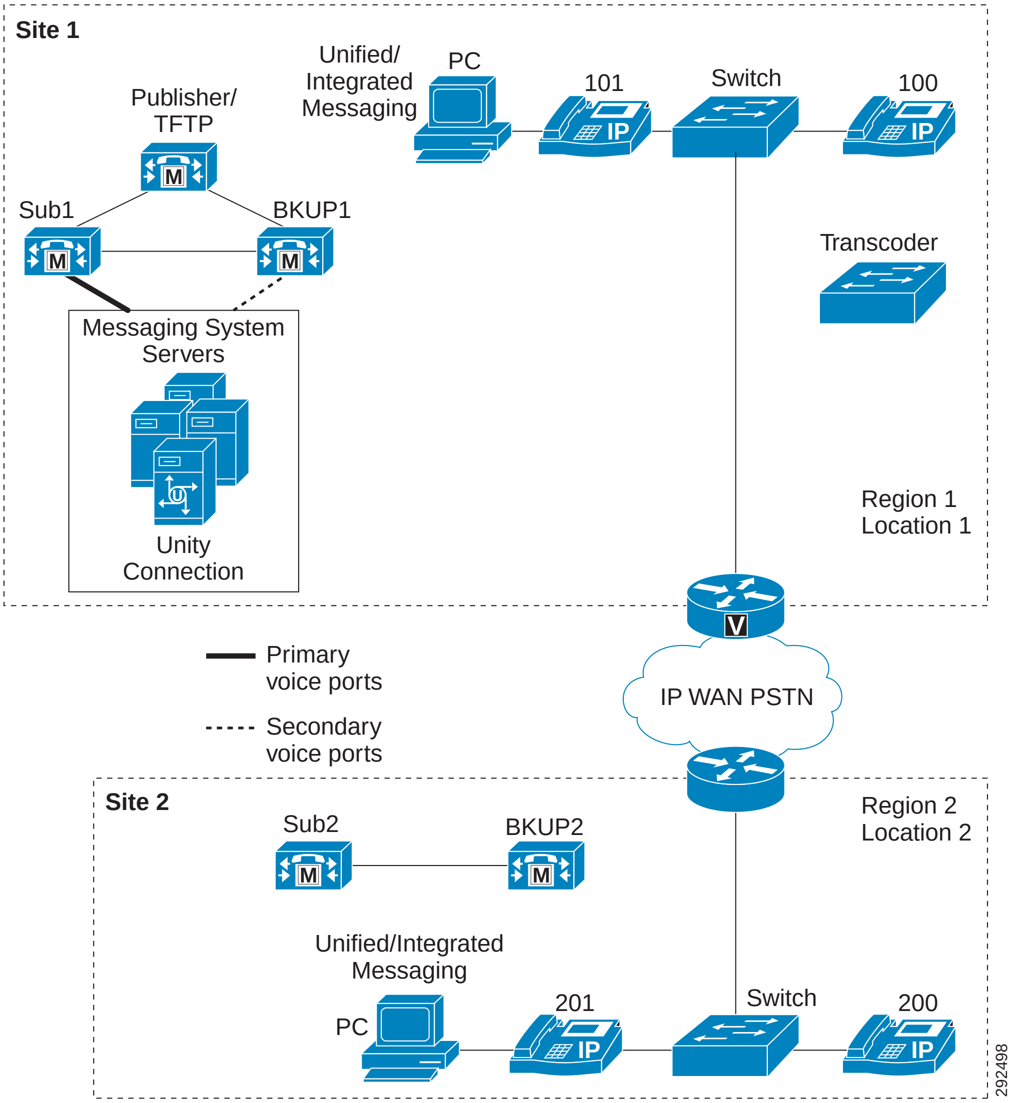
Centralized Messaging with Clustering Over the WAN

This section addresses Cisco Unity Connection design issues for deploying centralized messaging with Unified CM clustering over the WAN with local failover. In the case of a WAN failure with this model, all remote messaging sites will lose voicemail capability until the WAN is restored. (See Figure 21-4.)

Clustering over the WAN supports local failover. With local failover, each site has a backup subscriber server physically located at the site. This section focuses on deploying Cisco Unity Connection centralized messaging with local failover for clustering over the WAN.

For additional information, refer to the section on Clustering Over the IP WAN, page 5-33.

Figure 21-4 Cisco Unity Connection Centralized Messaging and Clustering Over the WAN with Local Failover



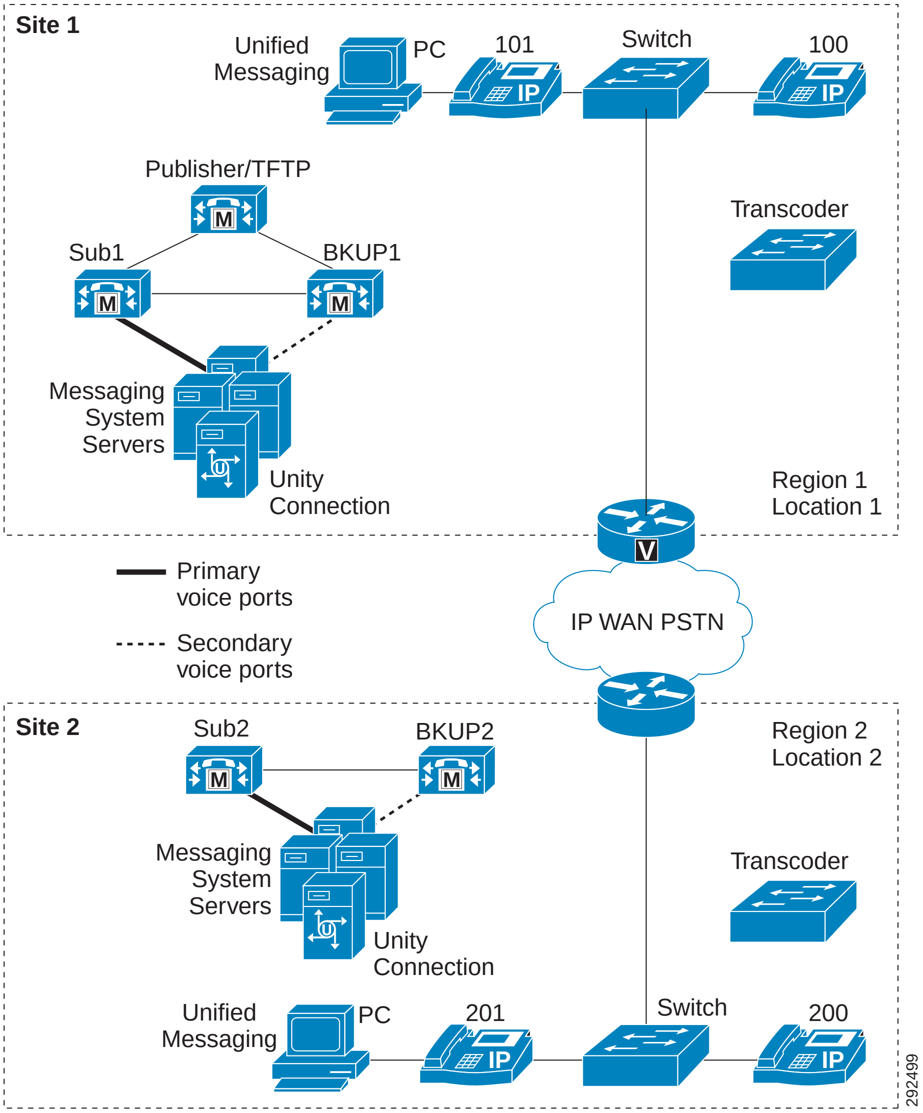
For minimum bandwidth requirements between clustered servers see the section on Local Failover Deployment Model, page 5-37.

Clustering over the WAN with Unified CM supports up to eight sites, as does Cisco Unity Connection. The voicemail ports are configured only at the site where the Cisco Unity Connection messaging system is located (see Figure 21-4). Voicemail ports do not register over the WAN to the remote site(s). Messaging clients at the other site(s) access all voicemail resources from the primary site. There is no benefit to configuring voice ports over the WAN to any of the remote sites because, in the event of a WAN failure, remote sites would lose access to the centralized messaging system. Because of bandwidth consideration, the voicemail ports should have TRaP disabled and all messaging clients should download voicemail messages to their local PCs (unified messaging only).

Distributed Messaging with Clustering Over the WAN

Local failover sites that also have Cisco Unity Connection messaging server(s) deployed would have voice ports registered to the local Unified CM subscriber server(s), similar to the centralized messaging model. For information about configuring the voice ports, see Voice Port Integration with a Unified CM Cluster, page 21-36, and Voice Port Integration with Dedicated Unified CM Backup Servers, page 21-38.

Figure 21-5 Cisco Unity Connection Distributed Messaging and Clustering over the WAN



In a purely distributed messaging implementation with clustering over the WAN, each site in the cluster would have its own Cisco Unity Connection messaging server with messaging infrastructure components. If not all of the sites have local Cisco Unity Connection messaging systems but some sites have local messaging clients using a remote messaging server(s), this deployment would be a combination model with both distributed messaging and centralized messaging. (See Combined Messaging Deployment Models, page 21-11.) In the event of a WAN failure in this model, all remote sites that use centralized messaging will lose voicemail capability until the WAN is restored.

Each site that does not have a local messaging server must use a single messaging server for all of its messaging clients, but all such sites do not have to use the same messaging server. For example, suppose Site1 and Site2 each have a local messaging server. Site3 can then have all of its clients use (register with) the messaging server at Site2, while Site4 can have all of its clients use the messaging server at Site1. Transcoder resources are required at sites that have local Cisco Unity Connection messaging server(s).

As with other distributed call processing deployments, calls going between these sites are transparent to gatekeeper call admission control, therefore you must configure regions and locations in Unified CM to provide call admission control. (See Managing Bandwidth, page 21-28.)

The distributed Cisco Unity Connection servers may also be networked digitally. For more information on this topic, refer to the *Cisco Unity Connection Networking Guide*, available at http://www.cisco.com. The networking guides are specific to the particular messaging store deployed.

Messaging Redundancy

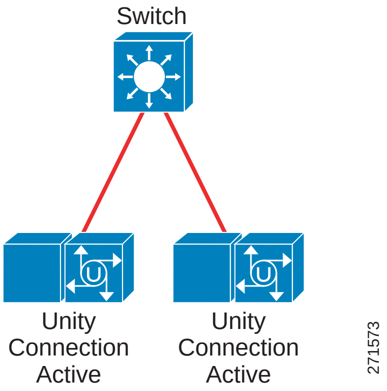
Messaging redundancy is discussed in this section as it refers to Cisco Unity Connection. Cisco Unity Express does not support messaging redundancy.

Cisco Unity Connection

Cisco Unity Connection supports messaging redundancy and load balancing in an active-active redundancy model consisting of two servers, a primary and a secondary, configured as an active/active redundant pair of servers, where both the primary and secondary servers actively accept calls as well as HTTP and IMAP requests. For more information, refer to the *Design Guide for Cisco Unity Connection*, available at http://www.cisco.com.

Figure 21-6 illustrates Cisco Unity Connection active/active messaging redundancy.

Figure 21-6 Redundancy of Cisco Unity Connection Messaging



Cisco Unity Connection SIP trunk implementation requires call forking for messaging redundancy functionality. Cisco Unified Communications Manager (Unified CM) supports the multi-destination SIP trunk feature. With this multi-destination SIP trunk feature, administrators can define full-mesh trunking between Cisco Unified CM and Cisco Unity Connection to achieve redundancy. Also, two separate SIP trunks can be configured, one for each server in a pair, and they can be added to the same route group associated to the same route list.The route group should be configured in top-down order so that calls are sent to the primary Unity Connection and overflow calls are sent to secondary Unity Connection server.

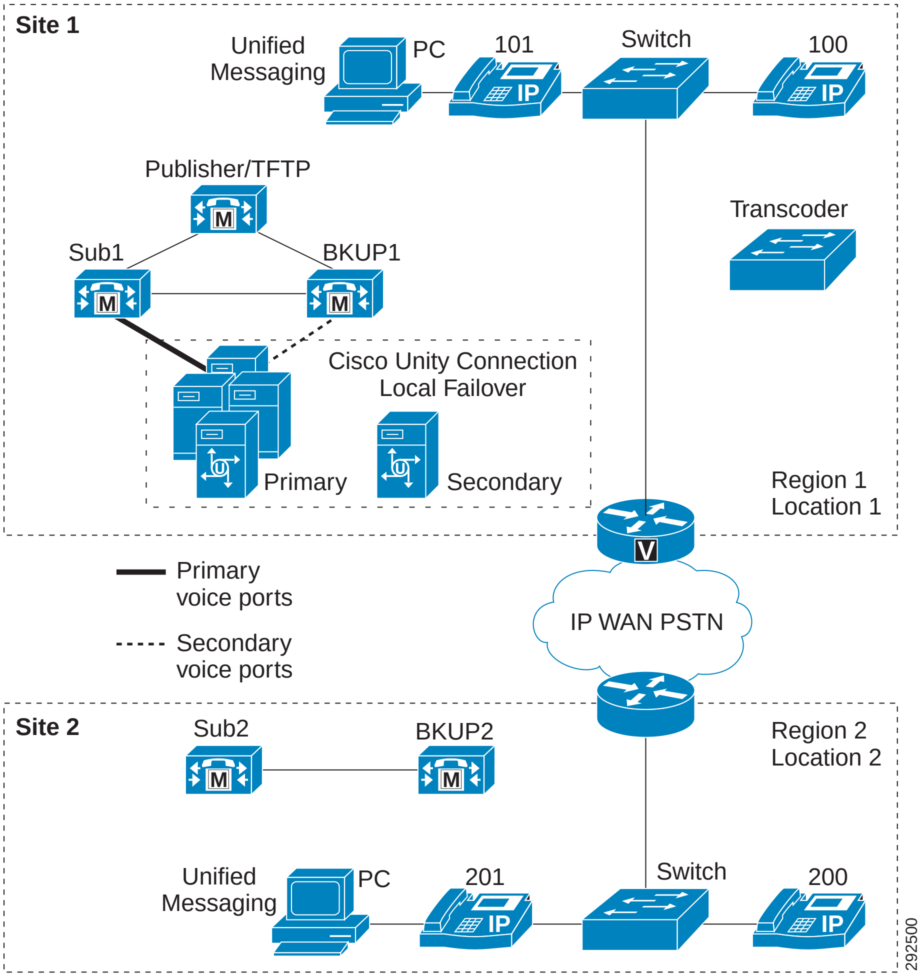
Note SIP OPTIONS Ping must be enabled on the Unified CM SIP trunk in order for Cisco Unity Connection failover to work properly.

Cisco Unity Connection Failover and Clustering Over the WAN

When deploying Cisco Unity Connection local failover with clustering over the WAN, apply the same design practices described in Centralized Messaging with Clustering Over the WAN, page 21-12, and Distributed Messaging with Clustering Over the WAN, page 21-14. The voice ports from the primary Cisco Unity Connection server should not cross the WAN during normal operation.

Figure 21-7 depicts Cisco Unity Connection local failover. Note that the primary and secondary Cisco Unity Connection servers are both physically located at the same site. Cisco Unity Connection failover supports up to the maximum number of remote sites available with clustering over the WAN for Unified CM.

Figure 21-7 Cisco Unity Connection Local Failover and Clustering Over the WAN

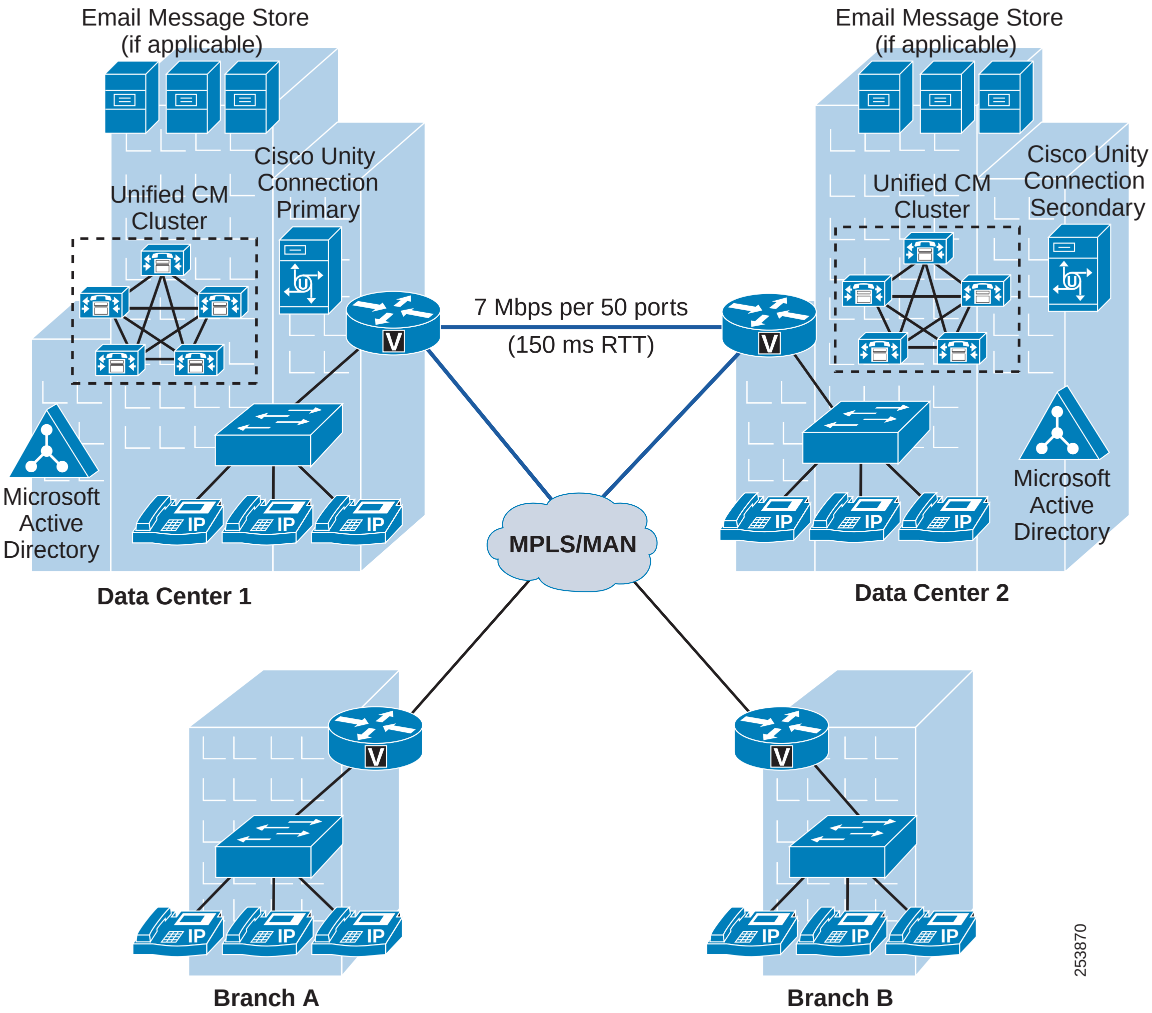


For information on configuring Cisco Unity Connection failover, refer to the *Cisco Unity Connection Failover Configuration and Administration Guide*, available at http://www.cisco.com.

Cisco Unity Connection Redundancy and Clustering Over the WAN

Cisco Unity Connection supports active/active clustering for redundancy and can be deployed over the WAN. The active/active or "high availability" configuration provides both high availability and redundancy. Both servers in the active/active pair run the Cisco Unity Connection application to accept calls as well as HTTP and IMAP requests from clients. Each of the servers from the cluster can be deployed over the WAN at different sites following required design consideration. Figure 21-8 depicts a Cisco Unity Connection active/active deployment for geographically separated data centers

Figure 21-8 Cisco Unity Connection with High Availability Between Two Sites



The following requirements apply to deployments of Cisco Unity Connection servers over different sites:

- Maximum of 150 ms RTT between an active/active pair at different sites.
- Minimum of 7 Mbps bandwidth is required for every 50 ports. (For example, 250 ports require 35 Mbps.)

Note Bandwidth and latency requirements may differ for different versions of Cisco Unity Connection.

For a complete set of requirements, refer to the latest version of the *System Requirements for Cisco Unity Connection*, available at

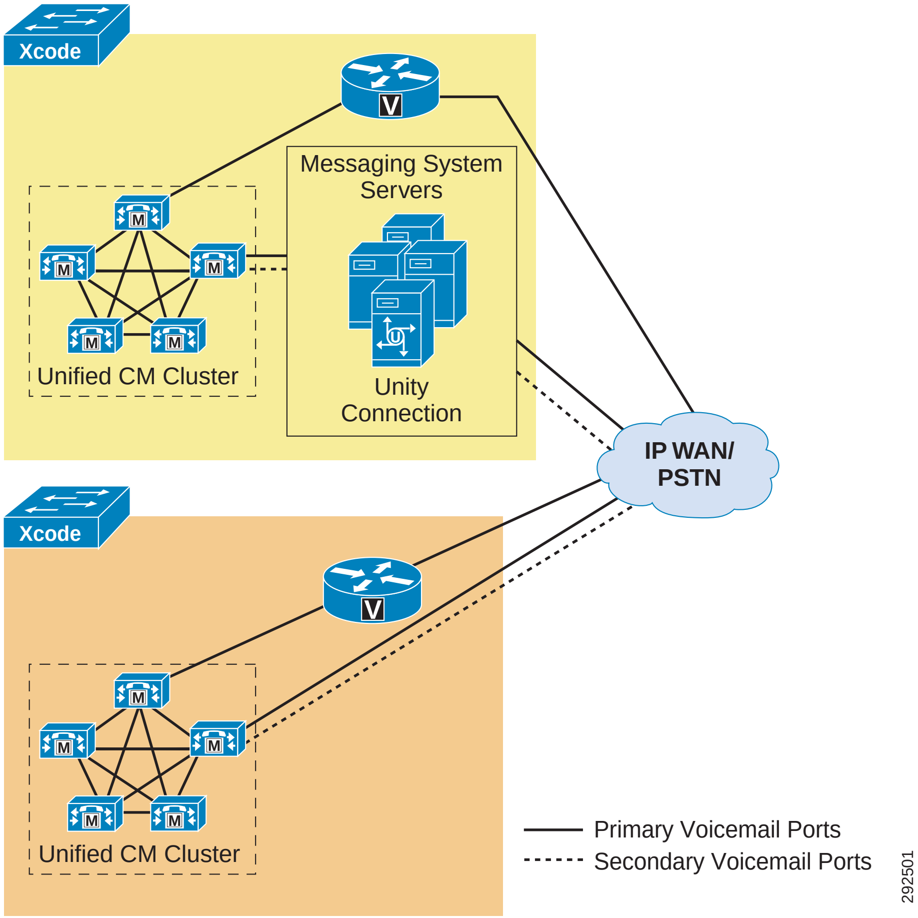
http://www.cisco.com/en/US/products/ps6509/prod_installation_guides_list.html

Note The Cisco Unity Connection cluster feature is not supported for use with Cisco Business Edition 3000 and 5000.

Centralized Messaging with Distributed Unified CM Clusters

Cisco Unity Connection can also be deployed in a centralized messaging configuration with multiple Unified CM clusters (see Figure 21-9). See the section on Integration with Cisco Unified CM, page 21-31, for details on multiple integrations and MWI considerations with multiple Unified CM clusters.

Figure 21-9 Integrating Cisco Unity Connection with Multiple Unified CM Clusters



For the configuration in Figure 21-9, messaging clients at both Cluster 1 and Cluster 2 sites use the Cisco Unity Connection messaging infrastructure physically located at Cluster 1.

Cisco Unity Express Deployment Models

This section begins with a quick overview of Cisco Unity Express, covering product related information. Next the deployment models section presents three supported deployment models with Cisco Unity Express, focusing on distributed voice messaging with both centralized and distributed call processing followed by some deployment characteristics and design guidelines. Lastly, this section discusses the signaling call flows and the various protocols used between Cisco Unity Express and Unified CM as well as between Cisco Unity Express and Unified SRST or Unified CME in SRST mode.

Overview of Cisco Unity Express

Cisco Unity Express is Linux-based software running on a Cisco Network Module in Cisco Integrated Services Routers (ISRs). It is an entry-level auto-attendant (AA) and voicemail solution that can be deployed with Cisco Unified Communications Manager (Unified CM), Cisco Unified SRST, or Cisco Unified Communications Manager Express (Unified CME). In prior releases, Cisco Unity Express was limited to a co-resident deployment with Unified CME or a Survivable Remote Site Telephony (SRST) router. However, with the H.323-to-SIP call routing capability introduced in Cisco IOS Release 12.3(11)T, Cisco Unity Express and SRST or Unified CME can reside on two separate routers when deployed with Unified CM or Unified CME, respectively. Cisco Unity Express uses SIP to communicate with Cisco Unified Communications Manager Express (Unified CME) while Cisco Unity Express uses JTAPI to connect to Cisco Unified Communications Manager (Unified CM).

For more information on supported hardware platforms and capacity with Cisco Unity Express, refer to the product release note available at
http://www.cisco.com/en/US/products/sw/voicesw/ps5520/prod_release_notes_list.html.

For details on interoperability of Unified CM and Unified CME, see Interoperability of Unified CM and Unified CM Express, page 8-44.

For additional information on supported deployment models with Unified CME, refer to the appropriate Cisco Unified Communications Manager Express design documentation available at
http://www.cisco.com.

Deployment Models

Cisco Unity Express can be deployed as a single site or distributed voicemail and automated attendant (AA) solution for Cisco Unified Communications Manager (Unified CM) or Unified Communications Manager Express (Unified CME). However, Cisco Unity Express is supported with all of the Cisco Unified CM deployment models, including:

- Single-site deployments
- Multisite deployments with centralized call processing
- Multisite deployments with distributed call processing

Figure 21-10 shows a centralized call processing deployment incorporating Cisco Unity Express, and Figure 21-11 shows a distributed call processing deployment.

Cisco Unity Express sites controlled by Unified CME, as well as other sites controlled by Unified CM, can be interconnected with each other using H.323 or SIP trunking protocol. Although Cisco Unity Express can integrate with either Unified CM or Unified CME, it cannot integrate with both simultaneously.

Note Cisco Unity Express supports a centralized deployment model with up to 10 Unified CMEs.

Figure 21-10 Cisco Unity Express in a Centralized Call Processing Deployment

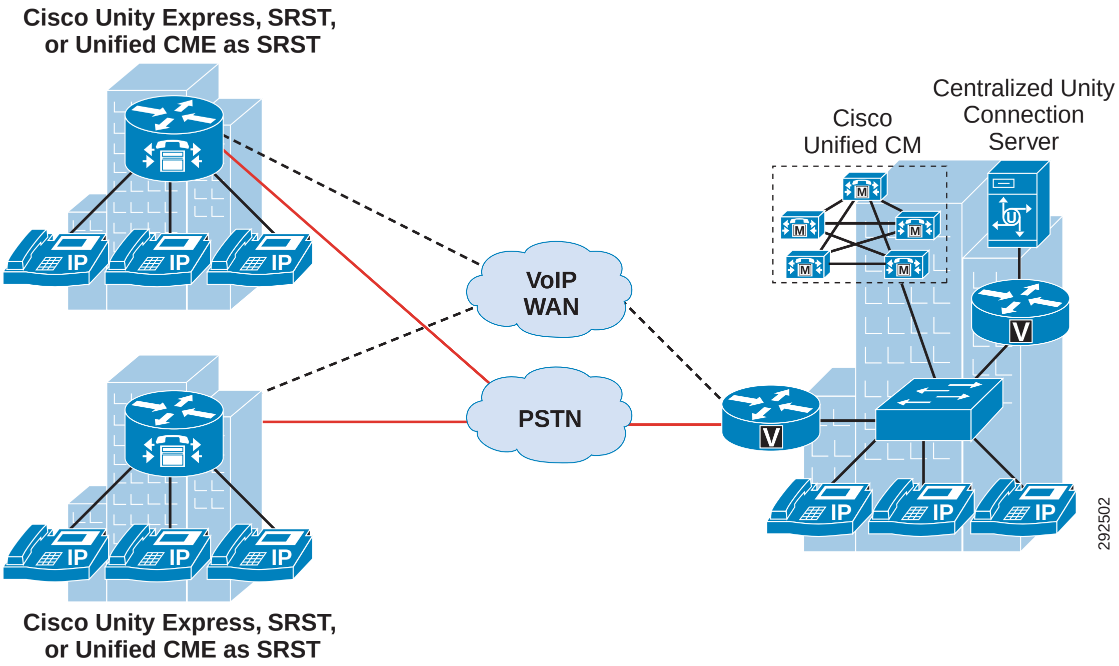
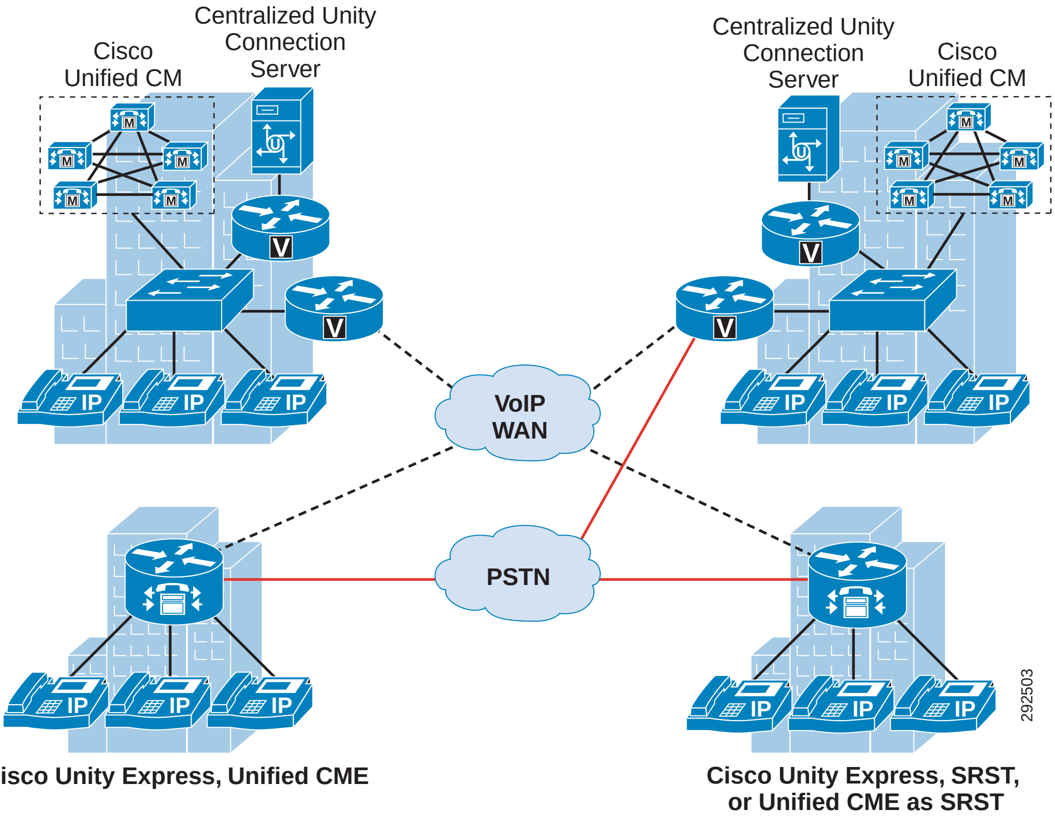


Figure 21-11 Cisco Unity Express in a Distributed Call Processing Deployment


Cisco Unified CM
Centralized Unity Connection Server
Centralized Unity Connection Server
Cisco Unified CM
M
U
V
IP
VoIP WAN
PSTN
292503
Cisco Unity Express, Unified CME
Cisco Unity Express, SRST, or Unified CME as SRST

The most likely deployment model to use Cisco Unity Express is the multisite WAN model with centralized call processing, where Cisco Unity Express provides distributed voicemail at the smaller remote offices and a central Cisco Unity Connection system provides voicemail to the main campus and larger remote sites.

Use Cisco Unity Express as a distributed voicemail solution if any of the following conditions apply to your Unified CM network deployment:

- Survivability of voicemail and AA access must be ensured regardless of WAN availability.
- Available WAN bandwidth is insufficient to support voicemail calls traversing the WAN to a central voicemail server.
- There is limited geographic coverage of the AA or branch site PSTN phone numbers published to the local community, and these numbers cannot be dialed to reach a central AA server without incurring toll charges.
- The likelihood is high that a PSTN call into a branch office will be transferred from the branch AA to a local extension in the same office.
- Management philosophy allows remote locations to select their own voicemail and AA technology.

The following characteristics and guidelines apply to Cisco Unity Express in either a centralized or distributed Unified CM deployment:

- A single Cisco Unity Express can be integrated with a single Unified CM cluster.
- Cisco Unity Express integrates with Unified CM using a JTAPI application and Computer Telephony Integration (CTI) Quick Buffer Encoding (QBE) protocol. CTI ports and CTI route points control the Cisco Unity Express voicemail and automated attendant (AA) applications.
- Cisco Unity Express provides voicemail functionality to Cisco Unified IP Phones running Skinny Client Control Protocol (SCCP). Cisco Unity Express 2.3 and later releases also provide support for Session Initiation Protocol (SIP) IP phones with Unified CM.
- The following CTI route points are defined on Unified CM for Cisco Unity Express:
  - Automated attendant entry point (Cisco Unity Express can contain up to five distinct AAs and may therefore require up to five different route points.)
  - Voicemail pilot number
  - Greeting management system (GMS) pilot number (Optional; if the GMS is not used, then this route point need not be defined.)
- The number of CTI ports and mailboxes supported for Cisco Unity Express on Unified CM depends on the hardware platform. For details, refer to the Cisco Unity Express data sheet available at:

  http://www.cisco.com/en/US/products/sw/voicesw/ps5520/products_data_sheets_list.html
- For Cisco Unity Express deployments that require more than the maximum number of supported mailboxes, consider using Cisco Unity Connection or other voicemail solutions.
- Each Cisco Unity Express mailbox can be associated with a maximum of two different extensions, if needed.
- The automated attendant function for any office deployed with Cisco Unity Express can be local to the office (using the AA application in Cisco Unity Express) or centralized (using Cisco Unity Express for voicemail only).
- Cisco Unity Express can be networked with other Cisco Unity Expresses or with Cisco Unity Connection via Voice Profile for Internet Mail (VPIM) version 2. Thus, a Cisco Unity Express subscriber can send, receive, or forward messages to or from another remote Cisco Unity Express or Cisco Unity Connection subscriber.
- Cisco Unity Express allows you to specify up to three Unified CMs for failover. If IP connectivity to all three Unified CMs is lost, Cisco Unity Express switches to Survivable Remote Site Telephony (SRST) call signaling, thus providing AA call answering service as well as mailbox access to IP phones and PSTN calls coming into the branch office.
- Cisco Unity Express automated attendant supports dial-by-extension and dial-by-name functions. The dial-by-extension operation enables a caller to transfer a call to any user endpoint in the network. The dial-by-name operation uses the directory database internal to Cisco Unity Express and does not interact with external LDAP or Active Directory databases.
- Centralized Cisco Unity Express with Unified CM is not supported.
- Cisco Unity Express is not supported in pure SIP networks that do not have either Cisco Unified CM or Unified CME controlling the SIP phones.
- Cisco Unity Express can be deployed on a separate Unified CME or SRST router or a separate PSTN gateway.
- When Cisco Unity Express is deployed on a router separate from Unified CME or SRST, configure the command **allow-connections h323 to sip** for H.323-to-SIP routing.

Figure 21-12 shows the protocols involved in the call flow between Unified CM and Cisco Unity Express.

Figure 21-12 Protocols Used Between Cisco Unity Express and Unified CM

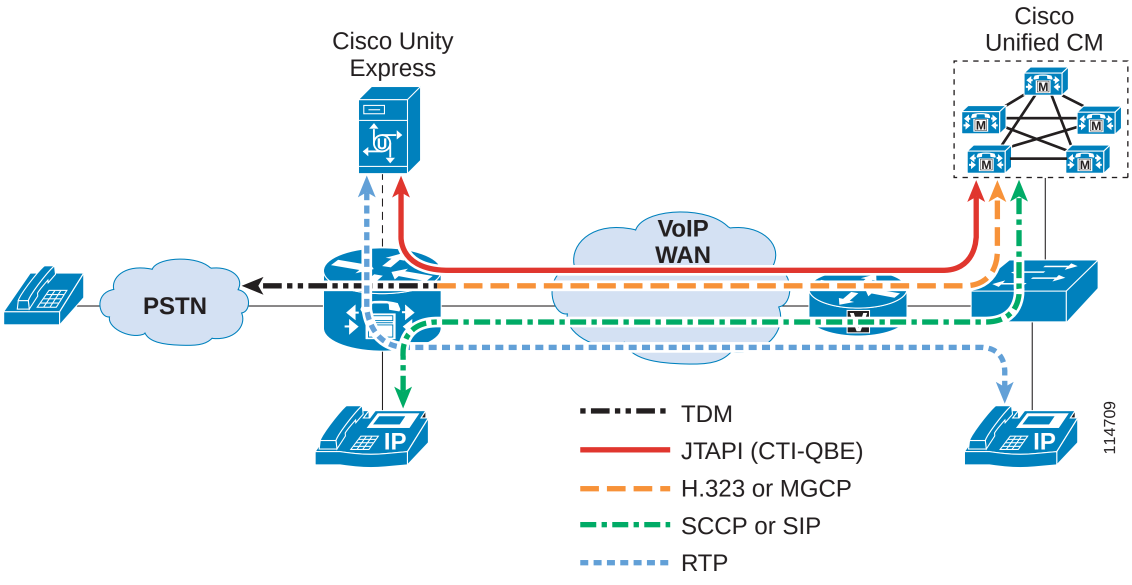


Figure 21-12 illustrates the following signaling and media flows:

- Phones are controlled via SCCP or SIP from Unified CM.
- Cisco Unity Express is controlled via JTAPI (CTI-QBE) from Unified CM.
- The Message Waiting Indicator (MWI) on the phone is affected by Cisco Unity Express communicating a change of mailbox content to Unified CM via CTI-QBE, and by Unified CM in turn sending a MWI message to the phone to change the state of the lamp.
- The voice gateway communicates via H.323, SIP, or MGCP to Unified CM.
- Real-Time Transport Protocol (RTP) stream flows carry the voice traffic between endpoints.

Figure 21-13 shows the protocols involved in the call flow between the router for SRST or Unified CME in SRST mode and Cisco Unity Express when the WAN link is down.

Figure 21-13 Protocols Used Between Cisco Unity Express and the Router for SRST or Unified CME in SRST Mode

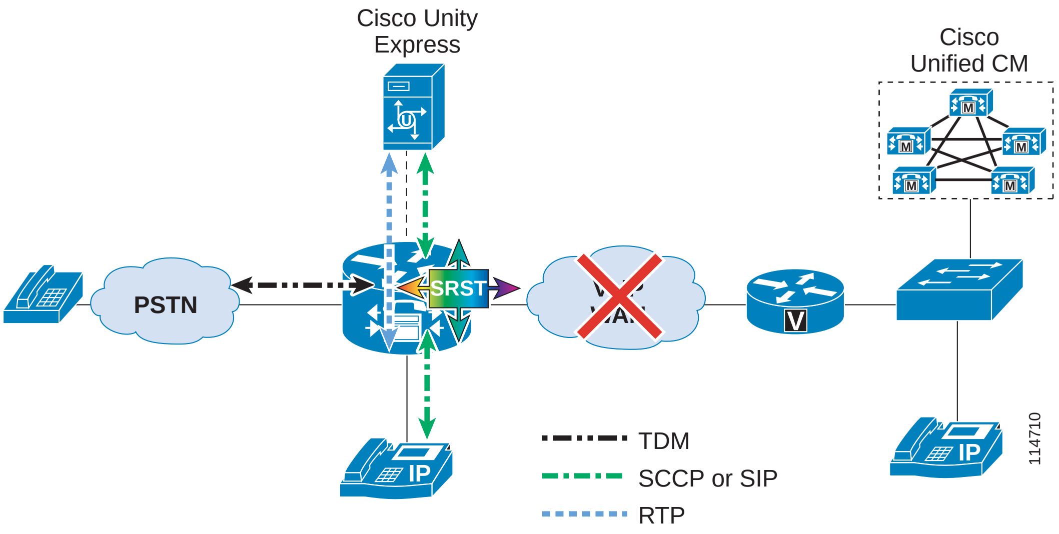


Figure 21-13 illustrates the following signaling and media flows:

- Phones are controlled via SCCP or SIP from the router for SRST or Unified CME in SRST mode.
- Cisco Unity Express communicates with the SRST router via an internal SIP interface.
- Although MWI changes are not supported in SRST mode with previous releases of Cisco Unity Express, voice messages can be sent and retrieved as during normal operation, but the MWI lamp state on the phone remains unchanged until the phone registers again with Unified CM. At that time, all MWI lamp states are automatically resynchronized with the current state of the users' Cisco Unity Express voicemail boxes. Cisco Unity Express 3.0 and later releases support MWI for SRST mode.
- Cisco Unity Express supports SIP Subscriber/Notify and Unsolicited Notify to generate MWI notifications, in both Unified CME and SRST modes.
- RTP stream flows carry the voice traffic between endpoints.
- SRST subscribes to Cisco Unity Express for MWI for each of the ephone-dns registered to receive MWI notifications.

Note Unified CM MWI (JTAPI) is independent of the SIP MWI methods.

Voicemail Networking

This section covers specific considerations for voicemail networking, including Cisco Unity Connection and Cisco Unity Express. For information specific to voicemail networking in Unity Connection, refer to the *Design Guide for Cisco Unity Connection*, available at http://www.cisco.com.

Voicemail networking is the ability to allow subscribers (voicemail users) to send, receive, reply to, and forward voicemail messages between systems such as Cisco Unity Connection and Cisco Unity Express using an embedded Simple Mail Transfer Protocol (SMTP) server and a subset of the Voice Profile for Internet Mail (VPIM) version 2 protocol. All three voicemail messaging products support interoperability between one another using VPIM messaging.

Cisco Unity Express Voicemail Networking

Cisco Unity Express communicates with Cisco Unity Connection by means of VPIM for message routing and SMTP for message delivery. Cisco Unity Express voicemail networking provides the following capabilities:

- Subscribers can receive, send, and forward messages to or from another remote Cisco Unity Express or Cisco Unity Connection for locations configured on the originating system.
- Subscribers can also reply to a remote message received from a remote system.
- Subscribers can be recipients of a distribution list or individual message originating from Cisco Unity Connection.

For more information on voicemail networking with a specific product, refer to the corresponding voicemail product documentation available at http://www.cisco.com.

Interoperability Between Multiple Cisco Unity Connection Clusters or Networks

Cisco Unity Connection (digital network, standalone servers, or cluster) can interoperate with another Cisco Unity Connection (digital network), thus enabling users to achieve directory sharing, easy administration, and other features, as well as expanding the total number of nodes (cluster or standalone server) up to 20. Consider the following points when deploying Cisco Unity Connection for interoperability with another Cisco Unity Connection network:

- A Cisco Unity Connection standalone server, cluster, or digital network can interoperate with another Cisco Unity Connection server or digital network.
- If any of the Cisco Unity Connection nodes in the digital network system is running Cisco Unity Connection 7.0, then the maximum number of users supported is 50,000.
- Each Cisco Unity Connection digital network can support a maximum of 10 servers.
- One Cisco Unity Connection can be a member of only one Cisco Unity Connection digital network.
- The maximum number of users and/or contacts can be 100,000 in any interoperating system, and a maximized system will allow only deletions and changes.
- Cisco Business Edition 3000 and 5000 are not supported.
- Each Cisco Unity Connection digital network must have one server defined as the bridgehead or site gateway. The site gateway is used to communicate with other digital networks.
- The Cisco Unity Connection server designated as the site gateway should be version 8.0 or higher.

For more information on these interoperability options, refer to the latest version of the *Networking Guide for Cisco Unity Connection*, available at

http://www.cisco.com/en/US/products/ps6509/prod_maintenance_guides_list.html

Cisco Unity Connection Virtualization

The Cisco Unified Computing System (UCS) is a next-generation data center platform that unites computing, networking, storage access, and virtualization into a cohesive system designed to reduce total cost of ownership (TCO) and increase business agility. Cisco Unity Connection supports virtualization over VMware with the Cisco Unified Computing system.

The following key design considerations apply to Cisco Unity Connection virtualization:

- Supports up to 20,000 users
- The Tested Reference Configurations include selected Cisco Unified Computing System (UCS) platforms. Other platforms may be supported with the specifications-based hardware support policy.
- VMware ESXi is required for virtualization.
- Servers in an active/active cluster should be on separate blades, preferably on different chassis.

Note One CPU core per physical server must be left idle and reserved for the ESXi scheduler.

For more information on deploying Cisco Unified Communications and Cisco Unity Connection in a virtualized system, refer to the documentation available at

http://www.cisco.com/go/uc-virtualized

General information about deploying Unified Communications on virtualized servers is also available in the section on Deploying Unified Communications on Virtualized Servers, page 5-46.

For Cisco Unity Connection virtualization, also refer to the latest version of the *Design Guide for Cisco Unity Connection* available at

http://www.cisco.com/en/US/products/ps6509/products_implementation_design_guides_list.html

Best Practices for Voice Messaging

This section discusses some general best practices and guidelines that were not mentioned previously yet are important aspects of the products and should be considered in the solution. They are separated into two groupings, with Cisco Unity Connection in one grouping and Cisco Unity Express in another.

Best Practices for Deploying Cisco Unity Connection with Unified CM

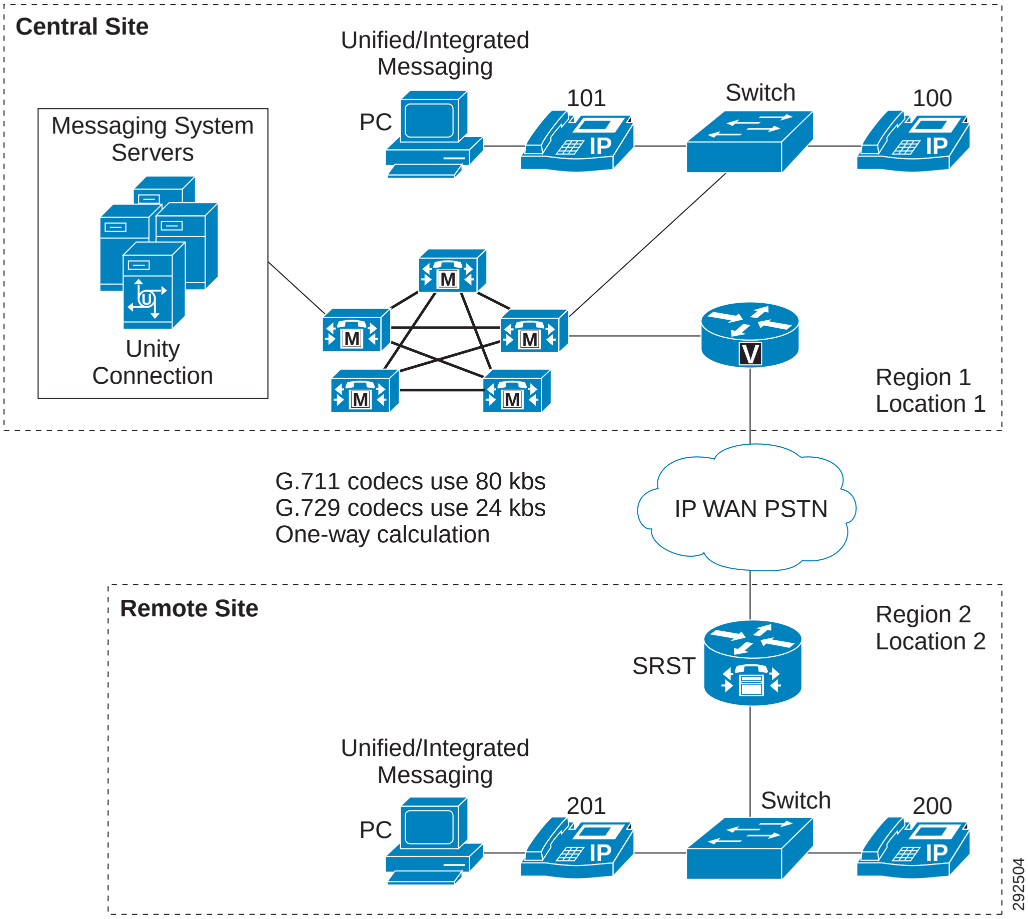
This section applies to Cisco Unity Connection. For Cisco Unity Express, see Best Practices for Deploying Cisco Unity Express, page 21-42.

Managing Bandwidth

Unified CM provides a variety of features for managing bandwidth. Through the use of regions, locations, and even gatekeepers, Unified CM can ensure that the number of voice calls going over a WAN link does not oversubscribe the existing bandwidth and cause poor voice quality. Cisco Unity Connection relies on Unified CM to manage bandwidth and to route calls. If you deploy Cisco Unity Connection in an environment where calls or voice ports might cross WAN links, these calls will be transparent to gatekeeper-based call admission control. This situation occurs any time the Cisco Unity Connection server is servicing either distributed clients (distributed messaging or distributed call processing) or when Unified CM is remotely located (distributed messaging or centralized call processing). Unified CM provides regions and locations for call admission control.

Figure 21-14 uses a small centralized messaging and centralized call processing site to illustrate how regions and locations work together to manage available bandwidth. For a more detailed discussion of regions and locations, refer to the chapter on Call Admission Control, page 11-1.

Figure 21-14 Locations and Regions



In Figure 21-14, regions 1 and 2 are configured to use G.711 for intra-region calls and G.729 for inter-region calls. Locations 1 and 2 are both set to 24 kbps. Location bandwidth is budgeted only in the case of inter-location calls.

An intra-region (G.711) call would not be budgeted against the available bandwidth for the location. For example, when extension 100 calls extension 101, this call is not budgeted against the 24 kbps of available bandwidth for Location 1. However, an inter-region call using G.729 is budgeted against both bandwidth allocations of 24 kbps for Location 1 and Location 2. For example, when extension 100 calls extension 200, this call would be connected but any additional (simultaneous) inter-region calls would receive reorder (busy) tone.

Native Transcoding Operation

In Cisco Unity Connection, native transcoding occurs when a call is negotiated between an IP endpoint and the Cisco Unity Connection server in one codec and is recorded or played out in another codec format. If a call is negotiated in G.729 and the system-wide recording format is done in G.711, then the server has to transcode that call natively. Cisco Unity Connection native transcoding does not use external hardware transcoders but instead uses the server's main CPU. This is what is meant by native transcoding.

Cisco Unity Connection Operation

In Cisco Unity Connection, a call in any codec format supported by Cisco Unity Connection SCCP or SIP signaling (G.711 mu-law, G.711 a-law, G.729, iLBC, and G.722) will always be transcoded to Linear PCM. From Linear PCM, the recording is encoded in the system-level recording format (Linear PCM, G.711 mu-law/a-law, G.729a, or G.726), which is set system-wide in the general configuration settings (G.711 mu-law is the default). In the rest of this chapter, we refer to the codec negotiated between the calling device and Unity Connection as the "line codec," and we refer to the codec set in the system-level recording format as the "recording codec."

Because transcoding is inherent in every connection, there is little difference in system impact when the line codec differs from the recording codec. The exception to this is when using iLBC or G.722. G.722 and iLBC require more computation to transcode, therefore they have a higher system impact. G.722 and iLBC use approximately twice the amount of resources as G.711 mu-law. The subsequent impact this has is that a system can support only half as many G.722 or iLBC connections as it can G.711 mu-law connections.

As a general rule, Cisco recommends leaving the default codec as G.711. If the configuration is constrained by disk space, then a lower bit rate codec such as G.729a or G.726 can be configured as the recording format; however, keep in mind that the audio quality will not have the fidelity of G.711 audio. Also, if G.722 is used by devices on the line, then linear pulse code modulation (PCM) is an option to improve the audio quality of the recording. This will, however, increase the disk usage and impact disk space.

There are also a few reasons to change the recording codec or to choose to advertise only specific line codecs. Consider the following factors when deciding on the system-level recording format and the advertised codecs on the SCCP or SIP integration:

- Which codecs will be negotiated between the majority of the endpoints and Cisco Unity Connection? This will help you decide on which codecs need to be advertised by Cisco Unity Connection and which do not. You can then decide on when you need Unified CM to provide hardware transcoding resources in lieu of doing computationally significant native transcoding in Cisco Unity Connection, such as when requiring a large number of clients connected to Cisco Unity Connection using G.722 or iLBC.
- Which types of graphical user interface (GUI) clients (web browsers, email clients, media players, and so forth) will be fetching the recordings, and which codecs do the GUI clients support?
- What quality of the sound is produced by the selected codec? Some codecs are higher quality than others. For example, G.711 has a higher quality than G.729a, and it is a better choice if higher audio quality is necessary.
- How much disk space does the codec use per second of recording time?

Table 21-5 summarizes the characteristics of the codec formats supported by Cisco Unity Connection.

Table 21-5 Codec Characteristics

Recording Format (Codec)	Audio Quality	Supportability	Disk Space (Bandwidth)
Linear PCM	Highest	Widely supported	16 kbps
G.711 mu-law and a-law	Moderate	Widely supported	8 kbps
G.729a	Lowest	Poorly supported	1 kbps
G.726	Moderate	Moderately supported	3 kbps
GSM 6.10	Moderate	Moderately supported	1.6 kbps

Refer to the *System Administration Guide for Cisco Unity Connection* for details on changing the codec advertised by Cisco Unity Connection. The choices for advertised codecs are G.711 mu-law, G.711 a-law, G.729, iLBC and G.722. There is also a list of preferences according to how they are ordered in the list (top-down). For SCCP integrations, the order of the codecs has no bearing because codecs are advertised and Unified CM negotiates the codec based on the location of the port and device in the negotiated call. For SIP integrations, however, the order list is significant. If the codec is preferred, then Cisco Unity Connection will advertise that it supports both protocols but will prefer to use the one specified over the other.

For information on how to change the system-level recording format in Cisco Unity Connection Administration, refer to the *System Administration Guide for Cisco Unity Connection*.

Integration with Cisco Unified CM

Cisco Unified CM can integrate with Cisco Unity Connection via SCCP or SIP. This section discusses some specifics of that integration regarding phones, SIP trunks, and voice ports.

In Cisco Unity Connection, users are associated to a phone system that contains one or more port groups. The port groups are associated with MWI ports; thus, the MWI requests are made through the ports associated to that specific port group. Cisco Unity Connection phone systems and port groups are configured with the System Administrator.

Cisco Unity Connection supports a maximum of 90 simultaneous phone systems and port groups.Cisco recommends using a maximum of 90 port groups if you are using only the touchtone conversation (telephone user interface, or TUI) and voice recognition (voice user interface, or VUI) features of Unity Connection. If you are using all other features such as calendaring and text-to-speech (TTS), then Unity Connection supports a maximum of 60 simultaneous phone systems. These features function the same way for both SCCP and SIP integrations. For details, refer to the appropriate Cisco Unity Connection administration guides available at

http://www.cisco.com/en/US/products/ps6509/index.html

In addition to the option of adding multiple clusters by adding additional integrations for each new Unified CM cluster in Cisco Unity Connection, Unified CM supports Annex M.1, Message Tunneling for QSIG, which gives administrators the ability to enable QSIG on intercluster trunks (ICTs) between Unified CM clusters. When QSIG is enabled on ICTs, Cisco Unity Connection needs to integrate with only one Unified CM cluster and designate ports only in this one cluster for turning MWIs on and off, even when supporting multiple clusters. The Annex M.1 feature in Unified CM allows for propagation of the MWI requests across the ICTs to the proper Unified CM cluster and phone within that cluster. All calls originating in other clusters can be forwarded to the Cisco Unity Connection server integrated to that one cluster. There is no need to designate MWI ports on the other clusters when Annex M.1 is enabled on the ICT.

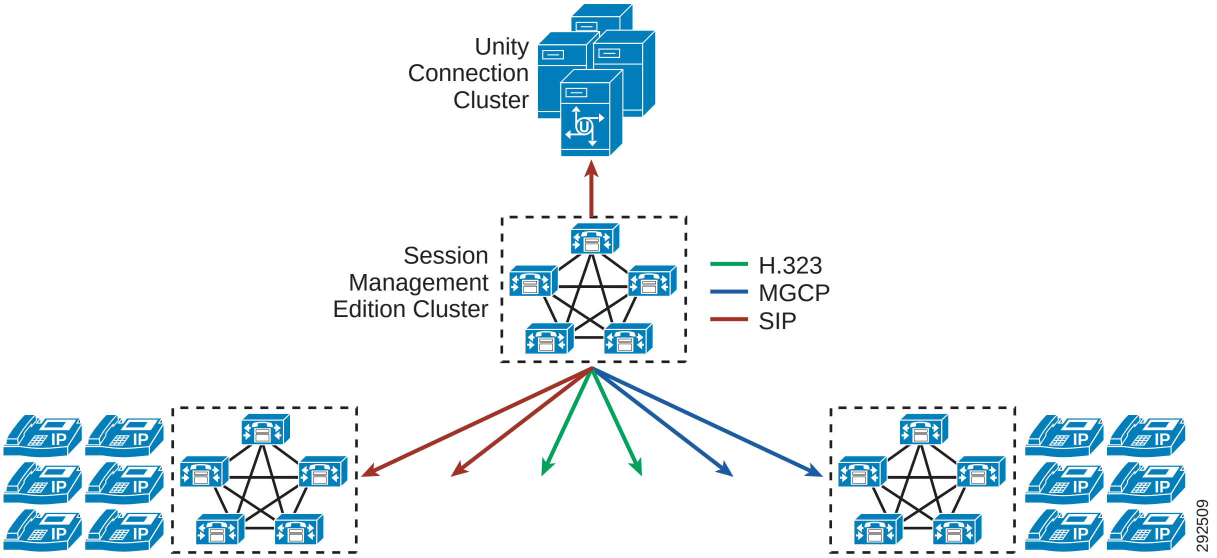
For more information on Annex M.1, refer to the protocol descriptions in the *Cisco Unified Communications Manager System Guide*, available at

http://www.cisco.com/en/US/products/sw/voicesw/ps556/prod_maintenance_guides_list.html

Integration with Cisco Unified CM Session Management Edition

Cisco Unity Connection can be integrated with Cisco Unified CM Session Management Edition to provide voice messaging services to the users associated with all leaf Unified Communications clusters. (See Figure 21-15.)

Figure 21-15 Cisco Unity Connection Deployment with Unified CM Session Management Edition



The following information must be sent on the intercluster trunks between Unified Communications leaf clusters and Unified CM Session Management Edition, and on the SIP trunk to Cisco Unity Connection:

- Original called party number or redirecting number
- Calling party number
- Reason for call forward

Non- Q.SIG Trunk

For a non-Q.SIG trunk, the following settings should be enabled to deliver the original called party number or redirecting number:

- Inbound and outbound redirecting number information element (IE) delivery on MGCP and H.323 gateways and H.323 trunks
- Inbound and outbound redirecting diversion header delivery on SIP trunks

Diversion information that is sent on non-Q.SIG MGCP, H.323, or SIP trunks picks up only the calling party transformations that are defined by the voice mailbox mask of the voicemail profile that is assigned to the redirecting DN. Any calling party transformations that are defined in a route pattern or route list, or through outbound calling party transformation calling search spaces (CSSs), are not applied to diversion information.

Q.SIG-Enabled Trunk

For Q.SIG-enabled SIP, MGCP and H.323 trunks, the original called party number is sent in Q.SIG diverting leg information application protocol data units (APDUs).

On Q.SIG-enabled H.323, MGCP, and SIP trunks all calling, called, and redirecting number information is always sent in the encapsulated Q.SIG message and not in the outer H.323 message or SIP headers. The sent diversion information does not pick up any calling party transformation and does not honor any voicemail mask setting. Q.SIG tunneling-enabled trunks do not support transport of the "+" character in Q.SIG APDUs. Because of this limitation, the user's voice mailbox number should be of the same format as the directory number used in the leaf Unified Communications system. For example:

- Users with directory numbers of the format 4YYYY should have a corresponding voice mailbox number of the same 4YYYY format.
- Users with directory numbers of the E.164 format +XX4YYY should have a corresponding voice mailbox number of the same E.164 +XX4YYYY format.

Cisco Unity Connection allows an alternate extension to be associated with the voice mailbox of the user. For example:

- Primary VM box number: 4YYYY
- Alternate VM box number in +E.164: +XX4YYY

Redirected Dialed Number Information Service (RDNIS) is not supported with Q.SIG-enabled H.323 or SIP trunks. The original called party or redirecting number is sent in a Q.SIG DivertingLegInformation2 APDU instead of via RDNIS.

E.164 Number Support with Cisco Unity Connection

Cisco Unity Connection supports the E.164 number format for the following fields:

- End users' primary extensions
- Transfer rule extensions for the end users
- System call handler extensions
- Directory handler extensions
- Interview handler extensions
- Notification device phone numbers for the end users
- Personal contact phone numbers for the end users
- System contact phone numbers for the Cisco Unity Connection System
- Personal call transfer rule (PCTR) phone numbers for the Cisco Unity Connection System
- Alternate extensions for the end users
- Restriction patterns for the Cisco Unity Connection System
- Message waiting indicator (MWI) extensions for the Cisco Unity Connection System

When importing users from LDAP with E.164-formatted primary phone numbers, use the regular expression and replacement pattern that together convert phone numbers into extensions. For more information on this, refer to the sections on converting phone numbers into extensions in the latest version of the *System Administration Guide for Cisco Unity Connection*, available at

http://www.cisco.com/en/US/products/ps6509/prod_maintenance_guides_list.html

If you want to import users from Cisco Unified Communications Manager (Unified CM) with E.164 formatted extensions through AXL integration, you will have to export the E.164 extensions from Unified CM into a comma-separated values (CSV) file and perform the necessary translations on the alternate extensions (in Excel, for example) prior to using the Bulk Administration Tool (BAT) to import them into Unity Connection. For more details on using the Cisco Unity Connection Bulk Administration Tool, refer to the latest version of the *User Moves, Adds, and Changes Guide for Cisco Unity Connection*, available at

http://www.cisco.com/en/US/products/ps6509/prod_maintenance_guides_list.html

Enhanced Message Waiting Indicator (eMWI)

Enhanced Message Waiting Indicator (eMWI) is an enhancement to traditional MWI, and it provides a visual indication of the number of voice messages. Traditional MWI works in a binary format by either enabling or disabling the message lamp on the phone whenever a new voice message arrives in or is deleted from a user's voicemail box. EMWI works with Cisco Unity Connection and is supported on the Cisco Unified IP Phones 8900 and 9900 Series SIP phones.

eMWI is a visual indication of unplayed messages in the user's voicemail box, with a colored indication depicting the status of the message. An unplayed message displays a red indication on the screen of the phone. eMWI is supported on Unified CM for Cisco Unity Connection through SIP and SCCP integrations. eMWI does not function when the system is running in SRST mode. In an integration with Cisco Unity Connection, only the messages stored on the Cisco Unity Connection servers will be indicated with eMWI, and any messages stored on an external IMAP server will not be indicated.

eMWI works in distributed call processing environments with Unified CM. In a system with distributed call processing and centralized voice messaging integration, where one cluster provides the connectivity to the voice messaging server through an intercluster trunk (H.323 or SIP), eMWI updates over the intercluster trunk are supported and are displayed on the end device. (See Figure 21-16.)

Note eMWI also works in a distributed call processing environment with centralized messaging over an intercluster trunk (H.323 or SIP).

Figure 21-16 Enhanced Message Waiting Indicator (eMWI)

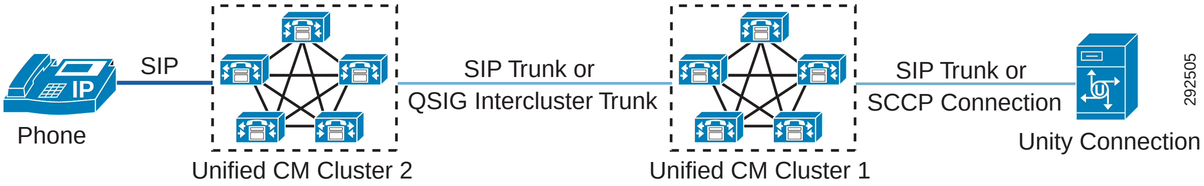
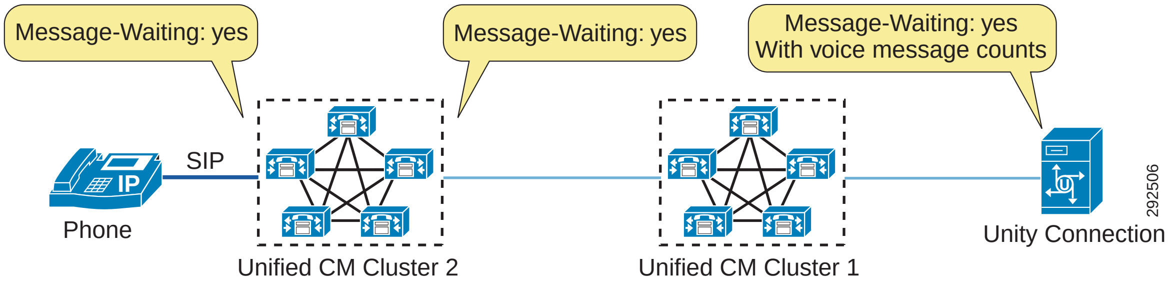


Figure 21-17 illustrates eMWI over an intercluster trunk (H.323 or SIP) in a distributed call processing environment with centralized voice messaging.

Figure 21-17 eMWI with Distributed Call Processing and Centralized Voice Messaging



As shown in Figure 21-17, Cluster 2 and its voice messaging solution support eMWI, but Cluster 1 does not. If an eMWI update with a voice message count is sent from the voice messaging solution intended for the Cluster 2 phone, Cluster 1 will forward only a standard MWI to Cluster 2 without the voice message count.

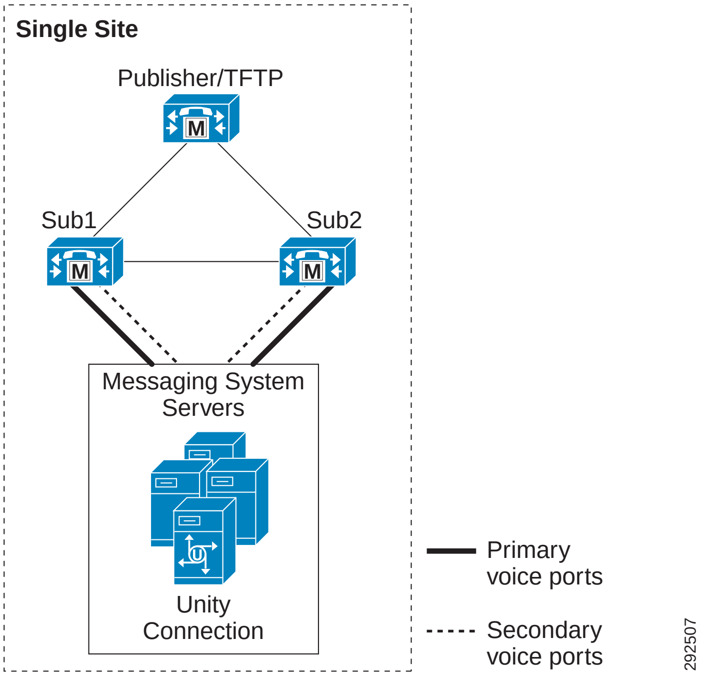
The following guidelines apply to eMWI:

- All clusters should support eMWI. If an intermediate cluster does not support eMWI, then the terminating cluster will receive a standard MWI only without voicemail counts.
- Standard MWI does not generate much traffic because it sends only a change of lamp state (ON or OFF). However, enabling eMWI can increase the amount of traffic because it also sends message counts from the messaging system. The amount of traffic depends on the number of messages and change notifications.

Voice Port Integration with a Unified CM Cluster

When deploying Cisco Unity Connection in a single-site messaging environment, integration with the Unified CM cluster occurs through the SCCP voice ports or SIP trunks. Design considerations must include proper deployment of the voice ports among the Unified CM subscribers so that, in the event of a subscriber failure (Unified CM failover), users and outside calls can continue to access voice messaging. (See Figure 21-18.)

Figure 21-18 Cisco Unity Connection Server(s) Integrated with a Unified CM Cluster (No Dedicated Backup Servers)



The Unified CM cluster in Figure 21-18 employs 1:1 server redundancy and 50/50 load balancing. During normal operations, each subscriber server is active and handles up to 50% of the total server call processing load. In the event of a subscriber server failure, the remaining subscriber server takes up the load of the failed server.

This configuration uses two groups of voicemail ports, with each group containing one-half of the total number of licensed voice ports. One group is configured so that its primary server is Sub1 and its secondary (backup) server is Sub2. The second group is configured so that Sub2 is the primary server and Sub1 is the backup.

Make sure that MWI-only ports or any other special ports are equally distributed between the two groups. During the configuration of the voice ports, pay special attention to the naming convention. When configuring the two groups of ports in Cisco Unity Connection, make sure that the device name prefix is unique for each group and that you use the same device name when configuring the voicemail ports in Unified CM Administration. The device name prefix is unique for each group of ports in this example, with group Sub1 using CiscoUM1 as the device name prefix and Sub2 using CiscoUM2 in this example.

For additional design information on the ratio of inbound to outbound voicemail ports (for MWI, message notification, and TRaP), refer to the *Cisco Unity Connection System Administration Guide* available at http://www.cisco.com.

Note The device name prefix is unique for each group of ports and must match the same naming convention for the voicemail ports configured in Unified CM Administration.

In Unified CM Administration, half of the ports in this example are configured to register using the unique device name prefix of CiscoUM1, and the other half are configured to register using the unique device prefix CiscoUM2. (See Table 21-6.) When the ports register with Unified CM, half will be registered with subscriber server Sub1, and the other half will be registered with Sub2, as shown in Table 21-6.

Table 21-6 Voicemail Port Configuration in Unified CM Administration

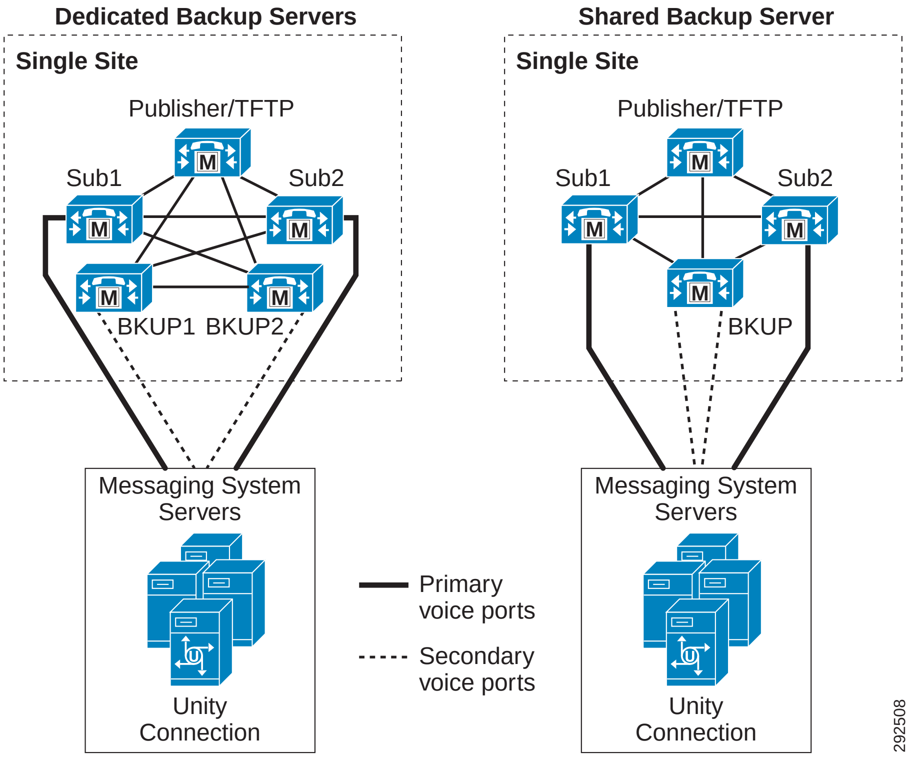
Device Name	Description	Device Pool	SCCP Security Profile	Status	IP Address
CiscoUM1-VI1	Unity Connection 1	Default	Standard Profile	Registered with sub1	1.1.2.9
CiscoUM1-VI2	Unity Connection 1	Default	Standard Profile	Registered with sub1	1.1.2.9
CiscoUM1-VI3	Unity Connection 1	Default	Standard Profile	Registered with sub1	1.1.2.9
CiscoUM1-VI4	Unity Connection 1	Default	Standard Profile	Registered with sub1	1.1.2.9
CiscoUM2-VI1	Unity Connection 1	Default	Standard Profile	Registered with sub2	1.1.2.9
CiscoUM2-VI2	Unity Connection 1	Default	Standard Profile	Registered with sub2	1.1.2.9
CiscoUM2-VI3	Unity Connection 1	Default	Standard Profile	Registered with sub2	1.1.2.9
CiscoUM2-VI4	Unity Connection 1	Default	Standard Profile	Registered with sub2	1.1.2.9

Note The naming convention used for the voicemail ports in Unified CM Administration must match the device name prefix used in Cisco UTIM, otherwise the ports will fail to register.

Voice Port Integration with Dedicated Unified CM Backup Servers

This Unified CM cluster configuration allows each subscriber server to operate at a call processing load higher than 50%. Each primary subscriber server has either a dedicated or shared backup server. (See Figure 21-19.) During normal operation, the backup server processes no calls; but in the event of failure or maintenance of a Subscriber server, the backup server will then take the full load of that server.

Figure 21-19 Cisco Unity Connection Server(s) Integrated with a Single Unified CM Cluster with Backup Subscriber Server(s)



Configuration of the voicemail ports in this case is similar to the 50/50 load-balanced cluster. However, instead of configuring the voice ports to use the opposite subscriber server as the secondary server, the individual shared or dedicated backup server is used. In the Unified CM cluster with a shared backup server, both of the secondary ports for the subscriber servers are configured to use the single backup server.

The voice port names (device name prefix) must be unique for each Cisco UTIM group and must be the same as the device names used on the Unified CM server.

To configure the voicemail ports on Cisco Unity Connection, use the Telephony Integration section of the Unity Connection Administration console. For details, refer to the Cisco Unity Connection administration guides available at http://www.cisco.com.

IPv6 Support with Cisco Unity Connection

The current requirements for IP addressing are surpassing the available set of IP address with IPv4, the current version of IP addressing. Therefore, most IP-based solutions are moving toward incorporating support for IPv6, which provides many more available IP addresses than IPv4. Cisco Unity Connection supports IPv6 addressing with Cisco Unified Communications Manager system integrations through SCCP or SIP. At a component level, dual-stack addressing (both IPv4 and IPv6) is supported over call control and media only.

Cisco Unity Connection supports following IPv6 address types:

- Unique Local Address
- Global Address

Note Voice messages are stored as .wav files and are independent of IPv6 or IPv4.

IPv6 support is disabled by default, but system administrators can enable IPv6 and configure IPv6 address settings either in Cisco Unified Operating System Administration or in the command line interface (CLI). Cisco Unity Connection can obtain an IPv6 address either through router advertisement, through DHCP, or from addresses configured manually either in Cisco Unified Operating System Administration or through the CLI. Cisco Unity Connection Administration and Cisco Personal Communications Assistant can be accessed using IPv6 addresses.

Note IPv6 addressing cannot be enabled during installation or upgrade of Cisco Unity Connection. Cisco Unity Connection does not support "IPv6 ONLY" server configuration. Cisco Unity Connection supports Unicast only for IPv6.

Cisco Unity Connection over IPv6 supports following functionality:

- Cisco Unity Connection offers auto-discovery functionality over IPv6, which allows Unity Connection to search for Microsoft Exchange servers to communicate with.
- Cisco Unity Connection can be integrated with an IPv6 Microsoft Exchange 2007 or 2010 server to enable the Single Inbox feature.
- Cisco ViewMail for Outlook (VMO) supports communication between Outlook and Cisco Unity Connection over IPv6.
- Voice messages received on Cisco Unity Connection can be accessed using any IMAP client such as Outlook over IPv6.
- Cisco Unity Connection can be integrated with LDAP over IPv6 to import the user information.
- Cisco Unity Connection also offers Telephone Record and Playback (TRaP) functionality over IPv6, which enables users to record or play back messages over an IPv6-enabled phone so that signaling can happen over IPv6.

Note Cisco Unity Connection does not provide support for dual-stack addressing mode (both IPv4 and IPv6) in Cisco Business Edition 3000 and 5000.

Single Inbox with Cisco Unity Connection

Cisco Unity Connection supports the Single Inbox feature with Microsoft Exchange 2003, 2007, and 2010 (clustered or non-clustered), thereby providing Unified Messaging for voicemail. Cisco Unity Connection can support all three of these Microsoft Exchange versions simultaneously or any one of them separately. Unity Connection also supports interoperability with the Microsoft Business Productivity Online Suite (BPOS)-Dedicated Services and Microsoft Office 365 cloud-based exchange server. Unity Connection uses Microsoft Exchange Online to enable the Single Inbox feature. For more information, refer to the latest version of the *Unified Messaging Guide for Cisco Unity Connectio*n, available at

http://www.cisco.com/en/US/partner/products/ps6509/prod_maintenance_guides_list.html

All voice messages, including those sent from Cisco Unity Connection ViewMail for Microsoft Outlook, are first stored in Cisco Unity Connection and are immediately replicated to the Microsoft Exchange mailbox for the recipient; however, replication is optional. Also, this feature can be configured per individual user.

Cisco Unity Connection support of Unified Messaging for voicemail involves several design considerations. The user’s email becomes a single container for all messages, including email and voicemail. If a message is moved to any other folder under the user’s Inbox, it will continue to show up in Cisco Unity Connection. However, if the user moves voice messages into Outlook folders that are not under the Inbox folder, the messages are deleted from Cisco Unity Connection but they can still be played by using ViewMail for Outlook because a copy still exists in the Outlook folder. If the user moves the messages back into the Inbox folder or into a folder under the Inbox folder, the message is synchronized back into the Cisco Unity Connection mailbox for that user. In addition, when a user deletes a voice message from Cisco Unity Connection or when Cisco Unity Connection automatically deletes a voice message because of message aging, the message is also deleted from Microsoft Exchange. Likewise, when a voice message is deleted from Microsoft Exchange, it is also deleted from Cisco Unity Connection.

If a message is marked as secured and private, the actual message is not replicated in Microsoft Exchange; instead, a placeholder with a brief description is created for the message. The only copy of actual message stays on Cisco Unity Connection an the user retrieves the message, it is played back from Cisco Unity Connection directly instead of from a local source, unlike in the case of a normal message. This also means that there is no local access to the audio file if it is accessed through voicemail from Outlook. Movement of the secure and private message to any folder other than Inbox and folders below Inbox would result in deletion of the message permanently, thereby leaving no opportunity for retrieval.

Note All voice messages remain on the Cisco Unity Connection Server regardless of the type of messaging deployment. Cisco Unity Connection is the authoritative source of voice messaging traffic, notifications, and synchronizations.

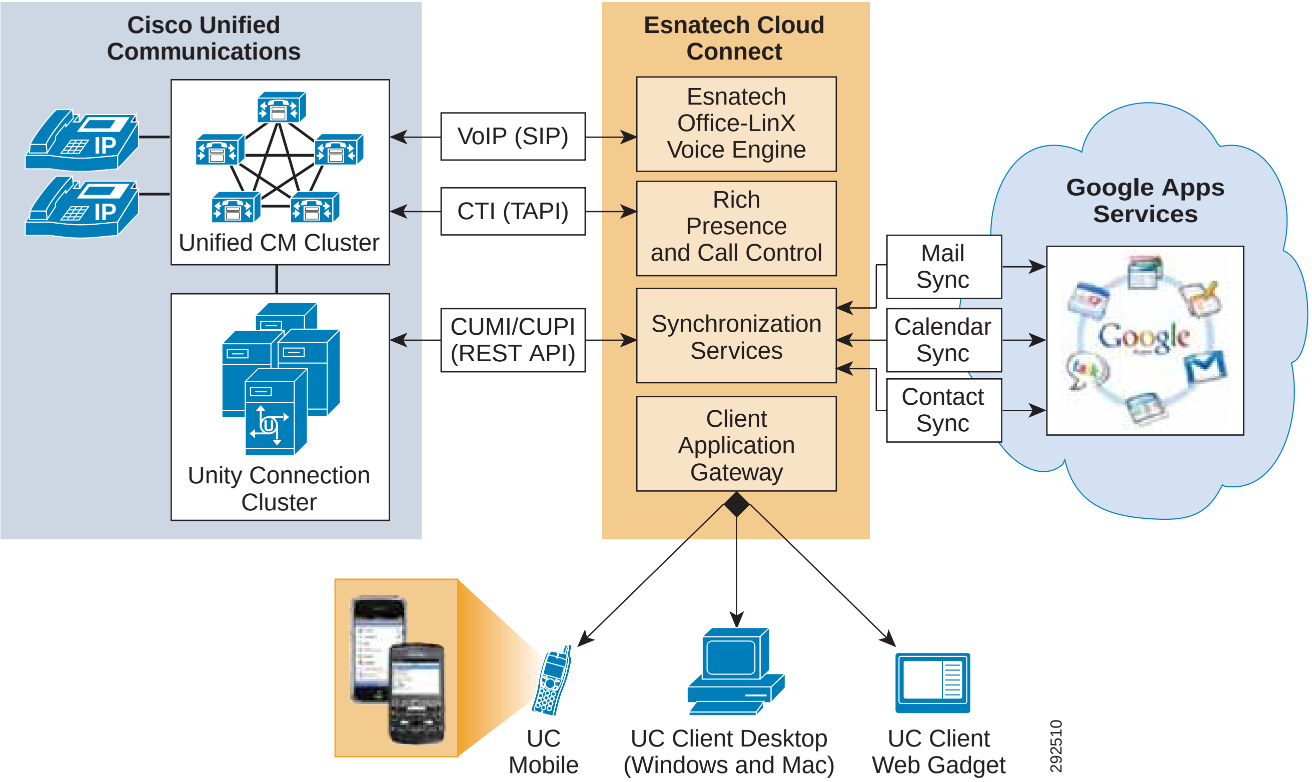
The amount of space a single voicemail message can acquire is configured on the Cisco Unity Connection server and is similar to message aging. The maximum size for a voicemail message is also configured on the Microsoft Exchange Server. Typically, the Microsoft Exchange Server maintains a larger size than Cisco Unity Connection that is synchronized to the mailbox. Hence, the minimum size of the message in Microsoft Exchange should be bigger than the maximum size in Cisco Unity Connection.

From a security aspect for communications between Cisco Unity Connection and Microsoft Exchange, HTTPS is chosen as the default option. HTTP is also supported but not recommended because it reduces security and might also need further configuration on Microsoft Exchange. At the same time, there is an option to validate the Microsoft Exchange certificate, provided that access to the certificate server is available.

Cisco Unity Connection can be integrated with IBM Lotus Domino using software from Cisco partners Esnatech and Donoma to enable the Single Inbox feature. Esnatech Office-LinX^{TM} Cloud Connect Edition and Donoma Unify for Lotus Notes both enable integration between Unity Connection and IBM Lotus Domino. For more detail information on deployment, installation, and configuration of these products, refer to the documentation available at http://www.esnatech.com/landing/cisco.htm and http://donomasoftware.com/donoma-unify-for-lotus-notes/.

Cisco Unity Connection also supports integrated voice and fax services with cloud-based applications such as Google Apps Gmail and VMware Zimbra. Unity Connection can be integrated with these email applications using Esnatech Office-LinX^{TM} Cloud Connect Edition. This solution allows for bidirectional synchronization of voice messages between Cisco Unity Connection and these email applications. (See Figure 21-20.)

Figure 21-20 Cisco Unity Connection Deployment with Cloud-Based Google Email Application



Esnatech Cloud Connect Edition uses the Cisco Unity Connection Messaging Interface (CUMI) API and the Cisco Unity Connection Provisioning Interface (CUPI) API for synchronization of subscriber and messaging information. Users can also initiate calls through Gtalk using the Click-to-Dial feature.This call control information gets exchanged with Cisco Unified Communications Manager through the CTI (TAPI) interface. For more detail information on deployment, installation and configuration, refer to the documentation available at http://www.esnatech.com/landing/cisco.htm.

Best Practices for Deploying Cisco Unity Express

When deploying Cisco Unity Express, use the following guidelines and best practices:

- Ensure that the IP phones having Cisco Unity Express as their voicemail destination are located on the same LAN segment as the router hosting Cisco Unity Express.
- If uninterrupted automated attendant (AA) and voicemail access is required for a site deployed with Cisco Unity Express, ensure that Cisco Unity Express, SRST, and the PSTN voice gateway are all located at the same physical site. Hot Standby Router Protocol (HSRP) or other redundant router configurations are not currently supported with Cisco Unity Express.
- Each mailbox can be associated with a primary extension number and a primary E.164 number. Typically, this number is the direct-inward-dial (DID) number that PSTN callers use. If the primary E.164 number is configured to any other number, use Cisco IOS translation patterns to match either the primary extension number or primary E.164 number so that the correct mailbox can be reached during SRST mode.

Voicemail Integration with Unified CM

- Each Cisco Unity Express site must be associated with a CTI route point for voicemail and one for AA (if licensed and purchased), and you must configure the same number of CTI ports as Cisco Unity Express ports licensed. Ensure that the number of sites with Cisco Unity Express does not exceed the CTI scalability guidelines presented in the chapter on Call Processing, page 8-1.
- Cisco Unity Express is associated with a JTAPI user on Unified CM. Although a single JTAPI user can be associated with multiple instances of Cisco Unity Express in a system, Cisco recommends associating each dedicated JTAPI user in Unified CM with a single Cisco Unity Express.
- If Unified CM is upgraded from a previous version, the password of the JTAPI user automatically gets reset on Unified CM. Therefore, after the upgrade, the administrator must make sure that the JTAPI password is synchronized between Cisco Unity Express and Unified CM so that Cisco Unity Express can register with Unified CM.
- The CTI ports and CTI route points can be defined in specific locations. Cisco recommends using locations-based call admission control between Unified CM and Cisco Unity Express. RSVP may also be used.
- Ensure proper Quality of Service (QoS) and bandwidth for signaling traffic that traverses the WAN between Cisco Unity Express and Unified CM. Provision 20 kbps of bandwidth for CTI-QBE signaling for each Cisco Unity Express site. See the chapter on Network Infrastructure, page 3-1, for more details.
- The CTI-QBE signaling packets from Unified CM to Cisco Unity Express are marked with a DSCP value of AF31 (0x68). Unified CM uses TCP port 2748 for CTI-QBE signaling.
- The Unified CM JTAPI library sets the proper IP Precedence bits in all outgoing QBE signaling packets. As a result, all signaling between Cisco Unity Express and Unified CM will have the proper QoS bits set.

Cisco Unity Express Codec and DTMF Support

Calls into Cisco Unity Express use G.711 only. Cisco recommends using a local transcoder to convert the G.729 calls traversing the WAN into G.711 calls. You can configure Unified CM regions with the G.711 voice codec for intra-region calls and the G.729 voice codec for inter-region calls.

If transcoding facilities are not available at the Cisco Unity Express site, provision enough bandwidth for the required number of G.711 voicemail calls over the WAN. Configure the Unified CM regions with the G.711 voice codec for calls between the IP phones and Cisco Unity Express devices (CTI ports and CTI route points).

Cisco Unity Express does not support in-band DTMF tones; it supports only DTMF relay. With Cisco Unity Express, DTMF is carried out-of-band via either the SIP or JTAPI call control channels. Cisco Unity Express 2.3 supports G.711 SIP calls with RFC 2833 into Cisco Unity Express.

JTAPI, SIP Trunk and SIP Phone Support

Cisco Unified CM supports SIP trunking protocol; however, Cisco Unity Express uses JTAPI to communicate with Unified CM. Cisco Unity Express supports both SCCP and SIP phones.

- Configure a SIP trunk for SRST and Unified CM for support of SIP phones (through JTAPI).
- Cisco Unity Express supports G.729 SIP calls via a transcoder, with the ability added in Cisco IOS Release 12.3(11)XW for RFC 2833 to pass through a transcoder.
- Cisco Unity Express supports delayed media (no SDP in the INVITE message) for call setup in case of a slow-start call from Unified CM.
- Cisco Unity Express supports both blind and consultative transfer, but the default transfer mode is consultative transfer (semi-attended) using REFER in SIP calls. Use the Cisco Unity Express command line interface to explicitly change the transfer mode to consultative transfer using REFER or blind transfer using BYE/ALSO. If REFER is not supported by the remote end, BYE/ALSO will be used.
- Cisco Unity Express supports outcall for voice message notifications. It also supports consultative transfers. During both of these call setups, Cisco Unity Express can receive 3*xx* responses to the INVITE. Cisco Unity Express processes only 301 (Moved Permanently) and 302 (Moved Temporarily) responses to the INVITE. This requires the URL from the Contact header from the 3*xx* response to be used to send a new INVITE. 305 (Use Proxy) responses are not supported.

Note For compatibility between Cisco Unified CM and Cisco Unity Express, refer to the *Cisco Unity Express Compatibility Matrix*, available at http://www.cisco.com/en/US/docs/voice_ip_comm/unity_exp/compatibility/cuecomp.htm.

For more information about Cisco Unity Express, refer to the product documentation available at http://www.cisco.com.

Third-Party Voicemail Design

This section discusses various options for deploying third-party voicemail systems with Cisco Unified Communications, and it covers both integration and messaging.

Note This section does not discuss how to size a third-party voicemail system for ports and/or storage. For this type of information, contact your voicemail vendor, who should be better able to discuss the individual requirements of their own system, based upon specific traffic patterns.

Integration

Integration is defined as the physical connection between a voicemail system and its associated PBX or call processing agent, and it also provides for the feature set between the two.

There are many voicemail vendors, and it is not uncommon for customers to want to continue to use an existing voicemail system when deploying Cisco Unified CM. With this requirement in mind, Cisco provides support for the industry standard voicemail protocol known as Simplified Message Desk Interface (SMDI). SMDI is a serial protocol that provides all the necessary call information required for a voicemail system to answer calls appropriately and is probably the most common method deployed for voicemail integration between dissimilar systems from various vendors.

Note Cisco does not test or certify any third-party voicemail systems. Within the industry, it is generally considered to be the responsibility of the voicemail vendor to test and/or certify their products with various PBX systems. Cisco does, of course, test its interfaces to such equipment and will support these interfaces regardless of which third-party voicemail system is connected.

An alternative to SMDI for voicemail integration is QSIG, which also allows a third-party PBX to connect to Unified CM through a Primary Rate Interface (PRI) T1/E1 trunk. Each method has its own advantages and disadvantages, and the method you employ will largely depend on how your voicemail system is integrated to your current PBX.

There are other methods for connecting voicemail systems to Unified CM (such as PRI ISDN trunks in conjunction with SMDI), but these methods are uncommon.

Today there are other potential methods of voicemail integration, such as H.323 or SIP. However, due to the varying methods of vendor implementation, features supported, and other factors, these third-party voicemail integrations will have to be evaluated on a per-customer basis. Customers are advised to contact their Cisco Account Team and/or Cisco Partner to discuss these options further.

Messaging

Messaging is defined as the exchange of messages between voicemail systems, and there are several open standards for this purpose.

The most common protocol deployed to allow messaging between dissimilar systems is Voice Profile for Internet Mail (VPIM). VPIM has seen several updates to its specification, and although Version 2 is not the latest, it still appears to be the most widely adopted. The messaging protocol prior to VPIM is Audio Messaging Interchange Specification - Analog (AMIS-A), and it is fairly rare in its adoption due mainly to its cumbersome user interface as well as the analog technology it employs and its lack of features.

CHAPTER 22

Cisco Collaborative Conferencing

Revised: April 30, 2013; OL-27282-05

Cisco offers a wide range of collaboration technologies that have the ultimate goal of allowing users to work in virtual collaborative environments that result in faster, more efficient decision-making processes and increased productivity. There are many technologies that fall under the large collaboration umbrella, but this chapter focuses on design guidance surrounding the Cisco offerings that allow for simultaneous communications through audio, video, and rich content sharing capabilities. This chapter also explores the differences in the various solutions and provides suggestions on when one solution may be a better fit than another.

Certain aspects are common to all the Cisco collaborative conferencing solutions. For instance, the capability to integrate with scheduling or calendaring systems so that the creation of a meeting is familiar and intuitive to users. Ties into LDAP directories for inviting attendees in the organization and a consistent authentication method are also critical. Users have the ability to host and attend virtual meetings, whether in the office or outside of the enterprise, to ensure continued productivity for users even when they are traveling outside the organization.

The Cisco collaborative conferencing solutions discussed in this chapter are available as on-premises, off-premises, or mixed deployments. This allows an organization to integrate with a Unified Communications solution in which they have already invested or, alternatively, to implement a service that is hosted "in the cloud." This is one of the more important distinctions between the various solutions, and it is the first decision point when determining which solution is the best fit for an organization. This chapter contains sections on the following topics:

- Cisco WebEx Software as a Service (SaaS)
- Cisco WebEx Meetings Server for private cloud
- Cisco Unified MeetingPlace

Each section defines the high-level architecture of the solution, followed by design guidance for high availability, capacity planning and other design considerations pertinent to the solution.

For more detail on the various Cisco collaborative client offerings and how they fit into collaborative conferencing solutions, see the chapter on Cisco Collaboration Clients and Applications, page 24-1.

What's New in This Chapter

This chapter incorporates new material to bring together design discussions surrounding Cisco's collaborative conferencing offerings. If you are reading this chapter for the first time, Cisco recommends reading the entire chapter.

Table 22-1 lists the topics that are new in this chapter or that have changed significantly from the previous release of this document.

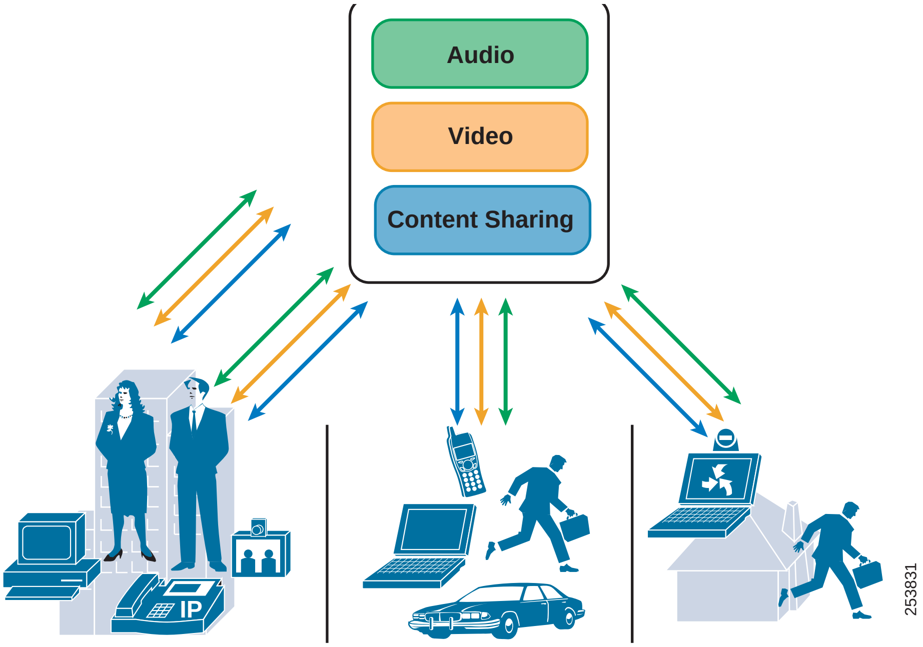
Table 22-1 New or Changed Information Since the Previous Release of This Document

New or Revised Topic	Described in:	Revision Date
Call control bandwidth for WebEx Meetings Server	Network Traffic Planning, page 22-19	April 30, 2013
Cisco WebEx Meetings Server	Cisco WebEx Meetings Server, page 22-13	October 31, 2012
Detailed capacity planning information has been moved to the chapter on *Unified Communications Design and Deployment Sizing Considerations*.	Unified Communications Design and Deployment Sizing Considerations, page 29-1	June 28, 2012
Cisco Unified Videoconferencing products have reached End-of-Sale (EoS) and End-of-Life (EoL) and are no longer covered in this document.	Refer to the EoS and EoL notices available at http://www.cisco.com/en/US/products/ps10463/prod_eol_notices_list.html	June 28, 2012
Other minor updates for Cisco Unified Communications System Release 9.0	Various sections throughout this chapter	June 28, 2012

Collaborative Conferencing Architecture

At a high level, collaborative conferencing involves receiving audio, video, and content from some or all of the attendees in a meeting, mixing those streams, and then sending the mixed audio, video, and content back to the attendees. Figure 22-1 illustrates a logical conference involving both internal and external participants, mobile and remote workers, or even attendees from other organizations.

Figure 22-1 Logical View of Collaborative Conferencing



These three aspects of a collaborative conference – audio, video and content sharing – are not exclusive. Cisco collaborative conferencing solutions integrate the three to create an enhanced user experience. Features such as the ability to determine active speakers, muting users from the content share interface, or choosing the video layout displayed in the content share, all imply that these three elements are integrated by the solution. All the collaborative conferencing solutions discussed in this chapter use the Cisco WebEx interface for content sharing. This provides a very consistent user experience across all the solutions.

When considering which solution is best for a given organization, many factors should be evaluated. Characteristics of an organization's users (number of remote workers, access capabilities, and video usage) as well as the range of available endpoints and their capabilities are important to consider. Video requirements such as high definition or interworking with an existing video infrastructure can also dictate a solution. The nature of the meetings themselves (for example, training scenarios, collaborative meetings, or how many meeting participants are external to the organization) is a critical characteristic to identify. Of course, initial cost, maintenance costs, and return on investment (ROI) all play a role as well.

One of the first delineations between the solutions is whether the resources performing each type of conferencing (or mixing) are located on-premises or off-premises. Access to cloud services, the size of the mobile workforce, and support staff levels are all considerations. Cisco WebEx software as a service (SaaS) offers a completely off-premises solution with an option to extend the cloud on-premises, while

Cisco Unified MeetingPlace is a hybrid (mix of on-premises and off-premises) with the option to pull the majority of resources on-premises. Organizations that have deployed Cisco Unified Communications will benefit most from leveraging an on-premises solution. The later sections in this chapter provide more detailed deployment options for each solution.

This document describes several approaches to providing a high-performance collaboration solution. These solutions can be broadly categorized as:

- Cloud-based (SaaS) service with on-premises acceleration
- On-premises solution with cloud-based augmentation

Table 22-2 summarizes available solutions from an on-premises cloud perspective.

Table 22-2 On-Premises, Cloud, and Hybrid Capabilities of Cisco Collaborative Solutions

	Audio		**Video**		**Content Sharing**	
Solution	On-premises	Cloud	On-premises	Cloud	On-premises	Cloud
Cisco WebEx Meetings Server	Yes	No	Yes[1]	No	Yes	No
Cisco WebEx SaaS	No	Yes	No	Yes[1]	No	Yes
Cisco WebEx SaaS with Cisco WebEx Node for Aggregation Services Router (ASR)	Yes (VoIP)	Yes	Yes	Yes[1]	Yes[2]	Yes
Cisco Unified MeetingPlace with WebEx SaaS[3]	Yes	No	Yes	No	No	Yes
Cisco Unified MeetingPlace with Cisco WebEx Node for Cisco MCS[3]	Yes	No	Yes	Yes	Yes[2]	Yes
Cisco Unified MeetingPlace with Cisco WebEx Node for Cisco ASR[3]	Yes	Yes	Yes	Yes[1]	Yes[2]	Yes
Cisco Unified MeetingPlace (audio/video only deployment)	Yes	No	Yes	No	No	No

[1] Cisco WebEx webcam video only and no support with standards-based video.

[2] Cisco WebEx Node for ASR and MCS requires a connection to the Cisco WebEx network.

[3] Cisco Unified MeetingPlace solutions may alternatively use the WebEx webcam video streaming capabilities of the cloud. However, Cisco does not recommend using both because there is no interoperability.

Cisco WebEx Software as a Service

Cisco WebEx is a collaboration solution that does not require any hardware to be deployed on-site. All services (audio, video, and content sharing) are hosted in the Internet or the cloud. This is often referred to as software-as-a-service (SaaS). Meetings can be initiated and attended from anywhere, anytime, and do not require connectivity back into the enterprise. This section describes solution characteristics and provides design guidance for deploying WebEx SaaS.

With respect to scheduling and initiating meetings, WebEx provides cloud-based web scheduling capability, but most organizations prefer to schedule from their corporate email system (Exchange, Lotus Notes, and so forth) or other enterprise applications. The WebEx Productivity Tools is a bundle of integrations with well known desktop tools incorporated into a single application. A WebEx administrator can control the specific integrations that are provided through the tool to their organization's user population. It can be installed automatically when accessing the WebEx sitename, or it can be pushed out locally using standard desktop management tools. For more information on WebEx Productivity Tool, refer to the WebEx *Productivity Tools FAQs*, available at

https://vnc.WebEx.com/docs/T26L/pt/mc0800l/en_US/support/productivitytools_faq.htm

There are three methods of creating WebEx user profiles for an organization in the cloud. Security considerations for the actual usernames and passwords, as well as for handling a large number of user accounts, should be considered. A WebEx administrator can create user profiles manually by bulk import of a CSV template or by a programmatic approach. A programmatic approach uses one or a combination of the WebEx APIs, URL, and XML, or a Federated SSO solution. The programmatic approach can be used by a customer portal, which is an application such as a CRM tool or a Learning Management System that integrates directly into WebEx. For more information regarding WebEx directory integration and authentication, refer to the WebEx *Approaches to Single Sign-On Developer Technical Note*, available at

http://developer.WebEx.com/c/document_library/get_file?folderId=11421&name=DLFE-213.pdf

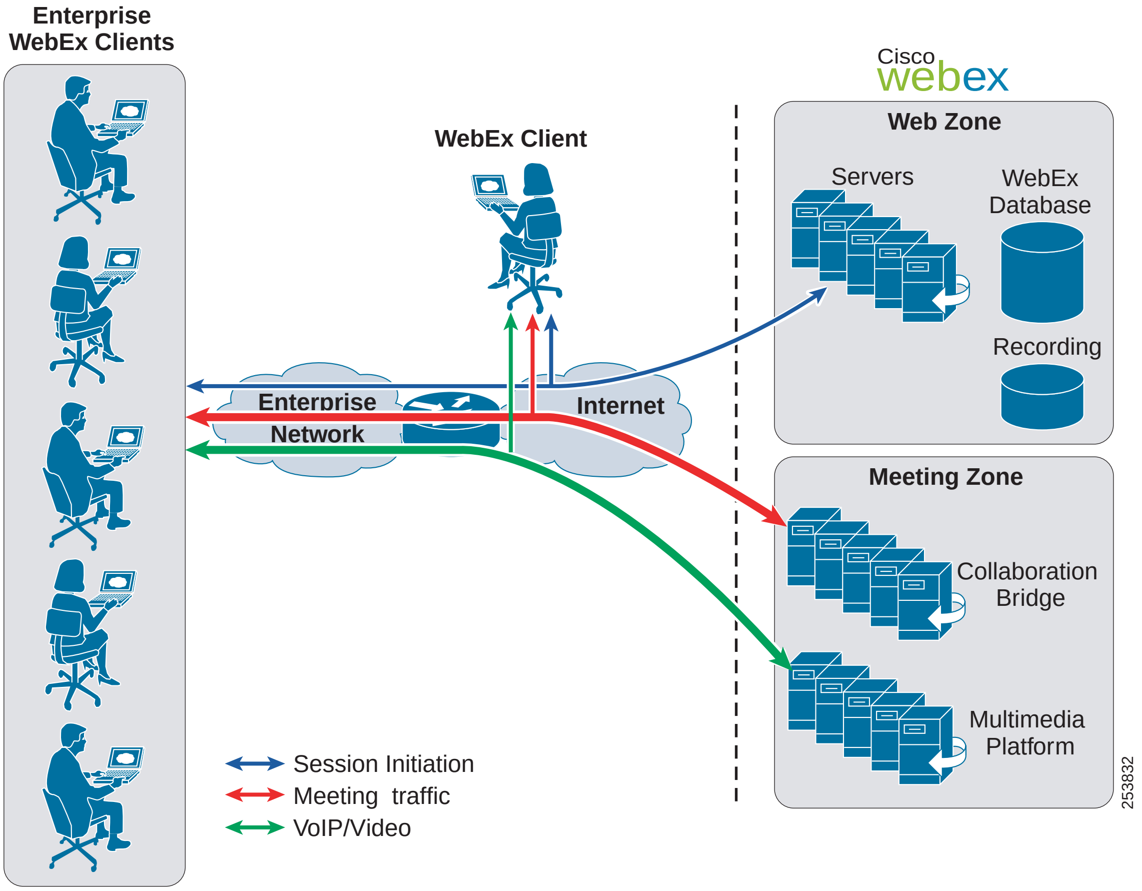
For integrating directly with an organization's LDAP directory, Federated SSO with Security Assertion Markup Language (SAML) is the preferred approach. For more information regarding Federated SSO, refer to the WebEx *Federated SSO Authentication Service Technical Overview*, available at

http://developer.WebEx.com/c/document_library/get_file?folderId=11421&name=DLFE-201.pdf

Architecture

An organization's IT department needs to understand the architecture of the Cisco Collaboration cloud-based solution. In the traditional WebEx deployment model shown in Figure 22-2, all the content, voice, and video traffic from every client traverses the internet and is mixed and managed in the cloud at the WebEx data center. The WebEx data center is logically divided into the Meeting Zone and the Web Zone. The Web Zone is responsible for things that happen before and after a web meeting. It incorporates tasks such as scheduling, user management, billing, reporting, and streaming recordings. The Meeting Zone is responsible for switching the actual meeting once it is in progress between the endpoints.

Figure 22-2 Traditional WebEx Deployment



The Meeting Zone consists of two subsystems. Within the Meeting Zone there are collaboration bridges that switch meeting content. The multimedia platform is responsible for mixing all of the VoIP and video streams within a meeting. To join a WebEx session, an attendee first connects to the Web Zone. The Web Zone traffic flows only before or after the meeting, is relatively low bandwidth, and is mainly non-real time. The real-time meeting content share flows to and from the Meeting Zone and can be bandwidth intensive. Its real-time nature can place a heavy burden on enterprise access infrastructure. For further details regarding network traffic planning, see Capacity Planning, page 22-9.

By default, all WebEx meeting data is encrypted using 128-bit SSL encryption between the client and Cisco's Collaboration Cloud. SSL accelerators within the cloud decrypt the content sharing information and send it to a WebEx conference bridge that processes the content and sends it back through an SSL accelerator, where it is re-encrypted and sent back to the attendees. All Web Zone and Meeting Zone traffic is encrypted using 128-bit SSL where SSL accelerators are used to off-load the SSL function from the Web and Meeting Zone servers.

After the meeting ends, no session data is retained in the WebEx cloud or an attendee's computer. Only two types of data are retained on a long-term basis: billing and reporting information and optionally network based recordings, both of which are accessible only to authorized enterprise users.

Some limited caching of meeting data is carried out within the Meeting Zone, and this is done to ensure that users with connectivity issues or who may be joining the meeting after the start time receive a current fully synchronized version of the meeting content.

Independent third parties are used to conduct external audits covering both commercial and governmental security requirements, to ensure the WebEx cloud maintains its adherence to documented security best practices. WebEx performs an annual SAS-70 Type II audit in accordance with standards established by the AICPA, conducted by Pricewaterhouse Cooper. The controls audited against WebEx are based on ISO-17799 standards. This highly respected and recognized audit validates that WebEx services have been audited in-depth against control objectives and control activities (that often include controls over information technology and security related processes) with respect to handling and processing customer data.

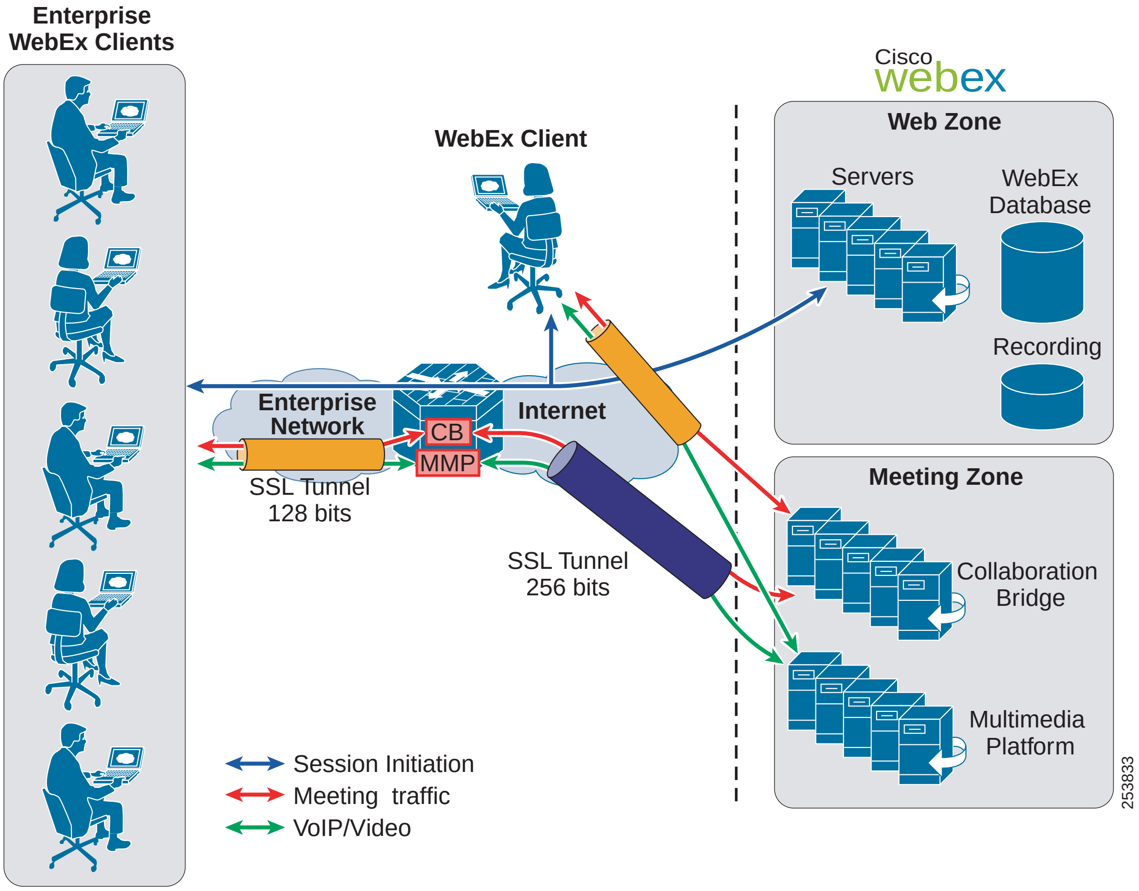
For customers that require enhanced security, there is also an option to perform end-to-end 256 bit AES encryption for collaboration bridge and multimedia content so that traffic is never decrypted in the cloud. In addition, PKI identity validation support is optionally available to further enhance the end-to-end AES encryption. End-to-end encryption results in some lost features such as NBRs. For more information on enhanced WebEx security options, refer to the *Security Overview of Cisco WebEx Solutions* available at

http://static.WebEx.com/fileadmin/WebEx09/files_en_us/pdf/whitepapers/cwe_securityoverview.pdf

Note Enhanced WebEx security options are available only for Meeting Center meetings. The WebEx security options come at no additional cost.

Starting with Cisco WebEx release WBS27, an organization can optionally accelerate WebEx meeting traffic using the WebEx Node for Aggregation Services Router (ASR) 1000 Series. Using a WebEx Node for ASR (a blade installed in the router), key components from the cloud can be extended onto a platform that resides on-premises within the enterprise, as shown in Figure 22-3. This moves an instance of the collaboration bridge and the multimedia platform onto the ASR, which provides performance and bandwidth improvements over a pure cloud-based solution. This is a fully cascaded solution that allows attendees within the enterprise to connect to the Node and external attendees to connect to the cloud. Failover and overflow from the Node(s) to the cloud are fully supported and transparent in operation. The WebEx Node's operation is unapparent to both the user and the WebEx site administrator. The WebEx Node for ASR works with standalone WebEx SaaS accounts and with MeetingPlace 8.5 Audio on-premises.

Figure 22-3 WebEx Deployment with WebEx Node for ASR



When an attendee joins a WebEx meeting, the Web Zone in the WebEx cloud serves the client entry page and tells the WebEx client where to connect. The clients always get passed the list of cloud-based Meeting Zones available for the meeting, represented as URLs. If WebEx Nodes for ASR have been provisioned for the organization's WebEx site, the node hostnames are also included in the list of Meeting Zones. The clients then ping all of the cloud and on-premises resources to determine which Meeting Zone instance is closest in terms of latency. Because the on-premises nodes are available through the corporate network, they should respond first, and the on-premises client connect to these resources. Clients also connect to the node using 128 bit SSL encryption. The nodes provide support for Meeting Center, Event Center, Training Center, and Support Center.

Note When deployed in multimedia mode, the WebEx Node for ASR is capable of mixing VoIP (from the WebEx client itself) and webcam video. Mixed Mode Audio involves PSTN callers and is always mixed in the cloud.

Comparing Figure 22-3 with the traditional WebEx deployment model depicted in Figure 22-2 indicates that session initiation still takes place in the Web Zone within the cloud, but the enterprise WebEx clients are using a conference bridge or multimedia platform in the WebEx Node in an ASR on the enterprise network, which saves internet bandwidth and improves performance. The WebEx Node for ASR cascades control traffic and meeting content or VoIP and video content back to the cloud over an SSL tunnel. This allows external participants to access the meeting and to support network based recording (NBR). The SSL tunnel is built when the WebEx node is started and all the connections are made outbound from the enterprise to the WebEx cloud.

Note A WebEx Node for ASR can be configured to act as either a content bridge or a multimedia node, but it does not support both functions simultaneously. To support both data and multimedia acceleration, a minimum of two WebEx Node blades are required. These can be deployed in the same ASR chassis or different chassis. There is no limit on the number of Nodes that may be deployed within an enterprise network.

For further details regarding network traffic optimization using WebEx Node for ASR, see Capacity Planning, page 22-9.

There is also the potential to deploy the WebEx Node for ASR in a multi-tenant capacity, in which two businesses working closely together with staff working on each other's premises could have the other's WebEx site defined on their ASR Nodes. This means that, when staff for Enterprise B access their company's WebEx site through Enterprise A, they can use the local ASR Node to accelerate their meeting while saving bandwidth for Enterprise A. This feature can also benefit organizations that have multiple WebEx sites.

Starting with Cisco WebEx release WBS27-FR20, Meeting Center uses the H.264 AVC/SVC codec to provide High Quality Video for the conference. Higher network bandwidth is needed for those deployments. For further details regarding network traffic optimization for High Quality Video, see Capacity Planning, page 22-9.

Note Cisco TelePresence integrates with WebEx using OneTouch. For details on Cisco TelePresence WebEx OneTouch, refer to the documentation at http://www.cisco.com/en/US/solutions/ns669/webex_engage.html.

High Availability

The WebEx cloud itself has a very high level of redundancy and is managed by Cisco. With respect to a WebEx Node for ASR, if a Node fails or becomes congested, then user meetings re-connect to the cloud. When clients ping the Meeting Zone URLs, they do not get a response back from the ASR node, therefore they connect to another Meeting Zone. If there are active meetings on a node and the node goes offline, there is a copy of the content cached in the cloud even if all attendees are internal. The WebEx clients reconnect to an alternate Meeting Zone, and the meeting continues with no intervention by the users.

Capacity Planning

For a given customer, the actual number of concurrent meetings is essentially unlimited. Different WebEx conferencing types have different capacities with respect to number of attendees. For a detailed product comparison table, refer to the *Cisco WebEx Web Conferencing Product Comparison*, available at

http://www.cisco.com/en/US/prod/ps10352/product_comparison.html

The capacity of the WebEx Node for ASR depends on the function for which it is implemented. When deployed as a collaboration bridge (web conferencing), the Node supports up to 500 attendees. If a node reaches its maximum attendee limit, a WebEx client either uses an alternative on-premises node or overflows directly to the cloud. There is no limit to the number of ASR nodes deployed, and web conferencing can be cascaded across multiple nodes for redundancy and capacity.

The sizing for the WebEx Node for ASR when used to switch VoIP and video locally is slightly more complex because there are different video and VoIP traffic types that impact the performance of the node to a lesser or greater extent. To help with sizing the node for multimedia conferencing, there is a points system that starts with 11,600 points, and points are decremented from this total according to the type and number of streams that flow through the node. Table 22-3 lists the different types of VoIP and video, and the points they consume. As is the case with the web conferencing version of the node, if a multimedia node runs out of capacity, a WebEx client simply connects to another available ASR node or to the cloud.   This alleviates capacity concerns during unexpected random busy periods that over-utilize a given node's capacity.

Table 22-3 WebEx Node for ASR Points Consumed Per Video or VoIP Type

Integrated VoIP or Video Type	Points per Use	Maximum Capacity If Using a Single Service
Active Video 360p + 5x90p	97	120
Active Video 180p	18	640
Active Video 180p + 6x90p	60	192
VoIP	19	600
Audio broadcast	6	1,933

Active Video means that the active speaker will appear in the main video window, and other attendees will be shown as thumbnail images, with the following resolutions:

- 360p: 640x360 resolution
- 180p: 320x180 resolution
- 90p: 160x90 resolution

Note Multi-point video points are deducted per attendee watching the video panel during a meeting. A maximum of 6 webcam video sessions can be displayed per WebEx client, but each attendee has control over which are shown.

Table 22-3 provides conservative estimates; however, it is difficult to predict usage precisely and to control user behavior. Cisco recommends provisioning enough resources to deal with the average load on the system, allowing for periods of peak usage to overflow to the cloud.

Network Traffic Planning

With the increased traffic out to the internet, it is important to consider network traffic planning. By evolving the WebEx architecture to include on-premises ASR nodes, performance can be optimized and significant savings in Internet access bandwidth can be achieved. Table 22-4 itemizes different traffic types that could load the enterprise network during a WebEx meeting. The only traffic type that is not native to WebEx is IP telephony, which might be used with either an on-premises or off-premises conferencing service integrated with WebEx.

Table 22-4 Bandwidth Estimates for WebEx Meeting Traffic

Traffic (Test Scenario)	Average (kbps)	Maximum (kbps)
Idle meeting: iPad (16G), iPhone (3G), and Blackberry Bold9700) use a WiFi network for data connectivity.	0.8 for PC, 8.9 for iPad, 0.17 for iPhone, 0.42 for Blackberry devices	3.7 for PC, 9 for iPad, 0.4 for iPhone, 0.45 for Blackberry devices
Desktop share (Slide presentation with 30 second transitions)	43 for PC, 95 for iPad, 67 for iPhone, 24.8 for Blackberry devices	598 for PC, 241 for iPad, 232 for iPhone, 29.9 for Blackberry devices
Presentation share (Slide presentation with 5 second transitions)	6.5 for PC, 30 for iPad, 23 for iPhone, 54.56 for Blackberry devices	7.5 for PC, 62 for iPad, 41 for iPhone, 55.28 for Blackberry devices
Video (Webcam with 352x288 resolution at 15 fps)	172	298
Video Standard Quality (6 thumbnail videos, each 160x90 resolution up to 10 fps)	350	500
Video High Quality Medium View (Webcam with 320x180 resolution up to 12 fps)	300	500
Video High Quality Large View (Webcam with 640x360 resolution up to 30 fps)	900	1,500
Video High Definition (Webcam with 1280x720 resolution up to 30 fps)	1,500	2,000

How users actually use WebEx will make quite a bit of difference in the amount of traffic generated by the meeting. For example, if attendees use native presentation sharing (where the document is loaded to the WebEx site prior to sharing), it generates far less data than if they share their desktops. For a large enterprise, this can be important to understand to ensure correct traffic engineering, especially at the choke points in the network, such as the Internet access points. A preliminary estimate should be made around the average number of meetings to be hosted during the busy hour, along with the average number of attendees. Then, depending on the type and characteristics of these meetings, some projections on bandwidth requirements can be made. For more information regarding network traffic planning, please see the *WebEx Network Bandwidth White Paper*, available at

http://www.WebEx.com/pdf/wp_bandwidth.pdf

As discussed, the WebEx Node for ASR can be implemented to pull the collaboration bridge and the multimedia platform engine on-premises. To help quantify the impact of an ASR Node, Table 22-5 and Table 22-6 show some examples of theoretical bandwidth savings. In the examples, fairly large customer deployments have been assumed, each having 1,000 concurrent peak meeting attendees distributed across a number of separate meetings with two different average numbers of attendees for each example. Example 1 uses desktop sharing, while example 2 uses presentation sharing. Both examples result in large reductions in the WebEx traffic bandwidth across the organization's internet access pipes.

Table 22-5 Parameters for Example Bandwidth Calculations

Parameters	Example 1	Example 2
Peak number of attendees	1,000	1,000
Average attendees per meeting	6	10
Percentage of internal attendees	80%	50%
Percentage of internal presenters	90%	90%
Average attendees receiving VoIP	10%	30%
Average attendees receiving video	30%	40%
Average meeting traffic bandwidth	43 kbps	6.5 kbps
Average video traffic bandwidth using 320x180 resolution High Quality Video	300 kbps	300 kbps
Average VoIP bandwidth	35 kbps	35 kbps

Table 22-6 Example Bandwidth Estimate Calculations

	Example 1		Example 2	
Traffic Types	Bandwidth without Node	Bandwidth with Node	Bandwidth without Node	Bandwidth with Node
Average meeting traffic bandwidth	34 Mbps	1 Mbps	22 Mbps	1 Mbps
Average single-point video traffic bandwidth	72 Mbps	15 Mbps	60 Mbps	12 Mbps
Average VoIP bandwidth	3 Mbps	1 Mbps	5 Mbps	1 Mbps

Note The example in Table 22-5 and Table 22-6 assumes that two WebEx Nodes for ASRs are deployed, one in collaboration bridge mode and one in multimedia mode.

Design Considerations

Observe the following design considerations when implementing a Cisco WebEx SaaS solution:

- Collaborative meeting systems typically result in increased top-of-the-hour call processing loads. Cisco partners and employees have access to capacity planning tools with parameters specific to collaborative meetings to help calculate the capacity of the Cisco Unified Communications System for large configurations. Contact your Cisco partner or Cisco Systems Engineer (SE) for assistance with sizing of your system. For Cisco partners and employees, the Cisco Unified Communications Sizing Tool is available at http://tools.cisco.com/cucst.
- The WebEx Node for ASR is typically located in a DMZ because it is serves as an extension of the WebEx cloud and is therefore managed from the cloud. However, there is no requirement for a DMZ, and the Node could be placed anywhere in the network. The WebEx cloud never makes any inbound connections to the Node; rather, secure connections are always initiated from the Node to the cloud on port 443.

- All connections from WebEx clients and WebEx Nodes are initiated out to the cloud. Typically, opening pinholes in network firewalls is not required as long as the firewalls allow intranet devices to initiate TCP connections to the Internet.
- If WebEx High Quality Video is integrated with a third-party audio bridge, video of the presenter will be displayed in the active speaker window rather than video of the active speaker on voice.
- For more details on the various Cisco collaborative client offerings and how they fit into collaborative conferencing solutions, see Cisco Collaboration Clients and Applications, page 24-1.

Cisco WebEx Meetings Server

Cisco WebEx Meetings Server is a highly secure, fully virtualized, private cloud conferencing solution that combines audio, video, and web conferencing in a single solution. Cisco WebEx Meetings Server addresses the needs of today's companies by presenting a comprehensive conferencing solution with all the tools needed for increased employee productivity as well as support for more dynamic collaboration and flexible work styles. Existing customers can build on their investment in Cisco Unified Communications and extend their existing implementation of Cisco Unified Communications Manager to include conferencing using the SIP architecture. In addition, Cisco WebEx Meetings Server leverages many capabilities from Cisco Unified CM to perform its functions; for example:

- Use the SIP trunk connection with Unified CM to conduct teleconferencing
- Utilize Unified CM's SIP trunk secure connection support for secure conferencing
- Integrate with legacy or third-party PBXs through Unified CM
- Leverage Unified CM's dual stack (IPv4 and IPv6) capability to support IPv6

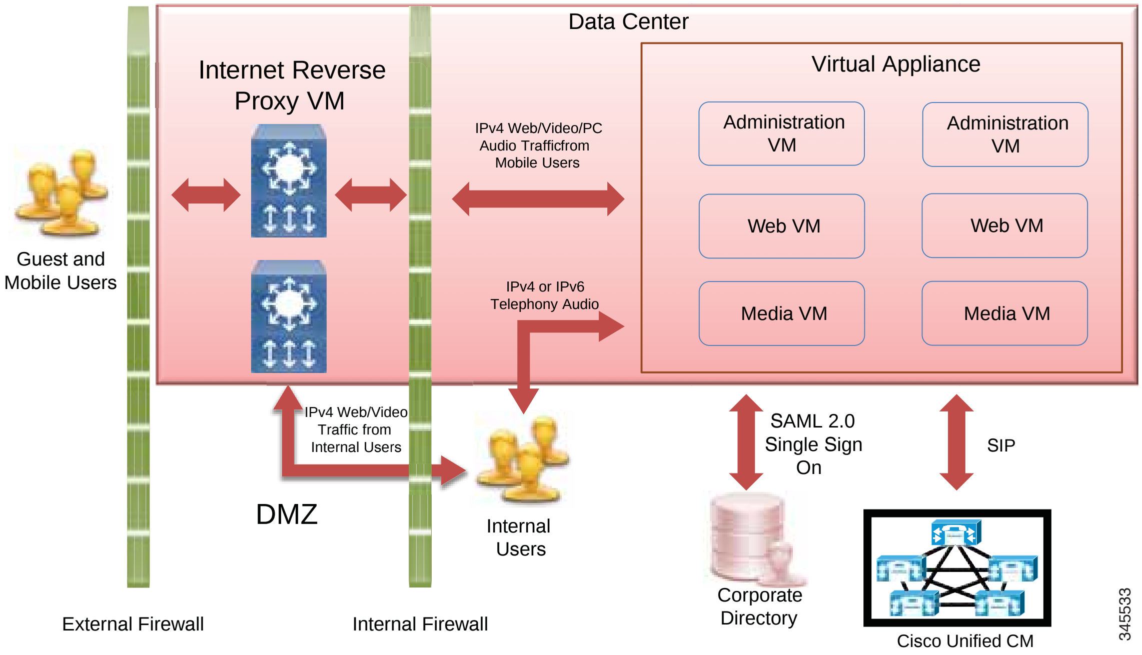
These capabilities are discussed in more detail in the following sections.

Architecture

Cisco WebEx Meetings Server is a fully virtualized, software-based solution that runs on Cisco Unified Computing System (UCS). It uses the virtual appliance technology for rapid deployment of services. Virtual appliance simplifies the task of managing the system. For example, using the hypervisor technology, system components can easily be moved around for maintenance, or system components can easily be rolled back to a working version if problem arises. The virtual appliance is distributed in the form of an industry standard format, Open Virtual Appliance (OVA). All the software components required to install WebEx Meetings Server are packaged inside the OVA. Traditionally, using an executable installer to install individual software components would take hours to deploy the software. However, using OVA can significantly reduce the amount of time required to deploy the software because all software components are pre-packaged inside the file. Thus, virtual appliance technology can help tremendously to reduce the deployment time for Cisco WebEx Meetings Server.

Figure 22-4 shows the high-level architecture for Cisco WebEx Meetings Server using the non-split horizon network topology. (For details on the non-split horizon network topologies, refer to the *Cisco WebEx Meetings Server Planning Guide*, available at http://www.cisco.com/en/US/products/ps12732/products_installation_and_configuration_guides_list.html.) Inside the virtual appliance, there could be one or more virtual machines (VMs) running. These are the administration, web, and media virtual machines. The administration and web virtual machines serve as the back-end processing for the administration and WebEx sites. These sites handle tasks that happen before and after the meeting, such as configuration, scheduling/joining meetings, and recording playback. The media virtual machine provides resource allocation, teleconference call control, and media processing (voice, video, and data) during the meeting. The number of virtual machines running inside the virtual appliance depends on the capacity desired and on whether high availability is needed. This provides various options for deployment size.

Figure 22-4 Cisco WebEx Meetings Server High-Level Architecture

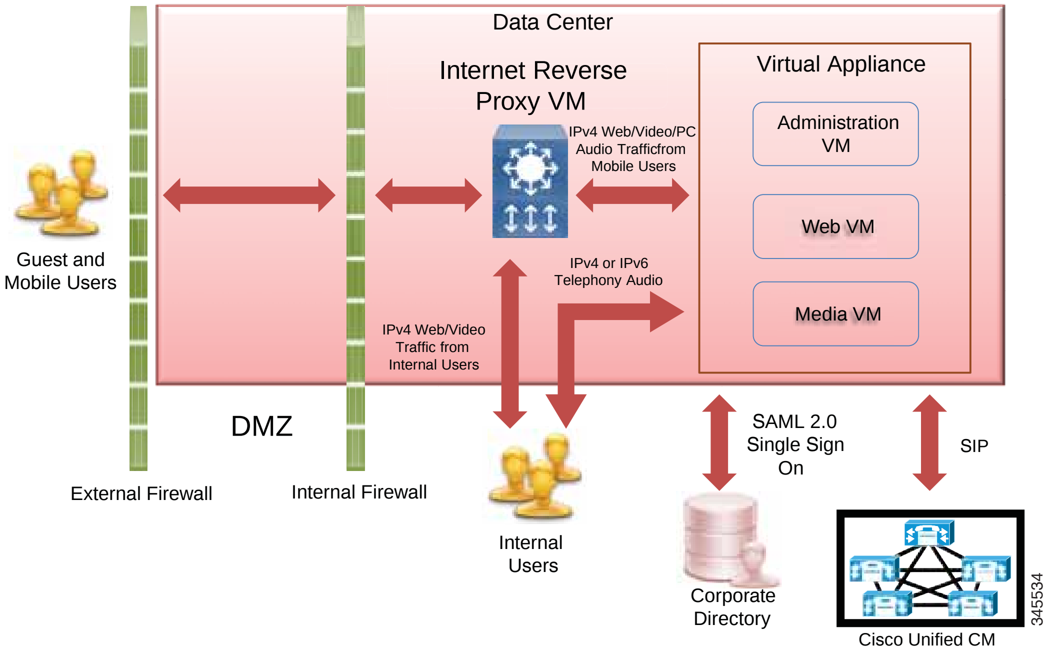


Cisco WebEx Meetings Server offers the option of deploying the Internet Reverse Proxy (or edge servers) in the DMZ to facilitate external access. This option provides two advantages. First, all external participants can securely access the WebEx conferences from the internet without going through a VPN. Second, mobile users can join the meetings from a mobile device anywhere as long as there is internet connectivity. Note that the Internet Reverse Proxy is mandatory if mobile client access is enabled.

Internet Reverse Proxy is used to terminate all inbound traffic from the internet inside the DMZ. The content is then forwarded to the internal virtual machines through an encrypted Secure Socket Layer (SSL) or Transport Layer Security (TLS) tunnel. This encrypted tunnel is established by the internal virtual machines connecting outbound to the Internet Reverse Proxy. Therefore, there is no need to open TCP ports inbound from the DMZ to the internal network on the internal firewall. However, some outbound ports from the internal network need to be opened on the internal firewall to allow communication with the Internet Reverse Proxy in the DMZ.

All end-user sessions are 100% encrypted using industry standard Secure Socket Layer (SSL) and Transport Layer Security (TLS). All traffic between the virtual machines is sent over the secure channel. Federal Information Processing Standard (FIPS) encryption can also be turned on by a single policy setting, providing US Department of Defense (DoD) level security. Alternatively, the Internet Reverse Proxy can be deployed behind the internal firewall as shown in Figure 22-5.

Figure 22-5 Internet Reverse Proxy Behind the Internal Firewall



For security concerns, an organization would typically take several months to get approval in deploying a component inside the DMZ. Using this methodology, it could eliminate any DMZ components and bypass the approval process to get the WebEx Meetings Server deployment done quickly. All internet traffic (HTTP on port 80 and SSL on port 443) to the external firewall should be forwarded to the internal firewall. This will minimize the number of ports that need to be opened in the external and internal firewalls. However, placing the Internet Reverse Proxy inside the internal network implies that inbound internet traffic will terminate in the internal network. Although direct internet access to the internal network could be controlled by the firewalls, not all organizations allow terminating internet traffic directly on their internal network. Ensure that this deployment does not violate your organization's IT policy before choosing this option.

In a large enterprise deployment, an organization would require the Single Sign On (SSO) capability to allow end users to sign in using their corporate credentials. Cisco WebEx Meetings Server can connect to the corporate LDAP directory using the industry standard SAML 2.0 for SSO.

Note Cisco WebEx Meetings Server supports Meeting Center only and does not support WebEx OneTouch.

Cisco Unified CM Integration

Cisco WebEx Meetings Server support both Cisco Unified CM and Session Management Edition (SME). Cisco Unified CM is a central piece of the WebEx Meetings Server architecture that allows the following:

- Attendees joining the teleconference by means of Cisco IP Phone or PSTN
- Integration of legacy or third-party PBXs with Cisco WebEx Meetings Server

Cisco Unified CM integrates with WebEx Meetings Server by means of SIP trunks to provide inbound and callback call control. Customer can choose to turn on security and run Transport Layer Security (TLS) and Secured Real-time Transport Protocol (SRTP) over the SIP trunk connection. A SIP trunk is configured in Unified CM with a destination address of the Load Balancer in WebEx Meetings Server, and then a route pattern (match the call-in access number configured in WebEx Meetings Server) must be used to route calls via the SIP trunk. A second SIP trunk is configured in Unified CM with a destination address of the Application Server in WebEx Meetings Server, and then a SIP route pattern must be used to route calls via the SIP trunk. When an attendee dials the access number to join the meeting, the first SIP trunk is used to send the call. After the call is connected and the caller enters the meeting ID, the Load Balancer issues a SIP REFER to Unified CM to send the caller to the Application Server that hosts the meeting via the second SIP trunk.

The system administrator can configure a SIP trunk in WebEx Meetings Server that points to a Unified CM to perform callback. Attendees can provide a callback number and have the system out-dial the number to the attendees to join the bridge. In the case of attendees requesting callback, the WebEx Meetings Server sends the SIP request to Unified CM along with the callback number via the configured SIP trunk. It is imperative for Unified CM to be able to resolve all dial strings received from a callback request to join the meetings. Callbacks may also be disabled system-wide by means of site administration settings. Unified CM is in control of all toll restrictions to various countries or other numbers that most enterprises will block, because WebEx Meetings Server does not have any toll restriction blocking itself.

WebEx Meetings Server supports the bidirectional SIP OPTIONS ping mechanism. The ping response from the remote end indicates that the remote end is active and whether it is ready to accept calls. Based on the response, WebEx Meetings Server or Unified CM can determine whether to send calls on the current SIP trunk or look for an alternate SIP trunk (if configured) to send calls. Note that SIP OPTIONS ping is supported in Cisco Unified CM 8.5 and later releases. Due to this reason, Cisco recommends using a compatible Cisco Unified CM version that supports SIP OPTIONS ping for Cisco WebEx Meetings Server deployment. For the list of compatible Unified CM versions, refer to the compatibility matrix in the *Cisco WebEx Meetings Server System Requirements*, available at

> http://www.cisco.com/en/US/products/ps12732/products_installation_and_configuration_guides_list.html

Note Cisco WebEx Meetings Server supports SIP trunk connection with Cisco Unified CM only.

Legacy PBX Integration

Some organizations that have a legacy PBX and are not ready to fully migrate to a Cisco Unified Communications solution, might want to use Cisco WebEx Meetings Server with their system for conferencing. Cisco Unified CM can be used to bridge the legacy PBX and Cisco WebEx Meetings Server together. Cisco WebEx Meetings Server can see only Unified CM and does not even know the PBX is behind Unified CM. As long as Unified CM can interoperate with the organization's PBX, Cisco WebEx Meetings Server can integrate with the organization's PBX. This integration can provide several benefits:

- Allow users in the legacy system to experience the new technology
- Allow an organization to adopt the new technology gradually, at its own pace
- Protect the customer's investment in existing technology while allowing them to migrate to Cisco technology gradually

For further details on PBX interoperability with Unified CM, refer to the documentation available at

http://www.cisco.com/en/US/solutions/ns340/ns414/ns728/networking_solutions_products_genericcontent0900aecd805b561d.html

IPv6 Support

Cisco WebEx Meetings Server supports IPv4 only or dual stack (IPv4 and IPv6) addressing for telephony audio, while telephony signaling remains at IPv4. Audio streams can be IPv4, IPv6, or a mix of IPv4 and IPv6 in the same meeting. Cisco WebEx Meetings Server supports Alternate Network Address Types (ANAT) to enable both IPv4 and IPv6 media addressing in the Session Description Protocol (SDP) during the SIP Offer and Answer exchange on the SIP trunk with Unified CM to establish a media connection using the preferred addressing scheme.

Both IPv4 and IPv6 devices can be used for teleconferencing. With IPv6 devices, Cisco WebEx Meetings Server leverages Unified CM's capacity to translate the IPv6 signaling to IPv4 and transport it over a SIP trunk to the Cisco WebEx Meetings Server. With the telephony media addressing, Cisco WebEx Meetings Server can convert between IPv4 and IPv6. Therefore, Cisco WebEx Meetings Server can support IPv6 without any expensive MTP resources.

With ANAT, Cisco WebEx Meetings Server can support IPv6 telephony audio without the support of IPv6 telephony signaling. However, ANAT must be supported on both ends of the Unified CM SIP trunk. Be sure to enable ANAT on the Unified CM SIP trunk, otherwise there will be a failure to establish the call when attendees request callback or attempt to dial in.

If the WebEx Meetings Server has IPv6 enabled, ANAT headers will be included in the media offer. WebEx Meetings Server will always answer with ANAT headers if the media offer includes ANAT headers. The following paragraphs describe the media address version selection process between the IPv6-enabled WebEx Meetings Server and the dual-stack Unified CM using the ANAT header.

When WebEx Meetings Server sends a call to Unified CM, the SDP offer contains both IPv4 and IPv6 media addresses. If the called device is IPv6, Unified CM chooses IPv6 for the media connection and answers with the IPv6 media address in the SDP; if the called device is dual-stack, Unified CM uses the **IP Addressing Mode Preference for Media** parameter to determine the address version in the answer SDP. If the parameter is set to IPv6, then IPv6 will be used for the media connection.

When Unified CM sends a call to the WebEx Meetings Server through the SIP trunk, WebEx Meetings Server receives the SDP offer with an ANAT header. If the SDP offer contains both IPv6 and IPv4 media addresses, WebEx Meetings Server answers with the higher precedence address version specified in the ANAT header, which would be IPv6 in this case. If the SDP contains only an IPv6 address, WebEx Meeting Server answers with an IPv6 media address.

For information on deploying IPv6 in a Cisco Unified Communications system, refer to the latest version of *Deploying IPv6 in Unified Communications Networks with Cisco Unified Communication Manager*, available at

http://www.cisco.com/go/ucsrnd

High Availability

Cisco WebEx Meetings Server uses the N+1 redundancy scheme to ensure system availability in the event of component failures. At the system level, virtual machines and components inside run in active/active mode. If one component goes down, the system restarts the component. Status information is exchanged between system components. Using this status information, the system is able to distribute the requests evenly among the active components. Depending on the deployment size, the number of virtual machines in the backup or redundant system might or might not be the same as in the primary system.

In the high availability system, when the virtual machine hosting the meeting goes down, affected meeting clients will automatically reconnect to the available service within a short period of time. However, depending on the nature of the failure and which component has failure, not all clients and meetings would be affected. For descriptions of system behavior during a component failure, refer to the latest version of the *Cisco WebEx Meetings Server Release Notes*, available at

http://www.cisco.com/en/US/products/ps12732/prod_release_notes_list.html

Virtual IP Address

Inside the high availability system, there is a second network interface in the active administration and Internet Reverse Proxy virtual machine that is configured with the virtual IP address. The administration and WebEx site URLs use this virtual IP address to access the administration and WebEx sites. In the event of failover, the virtual IP address is moved over to the new active virtual machine. Thus, it provides access redundancy to the administration and WebEx site.

Disaster Recovery for Dual Data Center Design

For disaster recovery deployments where the backup WebEx Meetings Server system needs to be in a different geographic location, it is possible to deploy an identically configured recovery system in the second data center. The recovery system is pre-installed and should be shut down or put into maintenance mode while the WebEx Meetings Server system is operational in the primary data center. If a disaster occurs and the primary data center is down, the recovery system should be brought up and restored using the most current system backup from the WebEx Meetings Server in the primary data center.

Consider the following information when using the disaster recovery option:

- Primary and recovery systems are independent of each other and do not connect together in any way.
- The recovery system should have access to the system backup from the primary system to perform restoration.
- Set up a Unified CM subscriber local to the recovery system to handle teleconferencing.

For detail information on disaster recovery requirements and procedures, refer to the *Cisco WebEx Meetings Server Administration Guide*, available at

http://www.cisco.com/en/US/products/ps12732/prod_installation_guides_list.html

Capacity Planning

The capacity of WebEx Meetings Server depends on the platform of choice and the number of conferencing nodes running in the deployment. For capacity planning details, see the section on Collaborative Conferencing, page 29-47.

Storage Planning

If recording meetings is a requirement, sufficient disk space should be allocated on the Network Attached Storage (NAS) device to store the recordings. For disk space allocation detail, refer to the *Meeting Recordings* section in the *Cisco WebEx Meetings Server Planning Guide*, available at

http://www.cisco.com/en/US/products/ps12732/products_installation_and_configuration_guides_list.html

Network Traffic Planning

Network traffic planning for WebEx Meetings Server collaboration consists of the following elements:

- Call control bandwidth

  Call control bandwidth is extremely small but critical. Co-locating the WebEx Meetings Server with Unified CM helps protect against issues with call control. Remote locations need proper QoS provisioning to ensure reliable operation. Call control bandwidth is used for establishment of calls between WebEx Meetings Server and Unified CM, and the amount of bandwidth required for each call depends on how the attendees join the meeting. For an attendee dialing into the meeting, the call consumes approximately the same amount of bandwidth as making two SIP calls. For an attendee requesting callback, the call consumes approximately the same amount of bandwidth as making one SIP call. For details about call control bandwidth estimation for SIP calls and QoS provisioning, see the chapter on Network Infrastructure, page 3-1.

- Real-Time Transport Protocol (RTP) traffic bandwidth

  RTP traffic consists of voice and video traffic. Voice bandwidth calculations depend on the audio codec used by each device. (See the chapter on Network Infrastructure, page 3-1, for bandwidth consumption by codec type.) Video bandwidth can be calculated the same way as WebEx SaaS. (See Network Traffic Planning, page 22-10.)

- Web collaboration bandwidth

  Web collaboration bandwidth for WebEx Meetings Server can be estimated the same way as WebEx SasS. (See Network Traffic Planning, page 22-10.)

Design Consideration

The following additional design considerations apply to WebEx Meetings Server deployments:

- For scenarios where any WebEx Meetings Server components are separated by network firewalls, it is imperative to ensure the correct pinholes are opened for all required traffic.
- Collaborative meeting systems typically result in increased top-of-the-hour call processing load. Capacity planning tools with specific parameters for WebEx Meetings Server are available to Cisco partners and employees to help calculate the capacity of the Cisco Unified Communications System for large configurations. Contact your Cisco partner or Cisco Systems Engineer (SE) for assistance with sizing of your system. For Cisco partners and employees, the Cisco Unified Communications Sizing Tool is available at http://tools.cisco.com/cucst.
- Using Transport Layer Security (TLS) and Secured Real-time Transport Protocol (SRTP) have no effect to the WebEx Meetings Server capacity. However, using TLS and SRTP does have an impact on Cisco Unified CM capacity.
- WebEx Meetings Server has no built-in line echo cancellation. Use an external device such as a Cisco Integrated Service Router (ISR) to provide echo cancellation functionality.
- For more details on the various Cisco collaborative client offerings and how they fit into collaborative conferencing solutions, see the chapter on Cisco Collaboration Clients and Applications, page 24-1.
- Call admission control with WebEx Meetings Server is performed by Unified CM. With locations-based call admission control, Unified CM can control bandwidth to the WebEx Meetings Server system by placing the SIP trunk specific to WebEx Meetings Server in a location with a set amount of audio bandwidth allowed. Alternatively, Unified CM supports the use of Resource Reservation Protocol (RSVP), which can also provide call admission control. For further information regarding call admission control strategies, see the chapter on Call Admission Control, page 11-1.
- Cisco recommends marking both the audio streams and video streams from WebEx Meetings Server as AF41 (DSCP 0x22) to preserve lip-sync. These values are configurable in WebEx Meetings Server Administration.
- Web conferencing traffic is encrypted in SSL and is always marked best-effort (DSCP 0x00).

Reference Document

For network requirements, network topology, deployment size options, and other deployment requirements and options for WebEx Meetings Server, refer to the *Cisco WebEx Meetings Server Planning Guide*, available at

http://www.cisco.com/en/US/products/ps12732/products_installation_and_configuration_guides_list.html

Cisco Unified MeetingPlace

Cisco Unified MeetingPlace combines the benefits and capabilities of Cisco WebEx content sharing with the ability to host the audio and standards-based video portions of the collaboration meetings on-premises. Customers that have invested in Unified Communications solutions are able to leverage and extend their existing deployments to include audio and video conferencing using an all-SIP architecture. Unified MeetingPlace deployments vary depending on several options such as scalability, scheduling interface options, media resource options, and degree of high availability required. These options are discussed in more detail in this section.

There are two different deployment models available with Unified MeetingPlace architecture:

- Multinode Unified MeetingPlace Audio with WebEx Scheduling model for large global enterprises:
  - Provides scalability to 14,400 G.711 audio ports using multiple Conferencing Nodes
  - Provides active/active resiliency for audio conferences
  - Provides virtualization support on the Cisco UCS platform
  - Provides enhanced WebEx integration features
  - Provides optional support for WebEx Node for MCS or ASR 1000 for on-premises mixing of Web conferences for internal network users
  - Provides user-based licensing for Active Users and hardware-based server capacity for ports

  Note Multinode deployment support is available with Cisco Unified MeetingPlace 8.5 and later releases.

- Unified MeetingPlace Scheduling model:
  - Available to installed base of Unified MeetingPlace customers only
  - Available as audio/video only with no Web conferencing (no WebEx) to new or installed-base customers
  - Provides continuous meetings with blast outdial
  - Provides Cisco Unified Communications Manager Video Telephony ad-hoc support
  - Provides scalability to a maximum of 1,200 audio ports with Cisco Unified MeetingPlace Express Media Server (EMS) or 2,000 audio ports with Hardware Media Server (HMS) using G.711
  - Provides active/warm-standby resilience with manual failover

Note This chapter focuses on audio, video, and Web sharing solutions. However, Unified MeetingPlace also supports deployments utilizing audio only or audio and video only.

This section covers system-level design guidance of a Cisco Unified MeetingPlace system in the Cisco Unified Communications environment. This chapter does not cover any hardware requirements or software component configurations of Unified MeetingPlace that are not related to system design. For information on these topics, refer to the Unified MeetingPlace product documentation available at

http://www.cisco.com/en/US/products/sw/ps5664/ps5669/tsd_products_support_series_home.html

Note The implementation of any Cisco Unified MeetingPlace 8.*x* web conferencing solution requires the purchase of a WebEx site. The WebEx services are independent of Cisco Unified MeetingPlace licensing.

Unified MeetingPlace Architecture

This section provides a high-level overview of each Unified MeetingPlace component and its function in the solution.

Unified MeetingPlace Meeting Director Server

The Meeting Director node supports several functions for multinode deployments with a WebEx scheduling front end. This is a required component used to support multinode configurations only. The Meeting Director module includes a WebEx Telephony Service Provider (TSP) connection to the WebEx collaboration cloud for integration using outbound TCP 443 only for a two-way communication path for the audio commands. The Meeting Broker Director is responsible for distributing audio meetings between different conferencing nodes in a equal load sharing methodology. The Events Aggregator monitors conferencing node capacity and events happening in real time. UserSync is used to synchronize all profiles from WebEx Site if it is enabled.

A multinode system has one Primary Meeting Director node and one Secondary Meeting Director node for redundancy, which can be located in any customer data center behind a corporate firewall. If the Primary Meeting Director fails, the Secondary Meeting Director becomes active. Cisco recommends that you configure your Meeting Directors as regional masters and that you locate your Meeting Directors in different data centers to provide greater system resiliency.

A "combined node" provides both Meeting Director and conferencing functionality, and it is supported when there are fewer than four Conferencing Nodes in a system. With more than four Conferencing Nodes, both Meeting Directors must reside on a dedicated hardware server (Cisco MCS or UCS).

Unified MeetingPlace Application Server (Conferencing Node)

The Unified MeetingPlace solution centers around the Unified MeetingPlace Application Server, also referred to as a Conferencing node in a multinode configuration, which provides audio and video mixing functionality through SIP trunking from a Unified CM or Session Management Edition call control system. At least one conferencing node is required in order to host conferences. Additional conferencing nodes provide greater capacity and resiliency.

The Unified MeetingPlace Application server is installed on a Cisco Media Convergence Server (MCS) or Unified Computing System (UCS) platform running the Linux operating system and the IBM Informix Dynamic Server (IDS) database, and it acts as the audio/video conference node component that mixes audio and standards-based video conferences in an enterprise network. The Unified MeetingPlace Application server controls the media servers of the solution, and it communicates with the Unified MeetingPlace Meeting Director component in a multinode configuration. The Unified MeetingPlace Application server supports SIP back-to-back user agent (B2BUA) and sends/receives calls through a SIP trunk connection with Cisco Unified CM or Session Management Edition (SME) for call delivery for inbound and outbound callbacks. The Cisco Unified MeetingPlace Express Media Server is also an optional software component that can be installed co-resident on the Unified MeetingPlace Application server and it is the preferred media mixer for most customer scenarios. Optionally, the Hardware Media Server scales higher per node (maximum of 2,000 G.711 audio port per audio node).

Media Server

The Cisco Unified MeetingPlace Media Servers provide the audio and video conferencing functionality for the solution, and they come in two distinct options:

- Cisco Unified MeetingPlace Express Media Server (EMS)
- Hardware Media Server (HMS)

The Express Media Server is the preferred cost-effective option with Cisco Unified MeetingPlace, and it performs audio mixing and standards-based video switching in software that is co-resident on the Unified MeetingPlace Application Server. The EMS allows for a single-box software-only solution for a Cisco Unified MeetingPlace audio/video-only deployment, or it can be deployed in a multinode configuration. Media cannot be cascaded across EMS instances; therefore, the capacity of a Unified MeetingPlace EMS solution depends on the MCS or UCS platform on which it is installed, or whether you install multiple Unified MeetingPlace Application and Express Media servers for scalability in a multinode deployment. Scalability in a multinode deployment can provide a maximum of 14,400 G.711 ports and requires the use of a WebEx Scheduling model. There is no cascading capability across EMSs. Higher capacities per node are available from the HMS option and with the EMS multinode deployment option.

For ultimate capacity on Express Media Servers, G.711 audio-only provides the highest number of simultaneous ports for audio conferencing. If G.729 or G.722 audio codecs are needed, then capacity is much less. Also, if standards-based video mixing is used, again this lowers capacity depending on the type mixing and maximum bandwidth settings. For instance, a Cisco UCS B-Series Blade Server using G.711 audio-only can support a maximum of 1,200 ports. To enable maximum capacity, Cisco highly recommends providing network layer audio codec transcoding to G.711 in Cisco Integrated Services Routers (ISRs) for calls that transverse a WAN in G.729 or G.722 and terminate in a Unified MeetingPlace conferencing node or single system. For more information, see Capacity Planning, page 22-44.

A Hardware Media Server is a Cisco Unified MeetingPlace 3515 or 3545 outfitted with blades that are specific to the Unified MeetingPlace solution. There are audio blades and optionally standards-based video blades, both of which have on-board DSP resources to provide voice and video conferencing, respectively. The HMS is controlled by the Unified MeetingPlace Application server through SIP API and Unified MeetingPlace Media Control protocols. The HMS supports cascading of audio and video streams, therefore multiple HMS 3545 chassis can be deployed in a single location to achieve the capacity and high availability required. HMSs cannot be distributed throughout a network and must be located in the same data center as the Unified MeetingPlace Application server. HMS standards-based video provides "continuous presence," which is composed video with support for standard format up to 2 MB per video stream. HMS video also fully supports transcoding and transrating, important features in standards-based video to provide advanced video MCU functions. High definition formats are not currently supported, but HD video devices can join standard format meetings.

The Unified MeetingPlace Application server can be configured to use either the EMS or HMS, but the two cannot be used together in the same conferencing node. It is relatively easy to switch from one to the other, however. Use of either is transparent to the user except for differences in supported video formats and features such as active speaker or continuous presence, transrating, transcoding, video recording, video mute, or HD video capabilities. There are some major differences in features and capabilities between an EMS and HMS; therefore it is critical to review these differences before choosing between them. For more information, refer to the latest version of the *Planning Guide for Cisco Unified MeetingPlace*, available at

http://www.cisco.com/en/US/products/sw/ps5664/ps5669/products_implementation_design_guides_list.html

WebEx Node for MCS or ASR (Optional Component)

The design of a Unified MeetingPlace solution is affected by the nature of the meetings to be hosted on the system. For example, is there a requirement for meetings to include only internal participants, or are external attendees also allowed? All web conferencing for the Unified MeetingPlace solution is provided by WebEx; however, the WebEx Node for MCS or WebEx Node for ASR 1000 optionally allows an organization to pull content sharing resources on-premises if required. If all meetings involve external participants or if the customer would prefer to use the WebEx Collaboration cloud only, then the WebEx Node for MCS or ASR 1000 is not required. However, if there is a requirement to have internal meetings where all the audio, video, and content sharing remains on-premises, the WebEx node for MCS should be deployed. The WebEx Node for ASR provides on-premises mixing for both internal web conference attendees and/or WebEx webcam high quality video (HQ Video). The node essentially extends the WebEx cloud's collaboration bridge technology into a customer's organization by using dedicated MCS or ASR 1000 hardware and WebEx software. It does have direct communication with the Unified MeetingPlace Application server; however, it is still operated and managed through the WebEx site administration, thus it requires connectivity to the internet so that the node can initiate outbound TCP port 443 SSL connections to the organization's WebEx site.

The WebEx client finds the WebEx Node for MCS in the same way it does for the WebEx Node for ASR. The WebEx node names are provisioned in the cloud, and after initial connection to the WebEx site, a list of Meeting Zone URLs is passed to the client from the meeting entry page. For internal-only meetings, only WebEx Node for MCS hostnames are passed to the client. This ensures that all users will be connected to WebEx Node for MCSs internally and no meeting information is cascaded to the Collaboration Cloud for that meeting. For external meetings on WebEx Node for ASR or MCS, there are cloud-based URLs and WebEx Node for MCS hostnames for profiled users, and only cloud-based URLs for external users (guests). The client then pings all the Meeting Zones and connects to the URL with the least amount of latency. This means that all WebEx Nodes for MCS load-share, and you cannot specify certain users to use certain servers. Most likely, users will be connected to the closest node, but that might not be the case depending on network situation and congestion. External meeting guest users are always connected to the Collaboration Cloud, and internal users are on the closest WebEx Node for MCS or ASR 1000. The WebEx Nodes for MCS or ASR 1000 and cloud users can see content shared by anyone with Sharing assignment during a meeting.

Note The WebEx Node for MCS and WebEx Node for ASR are different products. WebEx Node for MCS provides only collaboration bridge functionality (no multi-media) and is specific to the Unified MeetingPlace 8.*x* solution. It cannot be used for a WebEx SaaS implementation. For more information on WebEx Node for ASR, which provides on-premises mixing for both web meetings and HQ Video, see Cisco WebEx Software as a Service, page 22-4.

Note Internal meetings hosted on the WebEx Node for MCS support only Meeting Center meetings. Event Center and Training Center meeting traffic can be aggregated on the WebEx Node for MCS, but it can be designated only as an external meeting. Internal meetings do not support WebEx HQ Video nor Network Based Recordings (NBR) since both of these services are provided in the cloud. Only meetings scheduled as "external" provide both WebEx HQ Video and NBR recordings. NBR with WebEx Node for MCS is not supported for WebEx scheduling deployments but it is supported for Unified MeetingPlace scheduling deployments.

Also, remember that the WebEx Node for MCS does not support HQ Video (webcam only) and WebEx VoIP switching. So unless WebEx webcam video is disabled for the site, it will propagate to the cloud and be switched there. Meetings scheduled as "internal" do not have a data connection to the WebEx collaboration cloud to get the webcam video, so users must schedule meetings as "external" to use both the bandwidth aggregation of the web conference and the webcam video mixed in the cloud. Customers should choose between using either Unified MeetingPlace standards-based video or WebEx HQ Video in the cloud. Additionally, WebEx Node for ASR can be deployed to provide bandwidth aggregation of both webex web conferencing meetings and WebEx HQ Video with webcam mixing on the ASR.

Customers can also choose to disable HQ Video for the WebEx site and instead use no video or Unified MeetingPlace standards-based video (H.323, SIP, and SCCP devices only) on native webcams.

WebEx Site

All Unified MeetingPlace 8.*x* web conferencing solutions require a WebEx site. A WebEx site for a given organization will have the format *companyXYZ*.**WebEx.com**. Enterprise customers may use Meeting Center only or a combination of all the WebEx centers, which is called Enterprise Edition and which supports Meeting Center (MC), Event Center (EC), Training Center (TC), and Support Center (SC). WebEx packages for Active Host, Named Host, Ports, or minutes are all supported with Cisco Unified MeetingPlace 8.5 and later releases, with or without WebEx Node.

Event Center and Training Center offer additional integration features. Event Center Audio Broadcast allows for efficient use of Unified MeetingPlace Audio. Only presenters in an event meeting are connected to the Unified MeetingPlace Audio system, and all participants (up to 3,000) join by means of a browser URL and can listen to the audio broadcast in streaming mode (not multicast). Unified MeetingPlace audio can support a maximum of 500 audio ports in a single large meeting with auto-mute if desired, but Cisco highly recommends using the Event Center Audio broadcast feature for large meetings for one-to-many functions. Training Center offers the use of audio/web breakout rooms and mute participants upon entry.

A single WebEx Site is tied to only one Unified MeetingPlace system. A Unified MeetingPlace system in the multinode deployment model requires using the WebEx Scheduling model only. Multiple WebEx Sites cannot be supported on one Unified MeetingPlace system, and multiple Unified MeetingPlace systems cannot be supported on one WebEx Site.

Cisco Unified MeetingPlace 8.5 and later releases with WebEx WBS27 FR 26 and above allow Unified MeetingPlace to be integrated without any need for provisioning. Existing WebEx customers that have this release can easily add Unified MeetingPlace Audio to their existing site without any provisioning requests or changes. In addition, this WebEx release also supports Dual Audio vendor, which will allow for either WebEx Audio and Unified MeetingPlace Audio on the same site or Unified MeetingPlace Audio and TSP Audio on the same site. There is an administrative portal to the WebEx site that is used to configure key parameters that tie the site to the Unified MeetingPlace deployment. For more information regarding the WebEx site configuration, refer to the *Administration Documentation for Cisco Unified MeetingPlace*, available at

http://www.cisco.com/en/US/products/sw/ps5664/ps5669/prod_installation_guides_list.html

Note For Unified MeetingPlace audio/video-only deployments, a WebEx site is not required.

WebEx Site Dual Audio Support

A WebEx site using Release 27 FR26 or above supports a new feature called Dual Audio Vendor support. This feature allows for the following configurations and integrations:

- WebEx Audio/VoIP + Unified MeetingPlace audio
- TSP audio + Unified MeetingPlace audio

The Dual Audio Vendor feature enables existing WebEx sites with TSP Audio or WebEx Audio to configure Unified MeetingPlace Audio as well, and do a phased migration from one type to the other, which allows future meetings already scheduled with the first audio provider to still be used while all new future meetings start using Unified MeetingPlace Audio. This also allows different regions of the world to use different audio systems based on profile default settings. For example, Singapore can use WebEx Audio while all North America users are set to use Unified MeetingPlace Audio only.

In addition, profiles can be configured to offer both audio providers, and users must know how to schedule using each provider per meeting. Specific WebEx session types can also be configured to use one type of audio provider based on the meeting type scheduled.

Dual Audio Vendor support does not provide automatic overflow from one to the other or combining of both audio systems together.

Unified MeetingPlace Audio currently does not support "mixed" audio conferencing with the WebEx VoIP feature. So if customers want to use WebEx Audio with VoIP, this dual vendor audio support would have to be configured, and users would have to know to choose the WebEx Audio/VoIP option to use this function.

User Based Licensing

Starting with Cisco Unified MeetingPlace 8.5, a user-based licensing model is used. In previous versions of Unified MeetingPlace, ports-based licensing was used. A user-based licensing model allows customers to purchase systems based on the "active" users on the Unified MeetingPlace system. Active users are defined as a profiled account that schedules meetings or hosts meetings on Unified MeetingPlace. System reports are available for monitoring active usage to see if the system has exceeded the purchased user count. Also, a minor SNMP alarm is sent if the active user count is above licensed user count. In no way will Unified MeetingPlace block a conference call or profiled host from having a meeting. Customers may provision as many users as they need without any issues by using the various provisioning options available through WebEx or native to Unified MeetingPlace. The Unified MeetingPlace database will support a maximum of 400,000 profiles.

Note A user license (audio, web, or video) is not granted to any particular user but, rather, is a system-wide resource shared by all users in the Unified MeetingPlace system.

System capacity for the total number of audio callers connected simultaneously is dependant entirely on the hardware server model and number deployed. Peak usage and future growth both must be factored in when designing a Unified MeetingPlace on-premises solution. If you deploy two Cisco UCS B-Series Blade Servers or C210 Series Rack-Mount Servers with Unified MeetingPlace Application and EMS software, you will have 1,200 G.711 ports per server or 2,400 total ports or 1,200 redundant ports that all profiled users and guests can utilize. Conferencing nodes have active/active load sharing of all meetings. If one server is down, the calls on that server are dropped and users can immediately dial back in or use Callback from the WebEx meeting room user interface, and that meeting will be reestablished automatically on the other server (or the least busy server in the region). Unified MeetingPlace supports up to 14 conferencing nodes with a total of 14,400 G.711 ports. If G.729, G.722, and/or standards-based video is used, it will reduce these capacity numbers.

Unified MeetingPlace supports both scheduled and reservationless meetings. Reservationless meetings are audio only (or audio/video only if video is enabled).

Scheduling Interface

The Cisco Unified MeetingPlace solution offers two scheduling interface options:

- WebEx Scheduling Model using Productivity Tools, One Click, and WebEx scheduling interfaces
- Unified MeetingPlace Scheduling Model using Outlook, Lotus Notes, Conference Manager, or Web scheduling interfaces

In many cases, user familiarity with a particular interface will influence the decision of which option to choose. If users are currently using a WebEx SaaS deployment and simply want to pull audio/video resources on-premises, or if this is a new Unified MeetingPlace installation, Cisco recommends the WebEx scheduling deployment model. The WebEx Scheduling model is required for multinode deployments of Unified MeetingPlace 8.5 or later releases. However, if Unified MeetingPlace is currently deployed, it might be beneficial to maintain the same scheduling interface. While there are certainly differences, both have a web-based user scheduling portal and both have their own integrations with common calendaring systems (Outlook or Lotus Notes). Also, WebEx scheduling supports Enterprise Edition meetings (Meeting Center, Event Center, and Training Center sessions), while Unified MeetingPlace scheduling supports Meeting Center sessions only. The Unified MeetingPlace scheduling model is not available for new customers deploying Unified MeetingPlace 8.5.

WebEx Scheduling Deployment

WebEx supports two deployment models:

- Single-Site WebEx Scheduling Deployments, page 22-27
- Multisite WebEx Scheduling Deployment, page 22-29

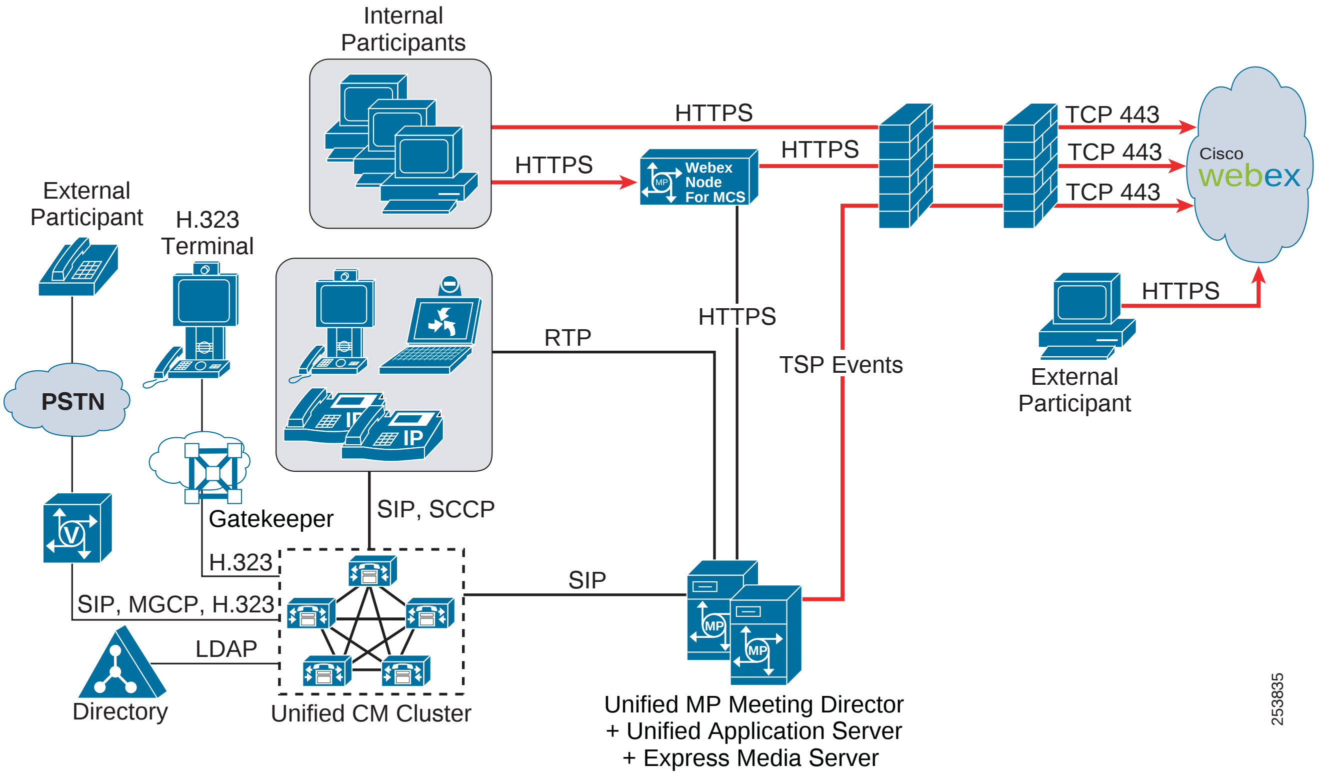
The WebEx Scheduling deployment model supports Meeting Center only or WebEx Enterprise Edition (EE), which includes Meeting Center, Event Center, and Training Center session types, all of which can integrate to Unified MeetingPlace Audio. Only Meeting Center meetings are mixed both on WebEx Node for MCS and in the cloud (for guest users). Event Center and Training Center are always considered external meeting types, and internal users join the WebEx node for MCS or ASR or cloud for those session types.

WebEx Scheduling utilizes all the current WebEx Productivity Tools (see Cisco WebEx Software as a Service, page 22-4), and all audio and WebEx recordings for external meetings are stored in the WebEx Collaboration cloud under the Network Based Recording site per host account.

Single-Site WebEx Scheduling Deployments

With WebEx scheduling, there are no Unified MeetingPlace Web servers required, and the click-to-attend URL in a meeting invitation takes users directly to the WebEx site. Figure 22-6 illustrates a high-level view of a sample Unified MeetingPlace solution with WebEx scheduling, dual Express Media Servers with active/active redundancy, and a WebEx Node for MCS. The WebEx Node for MCS is optional (required for internal-only meeting, or ASR can also provide both Web and HQ Video bandwidth aggregation), and alternatively an HMS could be used in place of an EMS.

Figure 22-6 Unified MeetingPlace Single-Site Solution with WebEx Scheduling, EMS, and WebEx Node for MCS



Note If WebEx Node for MCS is deployed, then only external meetings can support Network Based Recordings and HQ Video webcams with WebEx scheduling.

WebEx Node for MCS or WebEx Node for ASR 1000 are optional, based on whether customer requirements detail bandwidth aggregation and/or use of "internal" meetings only is available. Because the audio conferencing is occurring on-premises while the web conferencing is occurring both in the cloud and on the WebEx Node, all meeting-related service requests are exchanged and processed via telephony service provider (TSP) application programming interface (API) communications with Unified MeetingPlace or the WebEx Node API to the cloud. This effectively ties the systems together and allows for in-meeting controls such as the ability to mute attendees or to see active speakers. This TSP link is established by the Meeting Director outbound to the cloud via a TLS encrypted dedicated socket connection on TCP port 443 to the customer WebEx site.

Network Requirements

This hybrid architecture does not require any "inbound" ports to be opened through the firewall. The Meeting Director TSP supports only SOCKS proxy servers (not HTTP or HTTPS proxy). The WebEx Node for MCS or ASR does not support any type of web proxy systems and must be allow TCP 443 outbound to the cloud if deployed. Users joining WebEx meetings also use TCP 443 outbound only through firewalls to the WebEx Collaboration Cloud. WebEx publishes the IP ranges required if firewall settings to limit internet access are necessary.

Cisco recommends a maximum latency between all components of 300 ms round-trip time (RTT), wherever components may be deployed in the enterprise network. Standard VoIP network best practices also apply to deploying Unified MeetingPlace on-premises conferencing resources. SIP trunking latency between Unified MeetingPlace conferencing nodes from/to Unified CM must adhere to this same standard for optimal conferencing performance.

For all network requirements, refer to the latest version of the *System Requirements for Cisco Unified MeetingPlace*, available at

> http://www.cisco.com/en/US/products/sw/ps5664/ps5669/products_device_support_tables_list.html

Multisite WebEx Scheduling Deployment

Multisite deployments consist of sites and regions. Conferencing nodes, Meeting Director nodes, and optionally WebEx nodes are installed in data centers based on customer requirements for both capacity resiliency.

Sites are logical groups of nodes that have similar functions and capabilities. For example, a site might contain nodes with high-definition video capabilities. Sites are identified by a unique name in the system and can belong to only one region. A site contains one to all of the nodes in a region. You can configure a preferred site to host all meetings for specific user profiles.

Regions are groups of one or more sites. Regions are identified by a unique name in your system. You can have up to four regions in your system, and regions are also used to assign time zones.

A multinode Unified MeetingPlace Audio system has the following capacities:

- 14,400 G.711 audio ports
- 16 Cisco Unified MeetingPlace application server nodes consisting of two Meeting Director nodes and 14 conferencing nodes (12 with 1,200 G.711 ports = 14,400 ports, and 2 extra conferencing nodes for resiliency is supported)
- 1,200 ports per conferencing node (G.711) until the 14,400 limit is reached
- Maximum of four nodes per site
- Maximum of two sites per region (two sites with up to two nodes each, or one site with up to four nodes)
- Maximum of four regions

Note Capacities will be lower depending on G.729 or G.722 codec use, video use type, and bandwidth allowed.

WebEx Web Conferencing (required for scheduling and web conferencing) has the following capacities:

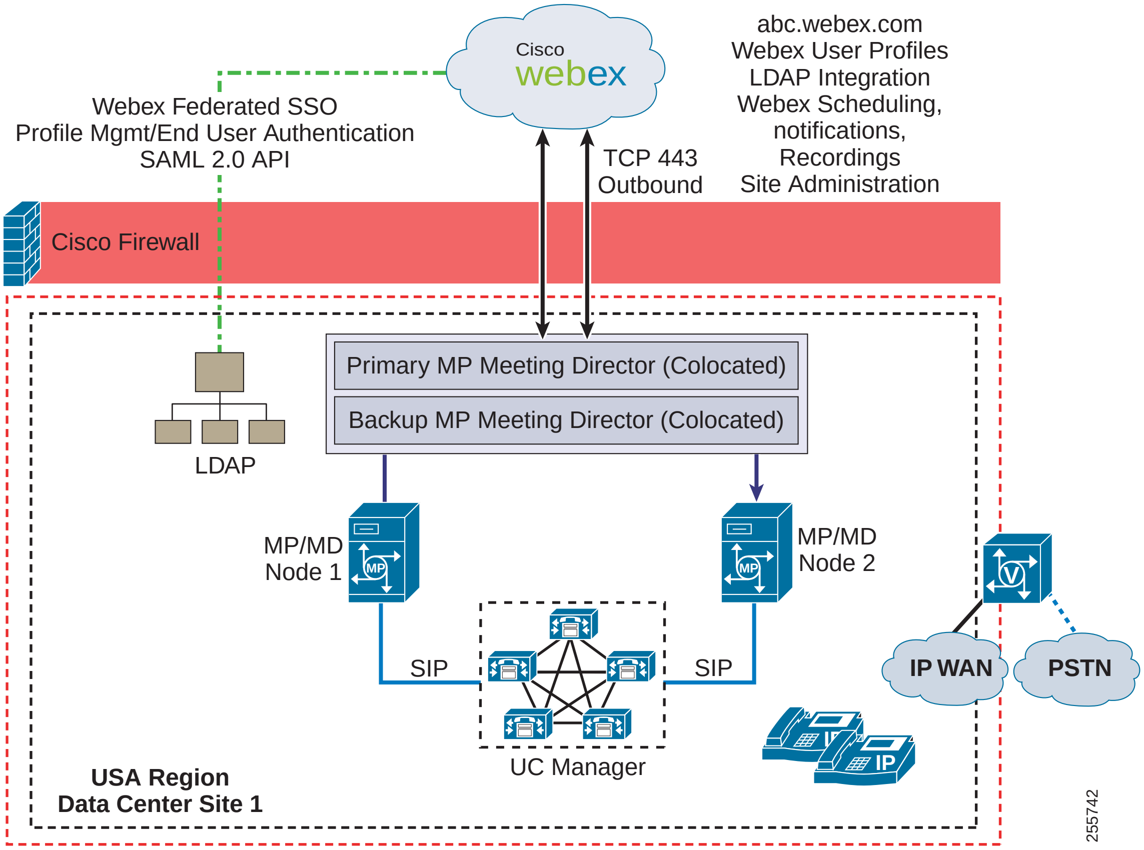
- 14,400 Web sessions (cloud and/or nodes)
- 2,000 internal Web sessions (using Cisco WebEx Node for MCS), consisting of up to 4 Cisco WebEx nodes with up to 500 sessions each
- Cisco WebEx Node for ASR supports:
  - Web conferencing per Shared Port Adapter (SPA), with up to 500 sessions each
  - HQ Video and VoIP per SPA (capacity based on usage)

Meetings are distributed evenly by configuring inbound SIP trunks to all Conferencing Nodes in a circular method in Unified CM or Session Management Edition. Callbacks initiated from within a WebEx meeting room are distributed by the Meeting Director who is monitoring all conferencing node traffic. The Meeting Director will start a new meeting on the least busy node in the region and based on the timezone of the host who scheduled that meeting. For inbound calls, the first person who joins the meeting will dictate which conferencing node they land on based on the SIP circular hunt mode. If that meeting ID is started on a different node within the same region or in a different region, a SIP Refer command will be initiated automatically to redirect that caller to the conferencing node where the host is assigned. All callers into the same meeting ID will be routed to one node in the system based on either timezone or the node on which the meeting was started by the first attendee. Thus, all users in the system will always dial their local Unified MeetingPlace dial-in numbers (or use callback) to join any meeting anywhere in the world. The SIP Refer will automatically redirect them to the proper node for that particular meeting, depending on the timezone of the host who scheduled that meeting. If a reservationless meeting ID is used, callbacks are distributed based again on the timezone where that host resides, but load sharing among multiple node is used for maximum capacity and resiliency.

Centralized Deployment Model with Multinode WebEx Scheduling

The example in Figure 22-7 consists of one region with active/active resiliency in a single site. This system requires two Cisco MCS or UCS servers to provide for two Meeting Director and/or EMS servers deployed in one sites and one region, which is a centralized deployment model. Scalability is 1,200 G.711 ports with active/active redundancy, and both servers equally share the meeting load from all timezones. Unified CM SIP trunk sizing needs to take into account only simultaneous peak SIP traffic, not 2,400 ports of SIP traffic. The Meeting Director is co-located with two different conferencing nodes. The 1,200 ports generally can support a ratio of 20 users to 1 port with typical conferencing usage patterns, so this configuration should be able to support a total of 24,000 users.

Figure 22-7 Unified MeetingPlace Multinode Deployment with WebEx Scheduling for One Region

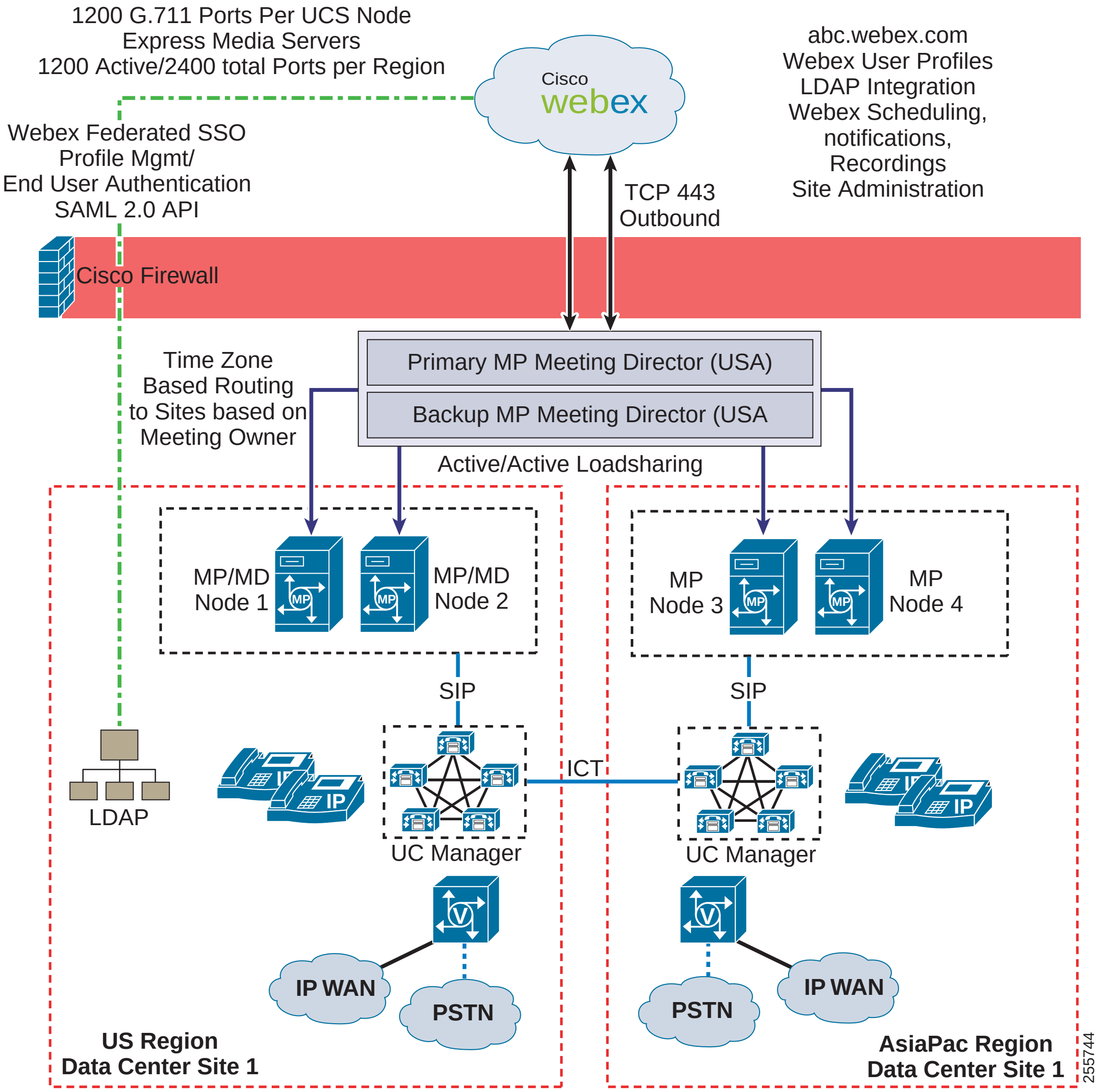


Two-Region Multinode Unified MeetingPlace Deployment Model with Webex Scheduling

The example in Figure 22-8 consists of two regions in a globally distributed design with active/active resilience in each region. Also, data center sites are configured based on customer data center design. All conferencing nodes in a region are load-balanced, and nodes in different sites or regions can fail-over to other regions by means of administration settings.

This system requires four Cisco MCS or UCS servers to provide for two Meeting Director and/or EMS servers and two Conferencing Nodes in two sites and two regions. Scalability is 1,200 G.711 ports per region with active/active redundancy. Unified CM SIP trunk sizing needs to take into account only simultaneous peak SIP traffic, not 2,400 ports of SIP traffic. The Meeting Director is co-located with two different conferencing nodes and can be located in either data center depending on customer requirements.

Figure 22-8 Unified MeetingPlace Multinode Deployment with WebEx Scheduling for Two Regions

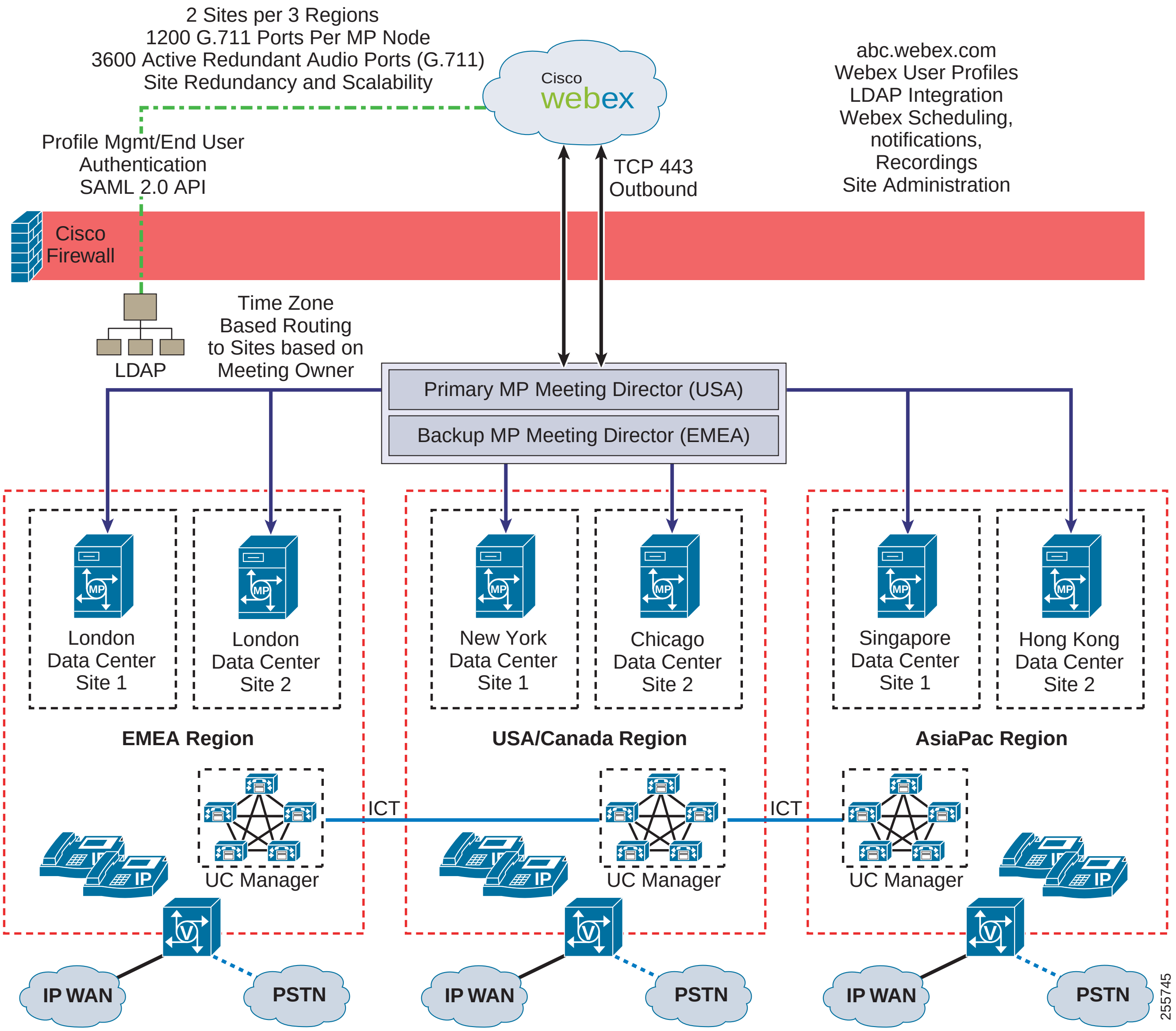


Unified MeetingPlace Multisite Solution with WebEx Scheduling and Three Regions

The example in Figure 22-9 consists of three regions in a globally distributed design with active/active resilience in each region. Also, separate data center sites are configured for site redundancy. All conferencing nodes in a region are load-balanced, and nodes in different sites or regions can fail-over to other regions by means of administration settings.

This system requires eight servers to provide for two Meeting Directors and six Conferencing Nodes. Scalability is 1,200 G.711 ports per region with active/active redundancy per region.

Figure 22-9 Unified MeetingPlace Multisite Solution with WebEx Scheduling for Three Regions



Video

There are two difference types of video available to customers:

- Unified MeetingPlace standards-based third-party room/desktop or Unified Communications Video (H.323, SIP, or SCCP)
- WebEx HQ Video for Meeting Center and Training Center using webcams only

Customers must choose between these two options because there is no interoperability available today between them. Do not enable both because doing so will cause confusion for end users.

With respect to standards-based Unified MeetingPlace video, when video is mixed by the Unified MeetingPlace components on-premises, the video is displayed on the standard room and desktop endpoints themselves. It is not seen in the WebEx video pod inside the web meeting, and Cisco recommends disabling the webcam HQ Video feature on the WebEx site, otherwise there could be a mix of video conferencing with endpoints and webcam video shown in the WebEx application with no tie between them. User-based licensing supports both audio and video usage on any Unified MeetingPlace system. Enabling video on Conferencing Nodes will affect capacity based on the video type and bandwidth used.

For information about standards-based video devices supported with Unified MeetingPlace, refer to the latest version of the *Compatibility Matrix for Cisco Unified MeetingPlace*, available at

http://www.cisco.com/en/US/products/sw/ps5664/ps5669/products_device_support_tables_list.html

Alternatively, if no Unified MeetingPlace video conferencing is deployed, users could take advantage of the WebEx HQ/HD Video capabilities using pure webcams-only mixed in the cloud, or if WebEx Node for ASR with video Shared Port Adapter (SPA) is deployed, then bandwidth aggregation can occur on-premises. WebEx HQ/HD Video cannot be used if WebEx Node for MCS is deployed and users have scheduled meeting as "internal," where there is no data sharing connection to the cloud. If meetings are scheduled as "external," then users can see the webcam video and still be connected to the WebEx Node for MCS for web meeting bandwidth aggregation.

Unified Communications Client Services Framework (CSF) devices and Cisco Unified Video Advantage are both webcam-only or SCCP/SIP video standards-based devices. How the client joins a meeting and which video option is enabled will determine the video experience for the end user. (See Table 22-7.)

Table 22-7 Supported Video Options

Video Type	WebEx HQ Video	MeetingPlace Video
Standards-based support for H.323, SIP, and SCCP	No	Yes
Webcam support	Yes	No
Webex Node for ASR	Yes	No
Internal premises-based	No	Yes
Global Access guest/users	Yes	No

WebEx Owned Profile Management

There are two ways to configure profile management: WebEx Owned Profiles or Unified MeetingPlace Owned Profiles.

WebEx owned profile management allows for profiles to be provisioned in the following ways:

- Account sign-up (automatically approved or with system administrator approval required)
- Manual account creation
- Import periodically from Excel spreadsheet file
- Federated single sign-on (SSO) option (accounts automatically created upon login)
- WebEx XML API (custom account management)

With WebEx Owned Profile enabled, Unified MeetingPlace automatically synchronizes all user profiles from the cloud through the X.509 encrypted link and creates users on Unified MeetingPlace Conferencing nodes. Users can then use the Profile Number and PIN code to access the reservationless audio-only meetings.

Note The Profile Number is eight digits in length and is assigned randomly when the user profile is created. The PIN code can be created by the user upon first logging in to the WebEx site. Optionally, the Profile Number can also be customized by retrieving it from the LDAP directory through the WebEx XML API by using a custom code for mapping LDAP fields to WebEx Profile fields.

Unified MeetingPlace then accesses profiled user information through an XML API User Synch module to automatically configure all users on Unified MeetingPlace Conferencing Nodes. When installing the Meeting Director primary server (the first one in the installation cycle), you choose the **WebEx Owned Profile** setting and the system then operates automatically to synchronize user profiles from the cloud through an X.509 encrypted link.

When WebEx Owned Profiles is enabled, the Unified MeetingPlace system uses a Profile Number and PIN code, which users enter only for reservationless audio-only meetings. When the user profile is newly created, WebEx Site with Unified MeetingPlace will atomically assign a random Profile Number to that user. Upon first logging in to the WebEx Site, that user is prompted to configure a PIN code. If customers want a specific number to be assigned to the users based on an LDAP field, then the WebEx XML API must be used for provisioning a custom code that uses LDAP fields to map to WebEx profile fields. The Profile Number and PIN length requirements are set in the Unified MeetingPlace System Administration parameters. Profile Numbers can be 4 to 8 digits in length, and PIN codes can be 5 to 24 digits in length.

Note WebEx Owned Profile is mandatory in order to enable the optional WebEx Federated Authentication Service (FAS) LDAP capability. For more information on FAS, refer to the WebEx *Federated SSO Authentication Service Technical Overview,* available at http://developer.webex.com/c/document_library/get_file?groupId=10465&folderId=11421&name=DLFE-201.pdf.

WebEx XML API

If you want to control the creation of the MeetingPlace Profile ID with a field that exists in the LDAP profile, then you must write a script to call the WebEx XML APIs for User Service and Create Users functions. One of the parameters for this XML API is the Unified MeetingPlace profile number (mpProfileNumber) assignment. Unified MeetingPlace profile numbers must be between 4 digits and 8 digits in length. Unified MeetingPlace profile numbers are used only with audio-only meetings or reservationless meetings that are audio-only, where the host must log into meetings with this profile number that is the meeting ID and PIN code to start the meeting. All other callers are in a waiting room on Unified MeetingPlace until the host logs in and starts the meeting.   Normal scheduled WebEx and Unified MeetingPlace combined meetings do not require the use of this profile number and PIN code to start them.

For more information on the XML API, refer to the Cisco WebEx Collaboration Cloud documentation available at

http://developer.webex.com/web/meetingservices/xmlapi

Unified MeetingPlace Owned Profile Management

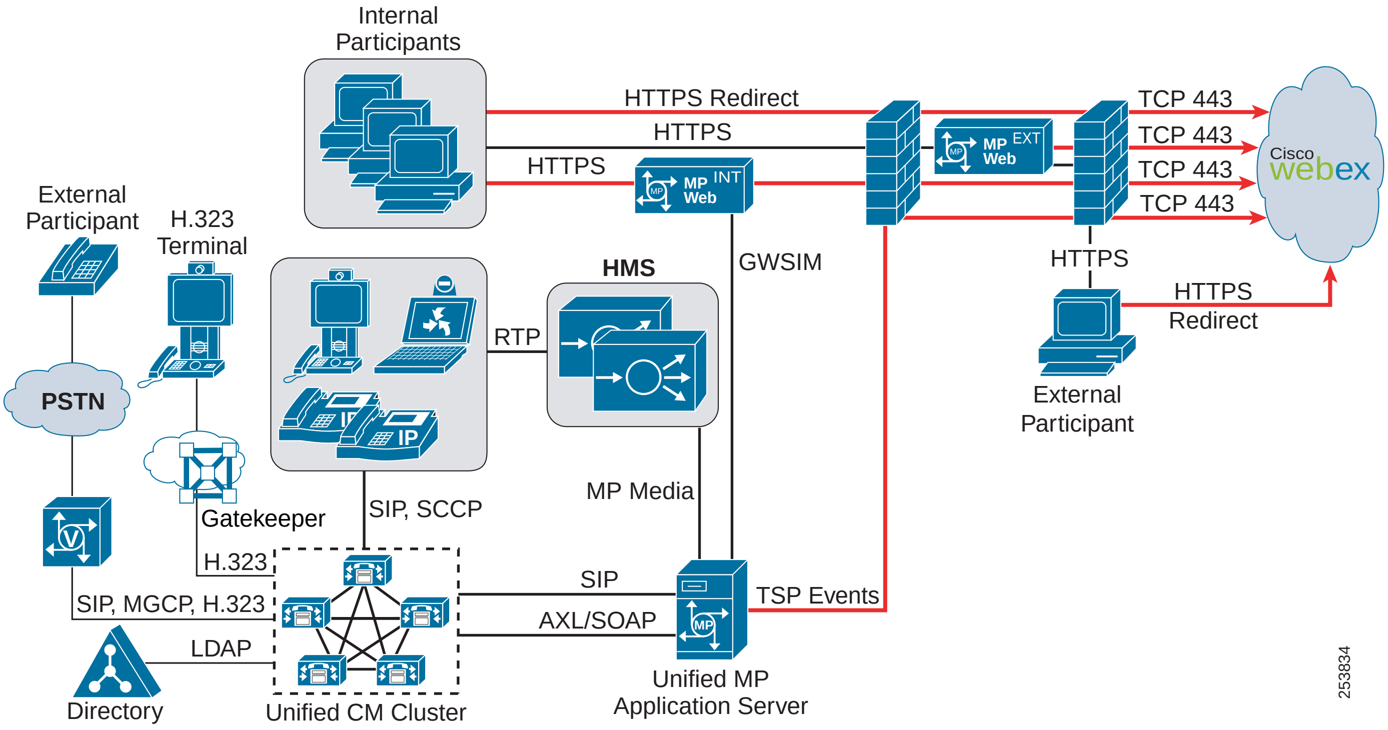
Unified MeetingPlace Owned profile management is available only for existing customers that wish to retain the use of current profiles for use with WebEx. New customers will not be able to provision the WebEx site using the Unified MeetingPlace-to-WebEx SSO integration, which is supported only on installed systems already provisioned in this manner.

If there is no SSO enabled between Unified MeetingPlace and WebEx, all WebEx host accounts must be provisioned by manual export from Unified MeetingPlace to the WebEx site by an administrator (to be updated periodically), and all end-user authentication is provided by the local WebEx host account passwords. WebEx host accounts may also be requested via the WebEx Site and then exported into the Unified MeetingPlace system for profile management. The SSO option must be chosen when ordering the WebEx Site for integration with Unified MeetingPlace on-premises, and it is available only for existing customers who already have Unified MeetingPlace and WebEx installed.

Unified MeetingPlace Scheduling Deployment

The Unified MeetingPlace scheduling deployment option requires the use of two Unified MeetingPlace Web Servers, solely for scheduling and attending meetings. They do not provide any web conferencing functionality. Figure 22-10 illustrates a high-level view of a sample Unified MeetingPlace solution with Unified MeetingPlace scheduling and HMS. Alternatively, an EMS could be used in place of the HMS, and a WebEx Node for MCS is not depicted but could optionally be added as well.

Figure 22-10 Unified MeetingPlace Solution with Unified MeetingPlace Scheduling and HMS



With Unified MeetingPlace scheduling, when users select the click-to-attend URL in an invitation, they first connect with a Unified MeetingPlace Web server customer-configured URL (HTTPS option recommended). The Unified MeetingPlace Web servers immediately initiate a connection to the organization's WebEx site and create a meeting, and the WebEx site returns a join URL which the MeetingPlace Web servers pass onto the clients in the form of a redirect to the WebEx Media Tone Network via secure HTTPS. This redirect behavior is completely transparent to the user, and user authentication is performed solely by the on-premises Unified MeetingPlace system, which is required to enable the SSO capability. The use of the on-premises WebEx Node for MCS is also available for internal users.

When a Unified MeetingPlace profiled user schedules a WebEx meeting or accesses the My WebEx link from the Unified MeetingPlace web user interface, WebEx automatically creates the user account based on the Unified MeetingPlace user profile with the SSO option enabled. The Unified MeetingPlace profile could be either from the local Unified MeetingPlace userID and password or from LDAP integration with Unified CM, which is the most commonly used. Several Unified MeetingPlace user profile attributes are inherited by WebEx, including username, password, first name, last name, telephone number, and email address. Because a WebEx Site is dedicated to a specific customer and the WebEx user profile is based on the Unified MeetingPlace user profile, there should not be any user profile conflicts. No WebEx host accounts are created manually because the Unified MeetingPlace SSO integration provides this function via the WebEx TSP link. Passwords are not sent over the TSP Link to WebEx. WebEx will trust all internal user traffic redirected by the Unified MeetingPlace Web servers. Guest users do not use any passwords or authentication to join WebEx meetings (except the WebEx Meeting Password if configured).

Note Internal WebEx meetings can be recorded with Unified MeetingPlace scheduling, but this requires a WebEx Node for MCS to be deployed on-premises.

Cisco Unified Communications Manager

Cisco Unified Communications Manager (Unified CM) is also a central piece of the architecture, and it provides inbound and callback by means of SIP trunks. A SIP trunk is configured in Unified CM with a destination address of the Unified MeetingPlace Application server(s), and then a route pattern(s) must be used to route calls via the SIP trunk to Unified MeetingPlace. Typically there are three phone numbers that are sent in email notifications for use for dial-in capabilities: Toll free (optional), toll number, and internal Unified CM DN for abbreviated dialing for internal callers. In Unified MeetingPlace there is a separate configuration for callback or outdial feature support by means of SIP trunks to a primary Unified CM subscriber, and subsequent subscribers are used if the primary is not accepting calls due to various conditions. The IP addresses or hostnames of multiple Unified CM call processing subscribers are listed for outbound call delivery in a hunt mode.

It is imperative that the Unified CM servers be able to resolve all dial strings received from a callback request within a WebEx Meeting room after joining. Callbacks may also be disabled system-wide on the WebEx Site by means of Site Administration settings. Unified CM is also in control of all toll restrictions to various countries or other numbers most enterprises will block, because Unified MeetingPlace does not have any toll restriction blocking itself.

In a multinode deployment the Unified CM or Session Management Edition systems are a critical component supporting Unified MeetingPlace in geographically disbursed enterprises. Unified CM clusters with intercluster trunks (ICTs) are required to accommodate Unified MeetingPlace conferencing servers with their unique assigned dial-in numbers and to resolve all calls based on dial plans between sites and to the PSTN for guest or outside mobile users. Guest users can either dial in or use the WebEx callback feature within a meeting room after joining. Multinode Unified MeetingPlace conferencing nodes in a region are configured in a route group in a circular method, where all inbound calls are distributed evenly between all nodes. Callbacks are initiated by the Meeting Director, which chooses the least busy conference node per region based on the timezone of the host of that meeting. The SIP Refer command is used to send dial-in callers to the conferencing node chosen to host that meeting ID.

Additional guidelines for redundancy are described in the section on High Availability, page 22-9. Third-party PBXs can be integrated with Unified MeetingPlace through Unified CM only. For further details on PBX interoperability with Unified CM, refer to the documentation available at

http://www.cisco.com/en/US/solutions/ns340/ns414/ns728/networking_solutions_products_genericcontent0900aecd805b561d.html

Unified MeetingPlace supports receiving both Early Offer (EO) and Delayed Offer (DO) SIP Invite messages. Unified MeetingPlace initiates EO SIP Invites for outbound calls, and Unified CM sends calls to Unified MeetingPlace by using DO SIP invites. Unified CM can be configured to use EO, but this might require the use of a media termination point (MTP) resource. For more information, see SIP Delayed Offer and Early Offer, page 14-21.

Note For Unified MeetingPlace audio/video deployments involving the Express Media Server (EMS), Unified MeetingPlace also supports call delivery by means of a Cisco IOS SIP gateway or Cisco Unified Border Element. LDAP synchronization capabilities are lost with this deployment. For more information, refer to the latest version of the *Planning Guide for Cisco Unified MeetingPlace*, available at http://www.cisco.com/en/US/products/sw/ps5664/ps5669/products_implementation_design_guides_list.html.

Recording

Another criterion for choosing a deployment model is where customers prefer meeting recordings to be stored and accessed. Meeting participants can start audio-only recording via a voice user interface such as a telephone, or they can start audio and web recording from a WebEx meeting room. Audio recording invokes a call event from the WebEx Collaboration cloud to the Unified MeetingPlace Media server via the PSTN voice gateways. For the Unified MeetingPlace scheduling deployment model, the recorded meetings are available from the Unified MeetingPlace Web user interface to download and play back with a WebEx recording playback program. The internal Unified MeetingPlace web server (with optional SAN/NAS) stores recordings that are scheduled as internal meetings. All internal meeting recordings (WebEx audio recordings, audio-only, or audio/video recordings) are stored on-premises. Video recordings are available only with the Hardware Media Server option and the Unified MeetingPlace Scheduling option.

Unified MeetingPlace Scheduling uses the WebEx Network Based Recording (NBR) storage for all meetings that are scheduled as external meetings. However, users access these external recordings via the same method as internal recordings, but the files are simply stored in a different location.

All Unified MeetingPlace and WebEx recordings are played back via the standard NBR recording playback program provided by download to the local users' PCs. All files are editable as well by WebEx editing tools for NBR recordings.

Other Architectural Considerations

Some integration options available with a Unified MeetingPlace Scheduling deployment model may require additional integration servers. Outlook and Exchange calendaring integration is inherently built into the Unified MeetingPlace Application server. However, Lotus Notes integration requires additional software that is co-resident on the Internal Unified MeetingPlace Web server, but other integrations do not require the deployment of the Internal Unified Meeting Web server.

For more information on available Unified MeetingPlace integrations, refer to the latest version of the *Planning Guide for Cisco Unified MeetingPlace*, available at

http://www.cisco.com/en/US/products/sw/ps5664/ps5669/products_implementation_design_guides_list.html

Deployment Options

The majority of Unified MeetingPlace deployments follow a single-site model. This section provides high-level details of each deployment option.

Single-Site Unified MeetingPlace Scheduling Deployment

This deployment model is for current customers who already have the Unified MeetingPlace Web components deployed. The other requirement for deploying this model include using the following features:

- Audio-only or audio/video-only deployments with no WebEx integration
  - Primary/warm standby redundancy is available with this deployment.
- Continuous meetings with blast outdial for audio-only meetings
  - Primary/warm standby redundancy is available with this deployment.
- Unified CM Video Telephony ad-hoc audio/video mixing for conference bridge resources
  - Multiple instances of Unified MeetingPlace in ad-hoc mode can be used per Unified CM cluster. Each Unified CM cluster requires its own Unified MeetingPlace audio-only server(s).
  - Multiple Unified MeetingPlace servers can be configured in hunt fashion on the conference bridge resource group configuration per cluster.
  - Standards-based video will affect overall capacity, depending on the type and bandwidth of video setting on Unified MeetingPlace.

Most deployments use the single-site deployment model, with all server components and users located at a single site interconnected by a single LAN. Solution components vary as discussed in the section on Architecture, page 22-5. Single-site deployments have the following common characteristics:

- The Express Media Server is automatically co-located with the Application server. The optional Unified MeetingPlace Hardware Media Server(s) must be located in the same data center with the active Unified MeetingPlace Application server.
- Network Time Protocol (NTP) must be implemented to allow Unified MeetingPlace components to synchronize their clocks to a network time server or network-capable clock. NTP is a critical network service for Unified MeetingPlace because it ensures accurate time for scheduling meetings. The external NTP source can be specified during Unified MeetingPlace Application server installation, and other Unified MeetingPlace components will synchronize with the application server automatically.
- For existing customer installations only, Unified MeetingPlace Scheduling audio, video, and web recordings and meeting attachments can optionally be stored on an external customer-provided SAN/NAS storage server.
- For deployments with Unified MeetingPlace Scheduling, you must deploy a single Unified MeetingPlace Web server for internal users and a single Unified MeetingPlace Web server located in the DMZ for external participants.
- For deployments with Unified MeetingPlace Scheduling, the round-trip delay between the active Unified MeetingPlace Application server and any Unified MeetingPlace Web server(s) in the solution must not be greater than 150 ms.

- For deployments of WebEx Node for MCS, Cisco recommended placing it on the internal network closest to participants involved in meetings. WebEx Node for MCS does not support HTTPS Proxy servers, therefore it must route directly outbound using TCP port 443 to have access to the WebEx Site.

For a detailed list of incoming and outgoing ports by component, refer to the latest version of the *System Requirements for Cisco Unified MeetingPlace*, available at

http://www.cisco.com/en/US/products/sw/ps5664/ps5669/products_device_support_tables_list.html

High Availability

This section describes redundancy considerations for the following Unified MeetingPlace components:

- Unified MeetingPlace Application Server
- Unified MeetingPlace Media Server (optional)
- Unified MeetingPlace Web Server
- WebEx Node for MCS
- Call Control

Unified MeetingPlace Application Server

Unified MeetingPlace in a multinode deployment with WebEx Scheduling automatically provides active/active resiliency, and customers can choose the level of redundancy per region and site. Regions can be configured to overflow to other regions if desired.

Unified MeetingPlace with the MeetingPlace Scheduling model allows for an active (primary) and a single warm standby Unified MeetingPlace Application server for failover. Each Unified MeetingPlace Application server in a failover deployment is configured with the same IP address associated to its physical network interface controller (NIC) and a unique IP address associated to a virtual network interface. The requirement for both Unified MeetingPlace Application servers to share the same IP address mandates both Application servers to be connected to the same virtual LAN (VLAN) or IP subnet. This is not an issue when both servers are placed in a single data center; however, a dual data center design is supported only if the same VLAN (IP subnet) spans both data centers. All Unified MeetingPlace components as well as Unified CM communicate with this shared IP address. The physical NIC (with the shared IP address) of the standby server remains disabled until the primary server fails and the manual failover process is initiated by IT personnel.

For network requirements in deploying either multinode or a standby server, refer to the failover information in the latest version of the *Planning Guide for Cisco Unified MeetingPlace*, available at

http://www.cisco.com/en/US/products/sw/ps5664/ps5669/products_implementation_design_guides_list.html

The virtual network interface is used for Informix database replication between the primary and standby servers. The database replication ensures that database tables related to users, groups, and meetings are synchronized between primary and standby servers. Cisco recommends placing the virtual network interfaces of the active and standby servers in the same VLAN. For further information regarding Unified MeetingPlace Application server redundancy, refer to the latest version of the *Planning Guide for Cisco Unified MeetingPlace*, available at

http://www.cisco.com/en/US/products/sw/ps5664/ps5669/products_implementation_design_guides_list.html

Another key requirement for a Unified MeetingPlace solution is that the active Unified MeetingPlace Application server must be co-located with the active Unified MeetingPlace Media server(s). Since the Express Media Server runs in software on the Unified MeetingPlace Application server itself, failover to the standby Unified MeetingPlace Application server results in using EMS capabilities on the standby. In the case of Hardware Media Servers, there are some considerations when looking at single data center designs compared to dual data center designs.

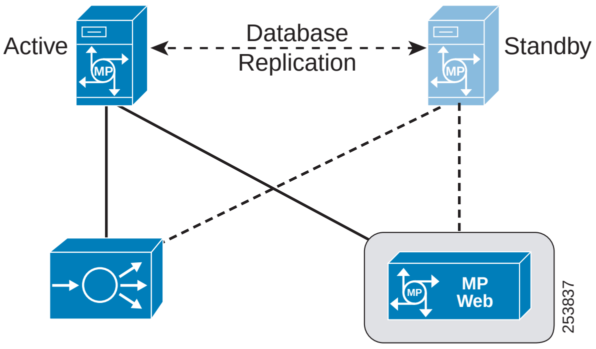
Single Data Center Design

In a single data center design, multinode resiliency is automatically available in an active/active mode, and meetings are evenly distributed by the Meeting Director component between both nodes. If failure occurs on one conferencing node, calls will be dropped, and when users dial back into that same meeting ID or use the WebEx Callback feature in the meeting room GUI, then those meetings are automatically established on another node in that region or they overflow to another region if configured. Up to four conferencing nodes per site may be deployed.

With the Unified MeetingPlace Scheduling model, failover of the Unified MeetingPlace Application server occurs within the same geographic location. For this type of deployment, there would typically be one set of Unified MeetingPlace Hardware Media servers shared by the primary and standby Unified MeetingPlace Application servers. If the primary Unified MeetingPlace Application server fails, the Unified MeetingPlace Media server(s) must be synchronized with the standby (now primary) server. Unified MeetingPlace Web server(s) would also be shared for a Unified MeetingPlace scheduling deployment. Figure 22-11 illustrates the failover process for the Unified MP Application server in a single data center deployment.

Note For highly redundant solutions, it is also possible to have a set of standby Unified MeetingPlace Media servers and Web Collaboration servers in a single data center. Unified MeetingPlace Web servers cannot be made redundant with Unified MeetingPlace 8.*x* systems. The WebEx Scheduling Deployment model offers a more reliable redundant deployment model.

Figure 22-11 Failover of a Unified MeetingPlace Application Server in a Single Data Center Deployment



Dual Data Center Design

In a dual data center design, the WebEx Scheduling model with multinode conferencing nodes provides active/active failover per region, or overflow to other regions can be configured as well. Four regions with two sites per region is supported with a maximum of 14 conferencing nodes deployed for active/active load sharing in multiple data centers, based on customer requirements. If a conferencing node fails, audio calls are dropped, and when users call back in or use the WebEx Callback GUI feature from within the meeting room, the meetings are automatically started on an active node with capacity. All conferencing nodes within a region can be used to distribute calls, and overflow to another region is based on optional system administration settings.

With the Unified MeetingPlace Scheduling model, failover of the Unified MeetingPlace Application server occurs between different geographic locations across an IP WAN. Again, although both servers are separated geographically, both the active and standby Application servers must be connected to the same VLAN to ensure proper failover operation. For this type of deployment, the standby Application server must be co-located with a redundant Unified MeetingPlace Hardware Media server(s) with which it is synchronized. If the identical number of Unified MeetingPlace Media server audio and video blades is not maintained in the standby data center, system capacity will be reduced during failover scenarios where the standby Application server is promoted to active.

Unified MeetingPlace Media Server

Since the Express Media Server runs in software on the Unified MeetingPlace Application server itself, in a multinode deployment model, any conferencing node in a region can be used for taking those additional meetings. A maximum of four servers per site, two sites per region, and four regions may be deployed for a globally distributed architecture.

Since the Express Media Server runs in software on the Unified MeetingPlace Application server itself, failover to the standby Application server will result in using EMS capabilities on the standby. EMSs do not support cascading or clustering to other EMS instances. A maximum of one primary and one failover Unified MeetingPlace Application and EMS server is supported with Unified MeetingPlace solutions with either Unified MeetingPlace Scheduling or WebEx Scheduling deployment models. Active RSNA failover is not supported with any WebEx integrations (only standalone audio/video deployments).

The Unified MeetingPlace Application Server automatically performs failover to alternate HMSs (audio or video blades) in the system. For example, if the Application Server detects a loss of connectivity with an audio blade, it removes it from the list of active audio blades so that subsequent audio sessions will connect to an active audio blade. To avoid reduction in Unified MeetingPlace Media Server capacity during an audio or video blade outage, one option is to add additional HMS audio and video blades to the solution. The Application Server will not exceed the number of sessions for which it is licensed. Another option is to revert to the standby Unified MeetingPlace Application Server with its own set of HMSs (as in a dual data center design). These two options are not mutually exclusive; a standby Unified MeetingPlace Application Server with its own set of HMSs can gain further redundancy by adding more audio or video blades.

For further information regarding Hardware Media Server failover, refer to the latest version of the *Planning Guide for Cisco Unified MeetingPlace*, available at

> http://www.cisco.com/en/US/products/sw/ps5664/ps5669/products_implementation_design_guides_list.html

Unified MeetingPlace Web Server

The Unified MeetingPlace Scheduling model uses only one Web server with audio-only configuration for recordings and/or the Web scheduling interface. For existing customers using WebEx Integration to migrate to Unified MeetingPlace 8.5 (or later release) and still using the Unified MeetingPlace Scheduling model, then use an additional Web server deployed in a DMZ. Each Cisco Unified MeetingPlace system can have a maximum of one internal Web server and one Web server in the DMZ if using WebEx Integration only. There are no redundancy options for these servers. Unified MeetingPlace Web servers are implemented only for solutions incorporating the Unified MeetingPlace scheduling interface. The Unified MeetingPlace Lotus Notes or Jabber integration also cannot be made redundant.

WebEx Node for MCS or ASR

A Unified MeetingPlace solution supports unlimited nodes if WebEx Node for ASR is used, but the maximum number of supported WebEx Nodes for MCS depends on the deployment options. For single Unified MeetingPlace Application server deployments, the solution supports a maximum of three WebEx Nodes for MCS. For multimode deployments with WebEx scheduling, the solution supports a maximum of four WebEx Nodes for MCS. Cascading meetings are supported across the WebEx Nodes for MCS and out to the WebEx Collaboration cloud. WebEx Nodes for MCS or ASR will each automatically provide a level of redundancy in case of a single node outage. After receiving a list of Meeting Zone URLs, the client then pings all the Meeting Zones URLs to determine the closest node. If a node does not respond, no clients will connect to this node. All internal users (even those that use VPN from remote locations) can connect to any of the WebEx Node for MCS servers.

If the WebEx Node for MCS or ASR that is hosting a meeting becomes unavailable, the next available WebEx Node for MCS or ASR automatically takes over. Any sharing and recordings will be stopped, and users will have to restart sharing and recording the meetings. When a customer has multiple WebEx Nodes for MCS or ASR active within a meeting with a subset of users on each node, content is cascaded between the WebEx Nodes for MCS or ASR. When there are three or more WebEx Nodes for MCS active in the same meeting, the cascade appears as a star with the WebEx Node for MCS that the host is on at its center. If a node fails, the clients automatically rejoin other nodes using the list presented to the client from WebEx within the client entry meeting window, with little or no effect to the end user. External scheduled meetings also allow for internal users to connect to the WebEx cloud as well, while internal scheduled meetings always stay internal on other redundant WebEx nodes (which can be distributed or co-located, depending on customer network design requirements). Audio calls remain intact on the Unified MeetingPlace system on-premises.

For more information on redundancy within the WebEx cloud, see High Availability, page 22-9.

Call Control

Unified MeetingPlace allows you to define multiple SIP outdial connections that point to Cisco Unified CM call processing subscribers. For redundancy, multiple SIP proxy servers should be configured to direct calls to call processing subscribers in the Unified CM cluster. These call processing subscribers should correlate with the Unified CM Group of the configured SIP trunk for Unified MeetingPlace calls in Unified CM. Note that the Unified MeetingPlace Application server will send outbound calls to SIP proxy server 1 only and will not send calls to SIP proxy server 2 unless communication with SIP proxy server 1 is lost. Only then will Unified MeetingPlace send a SIP INVITE message to the next available call processing agent in the list. Failure of the call processing agent should not affect existing calls. The existing media connection is torn down after the user disconnects.

Note The term SIP Proxy Server is simply the terminology seen on the Unified MeetingPlace Application Server configuration pages, and it does not imply that integration with any SIP Proxy server is supported.

For inbound calls, a single configured SIP trunk in Unified CM can be handled by up to three call processing subscribers found in its configured Unified CM Group. If the primary Unified CM call processing subscriber in the Unified CM Group is offline, the second one will take over initiating calls into the Unified MeetingPlace system. For more information, see Cisco Unified CM Trunks, page 14-1. For Unified MeetingPlace scheduling deployments with EMS, multiple Cisco IOS SIP gateways are required to provide redundancy for call delivery.

Capacity Planning

The capacity of a given Unified MeetingPlace solution depends on the design of the Cisco Unified Communications system (for example, audio codecs or video format used in conferencing) and the platform selected to run the Unified MeetingPlace solution components. For capacity planning details, see the sizing information in the section on Collaborative Conferencing, page 29-47.

Network Traffic Planning

Network traffic planning for Unified MeetingPlace collaboration consists of the following elements:

- Call Control Bandwidth

  Call control bandwidth is extremely small but critical. Co-locating the Unified MeetingPlace Application server with Unified CM helps protect against issues with call control. Remote locations need proper QoS provisioning to ensure reliable operation.

- Real-Time Transport Protocol (RTP) Traffic Bandwidth

  RTP traffic consists of voice and video traffic. The Unified MeetingPlace Media servers supports G.711, G.729, G.722, and iLBC as audio codecs, and it supports a wide range of video codecs and bandwidths. For further information regarding bandwidth calculations per codec type, refer to the chapters on Network Infrastructure, page 3-1, and IP Video Telephony, page 12-1.

- Web Collaboration Bandwidth

  Web collaboration bandwidth for a Unified MeetingPlace solution can be estimated the same way as for a WebEx SaaS solution. See Network Traffic Planning, page 22-10.

Design Considerations

The following design considerations apply to Unified MeetingPlace deployments:

- Only a single Unified MeetingPlace system is supported per WebEx site.
- For scenarios where any Unified MeetingPlace solution components are separated by network firewalls, it is imperative to ensure the correct pinholes are opened for all required traffic. For a detailed ports list, refer to the network requirements information in the latest version of the *System Requirements for Cisco Unified MeetingPlace*, available at

  http://www.cisco.com/en/US/products/sw/ps5664/ps5669/products_device_support_tables_list.html

- Collaborative meeting systems typically result in increased top-of-the-hour call processing load. Capacity planning tools with specific parameters for Unified MeetingPlace are available to Cisco partners and employees to help calculate the capacity of the Cisco Unified Communications System for large configurations. Contact your Cisco partner or Cisco Systems Engineer (SE) for assistance with sizing of your system. For Cisco partners and employees, the Cisco Unified Communications Sizing Tool is available at http://tools.cisco.com/cucst.
- For more detail on the various Cisco collaborative client offerings and how they fit into collaborative conferencing solutions, see Cisco Collaboration Clients and Applications, page 24-1.
- Call admission control with Unified MeetingPlace is performed by Unified CM. With locations-based call admission control, Unified CM can control bandwidth to the Unified MeetingPlace system by placing the SIP trunk specific to Unified MeetingPlace in a location with a set amount audio and/or video bandwidth allowed. Alternatively, Unified CM supports the use of

Resource Reservation Protocol (RSVP), which can also provide call admission control. For further information regarding call admission control strategies, see the chapter on Call Admission Control, page 11-1.

- Unified MeetingPlace supports the following standard dual-tone multi-frequency (DTMF) transmission methods: RFC 2833 and KPML DTMF. Unified CM supports RFC 2833, and it is the recommended method for DTMF Relay.
- SIP signaling traffic from the Unified MeetingPlace Application server is marked CS3 (DSCP 0x18). However other traffic from the Unified MeetingPlace Application server, such as communications with Unified MeetingPlace Web servers, Media Servers, or the WebEx Site, are marked best-effort (DSCP 0x00). If any of this traffic is traversing low-speed or congested links, QoS considerations should be taken into account.
- The audio streams from the Unified MeetingPlace Media servers are marked EF (DSCP 0x2E), and the video streams are marked AF41 (DSCP 0x22) by default. These values are configurable from Unified MeetingPlace Administration.
- Web conferencing traffic is encrypted in SSL and is always marked best-effort (DSCP 0x00).
- The Unified MeetingPlace Meeting Director TSP component initiates dual outbound TCP port 443 connections to the WebEx Site and also provides SOCKS proxy server support.
- The Unified MeetingPlace WebEx Node for MCS or ASR initiates an outbound TCP port 443 connection to the WebEx Site but does not support any HTTPS proxy server. The WebEx Node for MCS or ASR must be allowed to connect directly to the WebEx Site without a proxy.

CHAPTER 23

Cisco IM and Presence

Revised: August 31, 2012; OL-27282-05

Cisco IM and Presence consists of many components that enhance the value of a Cisco Unified Communications system. The main presence component of the solution is the Cisco IM and Presence Service, which incorporates the Jabber Extensible Communications Platform and supports SIP/SIMPLE and Extensible Messaging and Presence Protocol (XMPP) for collecting information regarding a user's availability status and communications capabilities. The user's availability status indicates whether or not the user is actively using a particular communications device such as a phone. The user's communications capabilities indicate the types of communications that user is capable of using, such as video conferencing, web collaboration, instant messaging, or basic audio.

The aggregated user information captured by the Cisco IM and Presence Service enables Cisco Jabber, Cisco Unified Communications Manager applications, and third-party applications to increase user productivity. These applications help connect colleagues more efficiently by determining the most effective form of communication.

This chapter explains the basic concepts of presence and instant messaging within the Cisco Unified Communications System and provides guidelines for how best to deploy the various components of the presence and instant messaging solution. The Cisco IM and Presence Service must be deployed with Cisco Unified Communications Manager (Unified CM) 9.0 or later releases; Cisco Unified CM 8.*x* and earlier releases interoperate with Cisco Unified Presence.

This chapter covers the following topics:

What's New in This Chapter

Table 23-1 lists the topics that are new in this chapter or that have changed significantly from previous releases of this document.

Table 23-1 New or Changed Information Since the Previous Release of This Document

New or Revised Topic	Described in:	Revision Date
The Cisco IM and Presence publisher node must be co-located with the Unified CM publisher, but the IM and Presence subscriber nodes can be deployed with clustering over the WAN.	Design Considerations for Cisco IM and Presence, page 23-38	August 31, 2012
Migration from Cisco Unified Presence to Cisco IM and Presence	Cisco IM and Presence Migration, page 23-26	June 28, 2012
Product name change from Cisco Unified Presence to Cisco IM and Presence Service, and from Cisco Unified Personal Communicator to Cisco Jabber	All sections of this chapter	June 28, 2012

Presence

Presence refers to the ability and willingness of a user to communicate across a set of devices. It involves the following phases or activities:

- Publish user status

  User status changes can be published automatically by recognizing user keyboard activity, phone use, or device connectivity to the network.

- Collect this status

  The published information is gathered from all the available sources, privacy policies are applied, and then current status is aggregated, synchronized, and stored for consumption.

- Consume the information

  Desktop applications, calendar applications, and devices can use the user status information to provide real-time updates for the end users to make better communication decisions.

Status combines the capabilities of what the device or user can do (voice, video, instant messaging, web collaboration, and so forth) and the attributes showing the state of the device or user (available, busy, on a call, and so forth). Presence status can be derived from automatic events such as client login and telephone off-hook, or it can be derived from explicit notification events for changing status such as the user selecting Do Not Disturb from a change-status pick list.

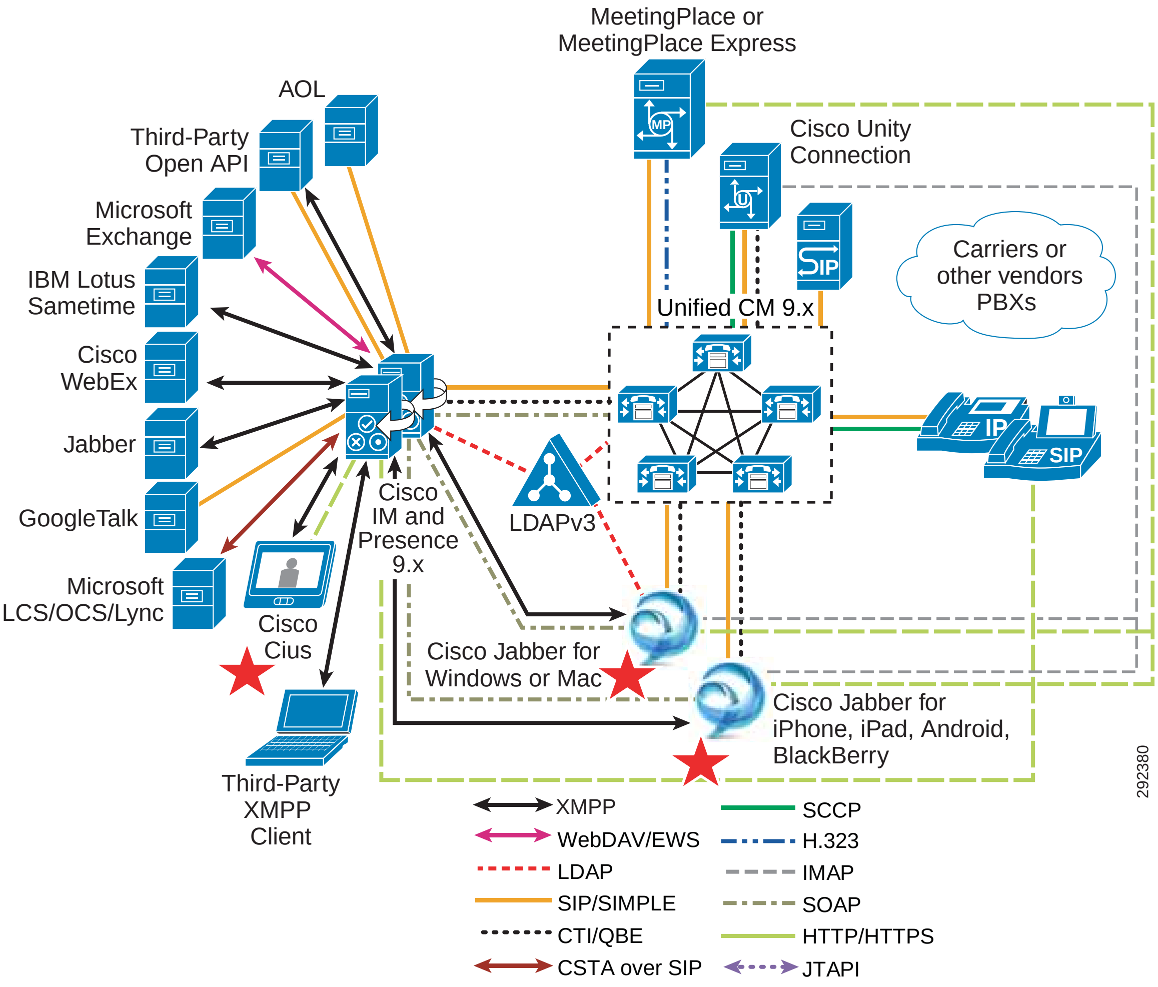
Terminology surrounding presence refers to a watcher, presence entity (*presentity*), and presence server. The presence entity publishes its current status to the presence server by using a PUBLISH or REGISTER message for SIP/SIMPLE clients, or by using an XML Presence Stanza for XMPP clients. It can be a directory number (DN) or a SIP uniform resource identifier (URI) that resides within or outside the communications cluster. A *watcher* (device or user) requests presence status about a presence entity by sending a message to the presence server. The presence server responds to the watcher with a message containing the current status of the presence entity.

Cisco IM and Presence Components

Cisco IM and Presence encompasses the following components, illustrated in Figure 23-1:

- Cisco IM and Presence Service
- Cisco Unified Communications Manager (Unified CM)
- Cisco Jabber
- Cisco Unified MeetingPlace or MeetingPlace Express
- Cisco Unity or Unity Connection
- Cisco Unified Videoconferencing or Cisco Unified MeetingPlace Express VT
- Lightweight Directory Access Protocol (LDAP) Server v3.0
- Cisco Unified IP Phones
- Third-party presence server
- Third-party XMPP clients
- Third-party applications

Figure 23-1 Cisco IM and Presence Interfaces



Cisco IM and Presence User

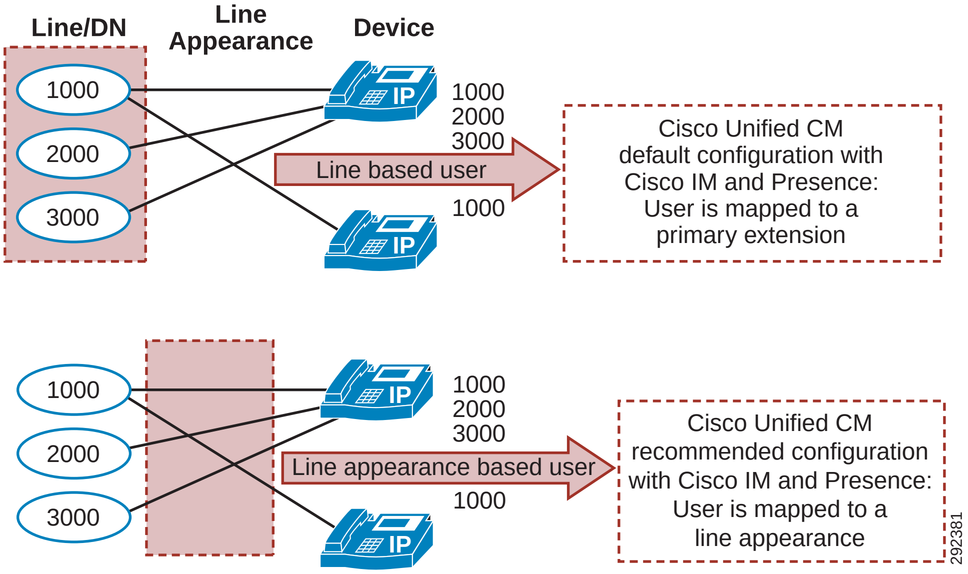
For presence, typically a user is described in terms of the user's presence status, the number of users on the system, or the user's presence capabilities.

As defined by Cisco IM and Presence, a user is specified in Cisco Unified CM by default as an *end user* and must be configured with a primary extension. The user is effectively tied to a directory number, and the presence status is reflected for the user's primary extension rather than for the device to which the user is associated. (See Figure 23-2.)

A user, specified in Unified CM as an *end user*, can be configured with a primary extension or associated with a line appearance. When using the CUP PUBLISH Trunk service parameter on Unified CM, you must associate the user with a line appearance rather than just a primary extension. With the line appearance, the user is effectively tied to a line appearance (directory number associated with a particular device), which allows for a more detailed level of granularity for aggregation of presence information. The user can be mapped to multiple line appearances, and each line appearance can have multiple users (up to 5). Cisco recommends associating the end user with a line appearance. (See Figure 23-2.)

A user can also have only IM and Presence capabilities. This deployment requires a Unified CM publisher for configuration, along with an IM and Presence cluster as well as Jabber clients. In this arrangement, telephony voice features would be provided through another vendor’s system or PBX.

Figure 23-2 Associating an End User with a Primary Extension or Line Appearance


Line/DN
Line Appearance
Device
1000
2000
3000
IP
1000
2000
3000
Line based user
1000
Cisco Unified CM default configuration with Cisco IM and Presence: User is mapped to a primary extension
1000
2000
3000
1000
2000
3000
Line appearance based user
1000
Cisco Unified CM recommended configuration with Cisco IM and Presence: User is mapped to a line appearance
292381

The concept of a *presence user* appears throughout this chapter; therefore, keep in mind the meaning of a user as defined for Cisco IM and Presence.

Busy Lamp Field (BLF)

All telephony presence requests for users, whether inside or outside the cluster, are processed and handled by Cisco Unified CM.

A Unified CM watcher that sends a presence request will receive a direct response, including the presence status, if the watcher and presence entity are co-located within the Unified CM cluster.

If the presence entity exists outside the cluster, Unified CM will query the external presence entity through the SIP trunk. If the watcher has permission to monitor the external presence entity based on the SUBSCRIBE calling search space and presence group (both described in the section on Unified CM Presence Policy, page 23-8), the SIP trunk will forward the presence request to the external presence entity, await the presence response from the external presence entity, and return the current presence status to the watcher.

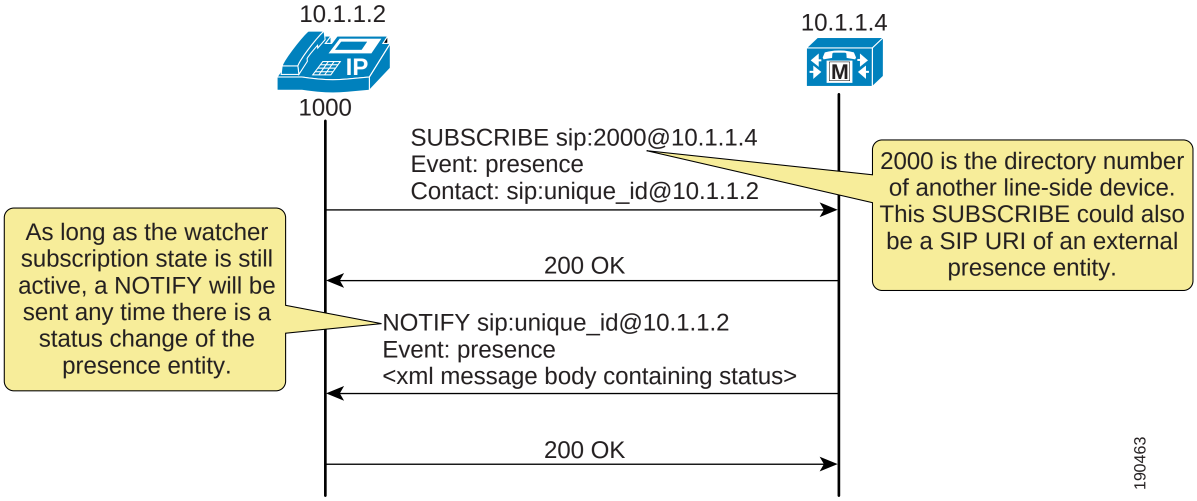
A watcher that is not in a Unified CM cluster can send a presence request to a SIP trunk. If Unified CM supports the presence entity, it will respond with the current presence status. If Unified CM does not support the presence entity, it will reject the presence request with a SIP error response.

Unified CM Presence with SIP

Unified CM uses the term *SIP line* to represent endpoints supporting SIP that are directly connected and registered to Unified CM and the term *SIP trunk* to represent trunks supporting SIP. SIP line-side endpoints acting as presence watchers can send a SIP SUBSCRIBE message to Unified CM requesting the presence status of the indicated presence entity.

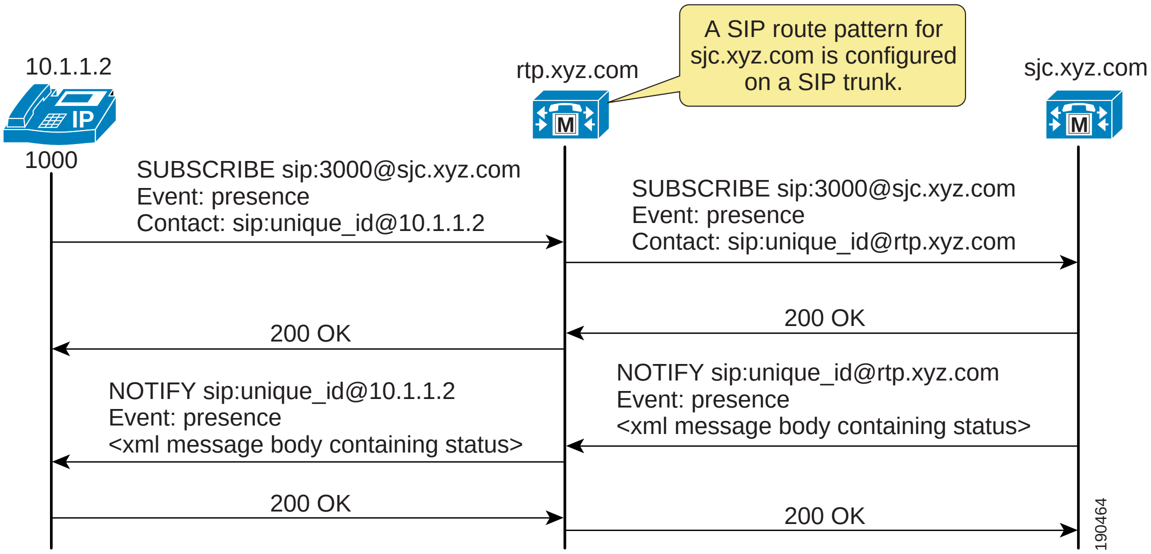
If the presence entity resides within the Unified CM cluster, Unified CM responds to a SIP line-side presence request by sending a SIP NOTIFY message to the presence watcher, indicating the current status of the presence entity. (See Figure 23-3.)

Figure 23-3 SIP Line SUBSCRIBE/NOTIFY Exchange


10.1.1.2
IP
1000
10.1.1.4
M
SUBSCRIBE sip:2000@10.1.1.4
Event: presence
Contact: sip:unique_id@10.1.1.2
2000 is the directory number of another line-side device. This SUBSCRIBE could also be a SIP URI of an external presence entity.
200 OK
As long as the watcher subscription state is still active, a NOTIFY will be sent any time there is a status change of the presence entity.
NOTIFY sip:unique_id@10.1.1.2
Event: presence
<xml message body containing status>
200 OK
190463

If the presence entity resides outside the Unified CM cluster, Unified CM routes a SUBSCRIBE request out the appropriate SIP trunk, based on the SUBSCRIBE calling search space, presence group, and SIP route pattern. When Unified CM receives a SIP NOTIFY response on the trunk, indicating the presence entity status, it responds to the SIP line-side presence request by sending a SIP NOTIFY message to the presence watcher, indicating the current status of the presence entity. (See Figure 23-4.)

Figure 23-4 SIP Trunk SUBSCRIBE/NOTIFY Exchange


A SIP route pattern for sjc.xyz.com is configured on a SIP trunk.
10.1.1.2
rtp.xyz.com
sjc.xyz.com
IP
M
M
1000
SUBSCRIBE sip:3000@sjc.xyz.com
Event: presence
Contact: sip:unique_id@10.1.1.2
SUBSCRIBE sip:3000@sjc.xyz.com
Event: presence
Contact: sip:unique_id@rtp.xyz.com
200 OK
200 OK
NOTIFY sip:unique_id@rtp.xyz.com
Event: presence
<xml message body containing status>
NOTIFY sip:unique_id@10.1.1.2
Event: presence
<xml message body containing status>
200 OK
200 OK
190464

SUBSCRIBE messages for any directory number or SIP URI residing outside the Unified CM cluster are sent or received on a SIP trunk in Unified CM. The SIP trunk could be an interface to another Unified CM or it could be an interface to the Cisco IM and Presence Service.

Unified CM Presence with SCCP

Unified CM supports Skinny Client Control Protocol (SCCP) line-side endpoints acting as presence watchers. There are no SCCP trunks. SCCP endpoints can request presence status of the indicated presence entity by sending SCCP messages to Unified CM.

If the presence entity resides within the Unified CM cluster, Unified CM responds to the SCCP line-side presence request by sending SCCP messages to the presence watcher, indicating the current status of the presence entity.

If the presence entity resides outside the Unified CM cluster, Unified CM routes a SUBSCRIBE request out the appropriate SIP trunk, based on the SUBSCRIBE calling search space, presence group, and SIP route pattern. When Unified CM receives a SIP NOTIFY response on the trunk, indicating the presence entity status, it responds to the SCCP line-side presence request by sending SCCP messages to the presence watcher, indicating the current status of the presence entity.

Unified CM Speed Dial Presence

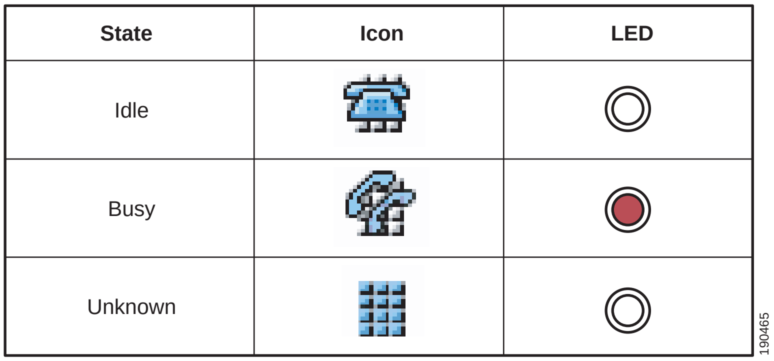
Unified CM supports the ability for a speed dial to have presence capabilities by means of a busy lamp field (BLF) speed dial. BLF speed dials work as both a speed dial and a presence indicator. However, only the system administrator can configure a BLF speed dial; a system user is not allowed to configure a BLF speed dial.

The administrator must configure the BLF speed dial with a target directory number that is resolvable to a directory number within the Unified CM cluster or a SIP trunk destination. BLF SIP line-side endpoints can also be configured with a SIP URI for the BLF speed dial, but SCCP line-side endpoints cannot be configured with a SIP URI for BLF speed dial. The BLF speed dial indication is a line-level indication and not a device-level indication.

For a listing of the phone models that support BLF speed dials, consult the Cisco Unified IP Phone administration guides available on http://www.cisco.com/.

Figure 23-5 lists the various types of BLF speed dial indications from the phones.

Figure 23-5 Indicators for Speed Dial Presence



State	Icon	LED
Idle	☎	○
Busy	☏	●
Unknown	▦	○

Unified CM Call History Presence

Unified CM supports presence capabilities for call history lists (the Directories button on the phone). Call history list presence capabilities are controlled via the **BLF for Call Lists** Enterprise Parameter within Unified CM Administration. The **BLF for Call Lists** Enterprise Parameter impacts all pages using the phone Directories button (Missed, Received, and Placed Calls, Personal Directory, or Corporate Directory), and it is set on a global basis.

For a listing of the phone models that support presence capabilities for call history lists, consult the Cisco Unified IP Phone administration guides available on http://www.cisco.com/.

The presence indicators for call history lists are the same as those listed in the Icon column in Figure 23-5; however, no LED indications are available.

Unified CM Presence Policy

Unified CM provides the capability to set policy for users who request presence status. You can set this policy by configuring a calling search space specifically to route SIP SUBSCRIBE messages for presence status and by configuring presence groups with which users can be associated to specify rules for viewing the presence status of users associated with another group.

Unified CM Subscribe Calling Search Space

The first aspect of presence policy for Unified CM is the SUBSCRIBE calling search space. Unified CM uses the SUBSCRIBE calling search space to determine how to route presence requests (SUBSCRIBE messages with the Event field set to Presence) that come from the watcher, which could be a phone or a trunk. The SUBSCRIBE calling search space is associated with the watcher and lists the partitions that the watcher is allowed to "see." This mechanism provides an additional level of granularity for the presence SUBSCRIBE requests to be routed independently from the normal call-processing calling search space.

The SUBSCRIBE calling search space can be assigned on a device basis or on a user basis. The user setting applies for originating subscriptions when the user is logged in to the device through Extension Mobility or when the user is administratively assigned to the device.

With the SUBSCRIBE calling search space set to <None>, BLF speed dial and call history list presence status does not work and the subscription messages is rejected as “user unknown.” When a valid SUBSCRIBE calling search space is specified, the indicators work and the SUBSCRIBE messages are accepted and routed properly.

Note Cisco strongly recommends that you do not leave any calling search space defined as <None>. Leaving a calling search space set to <None> can introduce presence status or dialing plan behavior that is difficult to predict.

Unified CM Presence Groups

The second aspect of the presence policy for Unified CM is presence groups. Devices, directory numbers, and users can be assigned to a presence group, and by default all users are assigned to the Standard Presence Group. A presence group controls the destinations that a watcher can monitor, based on the user’s association with their defined presence group (for example, Contractors watching Executives is disallowed, but Executives watching Contractors is allowed). The presence group user setting applies for originating subscriptions when the user is logged in to the device via Extension Mobility or when the user is administratively assigned to the device.

When multiple presence groups are defined, the Inter-Presence Group Subscribe Policy service parameter is used. If one group has a relationship to another group via the Use System Default setting rather than being allowed or disallowed, this service parameter's value will take effect. If the Inter-Presence Group Subscribe Policy service parameter is set to **Disallowed**, Unified CM will block the request even if the SUBSCRIBE calling search space allows it. The Inter-Presence Group Subscribe Policy service parameter applies only for presence status with call history lists and is not used for BLF speed dials.

Presence groups can list all associated directory numbers, users, and devices if you enable dependency records. Dependency records allow the administrator to find specific information about group-level settings. However, use caution when enabling the Dependency Record Enterprise parameter because it could lead to high CPU usage.

Unified CM Presence Guidelines

Unified CM enables the system administrator to configure and control user phone state presence capabilities from within Unified CM Administration. Observe the following guidelines when configuring presence within Unified CM:

- Select the appropriate model of Cisco Unified IP Phones that have the ability to display user phone state presence status.
- Define a presence policy for presence users.
  - Use SUBSCRIBE calling search spaces to control the routing of a watcher presence-based SIP SUBSCRIBE message to the correct destinations.
  - Use presence groups to define sets of similar users and to define whether presence status updates of other user groups are allowed or disallowed.
- Call history list presence capabilities are enabled on a global basis; however, user status can be secured by using a presence policy.
- BLF speed dials are administratively controlled and are not impacted by the presence policy configuration.

Note Cisco Business Edition can be used in ways similar to Unified CM to configure and control user presence capabilities. For more information, refer to the chapter on Call Processing, page 8-1.

Cisco IM and Presence Architecture

The Cisco IM and Presence Service uses standards-based SIP, SIP for Instant Messaging and Presence Leveraging Extensions (SIMPLE), and Extensible Messaging and Presence Protocol (XMPP) to provide a common demarcation point for integrating clients and applications into the Cisco Unified Communications System. Cisco IM and Presence also provides an HTTP interface that has a configuration interface through Simple Object Access Protocol (SOAP); a presence interface through Representational State Transfer (REST); and a presence, instant messaging, and roster interface through the Cisco AJAX XMPP Library (CAXL). The Cisco AJAX XMPP Library interface communicates to the Bidirectional-streams Over Synchronous HTTP (BOSH) interface on the Extensible Communications Platform within Cisco IM and Presence. The Cisco IM and Presence Service collects, aggregates, and distributes user capabilities and attributes using these standards-based SIP, SIMPLE, XMPP, and HTTP interfaces.

Cisco or third-party applications can integrate with presence and provide services that improve the end-user experience and efficiency The core components of the Cisco IM and Presence Service consist of: the Jabber Extensible Communications Platform (XCP), which handles presence, instant messaging, roster, routing, policy, and federation management; the Rich Presence Service, which handles presence state gathering, network-based rich presence composition, and presence-enabled routing functionality; and support for ad-hoc group chat storage with persistent chat and message archiving handled to an external database. If persistent chat is enabled, ad-hoc rooms are stored to the external PostgreSQL database for the duration of the ad-hoc chat. This allows a room owner to escalate an ad-hoc chat to a persistent chat; otherwise, these ad-hoc chats are purged from PostgreSQL at the end of the chat. If persistent chat is disabled, ad-hoc chats are stored in volatile memory for the duration of the chat.

Applications (either Cisco or third-party) can integrate presence and provide services that improve the end user experience and efficiency. In addition, Cisco Jabber is a supported client of the Cisco IM and Presence Service that also integrates instant messaging and presence status.

The Cisco IM and Presence Service also contains support for interoperability with Microsoft Live Communications Server 2005, Microsoft Office Communications Server 2007, and the Microsoft Office Communicator client for any Cisco Unified IP Phone connected to a Unified CM. The Microsoft Office Communicator client interoperability includes click-to-dial functionality, phone control capability, and presence status of Cisco Unified IP Phones.

Cisco IM and Presence Cluster

The Cisco IM and Presence Service uses the same underlying appliance model and hardware used by Unified CM as well as Unified CM on the Cisco Unified Computing System (UCS) platform, including a similar administration interface. For details on the supported platforms, refer to the *Cisco Unified Communications Compatibility Tool*, available at

http://tools.cisco.com/ITDIT/vtgsca/VTGServlet

A Cisco IM and Presence cluster consists of up to six servers, including one designated as a publisher, which utilize the same architectural concepts as the Unified CM publisher and subscriber. Within a Cisco IM and Presence cluster, individual servers can be grouped to form a subcluster, and the subcluster can have at most two servers associated with it. Figure 23-6 shows the basic topology for a Cisco IM and Presence cluster, while Figure 23-7 shows a highly available topology. The Cisco IM and Presence cluster can also have mixed subclusters, where one subcluster is configured with two servers while other subclusters contain a single server, as shown in Figure 23-8.

Figure 23-6 Basic Deployment of Cisco IM and Presence

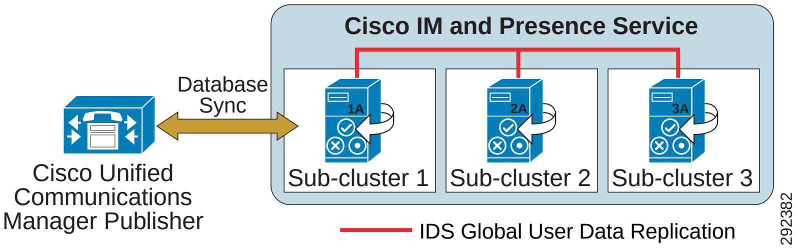

Cisco IM and Presence Service
Database Sync
1A
2A
3A
Cisco Unified Communications Manager Publisher
Sub-cluster 1
Sub-cluster 2
Sub-cluster 3
IDS Global User Data Replication
292382

Figure 23-7 High Availability Deployment of Cisco IM and Presence

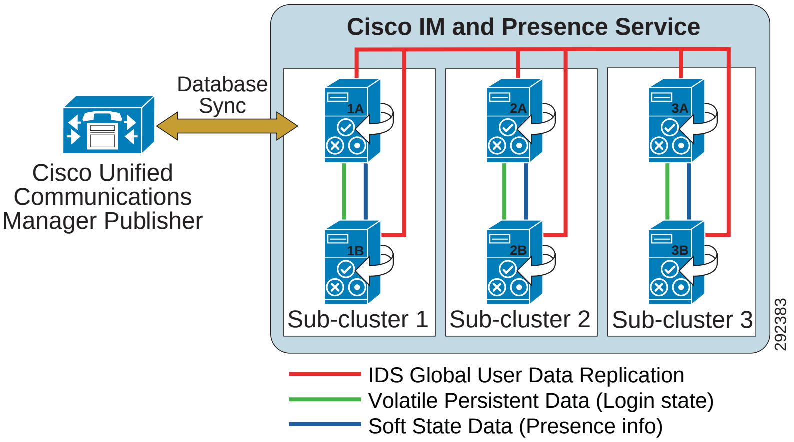
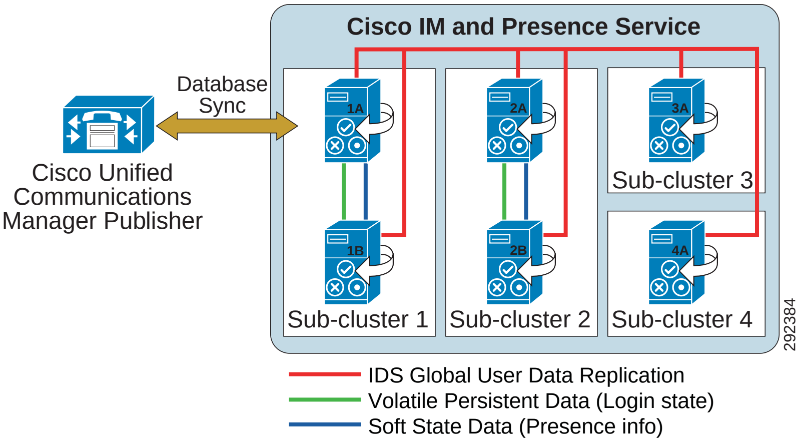

Cisco IM and Presence Service
Database Sync
Cisco Unified Communications Manager Publisher
1A
1B
2A
2B
3A
3B
Sub-cluster 1
Sub-cluster 2
Sub-cluster 3
292383
IDS Global User Data Replication
Volatile Persistent Data (Login state)
Soft State Data (Presence info)

Figure 23-8 Mixed Deployment of Cisco IM and Presence


Cisco IM and Presence Service
Database Sync
Cisco Unified Communications Manager Publisher
1A
1B
2A
2B
3A
4A
Sub-cluster 1
Sub-cluster 2
Sub-cluster 3
Sub-cluster 4
292384
IDS Global User Data Replication
Volatile Persistent Data (Login state)
Soft State Data (Presence info)

The Cisco IM and Presence Service utilizes and builds upon the database used by the Unified CM publisher by sharing the user and device information. A Cisco IM and Presence cluster supports only a single Unified CM cluster; therefore, a separate IM and Presence cluster is required for each Unified CM cluster.

Intracluster traffic participates at a very low level between Cisco IM and Presence and Unified CM and between the Cisco IM and Presence publisher and subscriber servers. Both clusters share a common hosts file and have a strong trust relationship using IPTables. At the level of the database and services, the clusters are separate and distinct, and each Cisco IM and Presence Service and Unified CM cluster requires separate administration. There is currently no Transport Layer Security (TLS) or IPSec utilization for intracluster traffic.

The Cisco IM and Presence Service interface with external systems sends SIP and XMPP traffic over UDP, TCP, or TLS. TLS mutual authentication requires the import and export of certificates between Cisco IM and Presence Service and the external system. TLS server authentication (Cisco IM and Presence Service presenting its TLS certificate to the client device for verification) validates the end user via digest authentication.

The Cisco IM and Presence publisher communicates directly with the Unified CM publisher via the AVVID XML Layer Application Program Interface (AXL API) using the Simple Object Access Protocol (SOAP) interface. When first configured, the Cisco IM and Presence publisher performs an initial synchronization of the entire Unified CM user and device database. All Cisco IM and Presence users are configured in the Unified CM End User configuration. During the synchronization, Cisco IM and Presence populates these users in its database from the Unified CM database and does not provide end-user configuration from its administration interface.

The initial Cisco IM and Presence database synchronization from Unified CM might take a while, depending on the amount of information in the database as well as the load that is currently on the system. Subsequent database synchronizations from Unified CM to Cisco IM and Presence are performed in real time when any new user or device information is added to Unified CM. For planning purposes, use the values in Table 23-2 as guidelines when executing the initial database synchronization with Unified CM using a single Cisco IM and Presence publisher.

Note Cisco IM and Presence supports synchronization of up to 160,000 users, equivalent to Unified CM. However, the maximum number of licensed presence users for a Cisco IM and Presence cluster is 45,000 in full Unified Communications mode and 75,000 in IM-only mode.

Table 23-2 Synchronization Times for a Cisco IM and Presence Publisher

Server Platform	Number of Users	Synchronization Time
Cisco MCS 7816 or OVA equivalent	500	5 minutes
Cisco MCS 7825 or OVA equivalent	1,000	5 minutes
Cisco MCS 7835 or OVA equivalent	1,000	5 minutes
	10,000	25 minutes
Cisco MCS 7845 or OVA equivalent	1,000	5 minutes
	10,000	20 minutes
	30,000	70 minutes

Note The numbers for the Cisco Unified Computing System (UCS) Open Virtualization Archive (OVA) platforms are equivalent to those for the MCS 7835 (2 vCPU, 4 GB RAM, 80 GB drive, 1 vNIC) and MCS 7845 (4 vCPU, 4 GB RAM, two 80 GB drives, 1 vNIC).

For planning purposes, use the values in Table 23-3 as guidelines when executing the initial database synchronization with Unified CM using a Cisco IM and Presence publisher and subscriber servers:

Table 23-3 Synchronization Times for a Cisco IM and Presence Publisher and Subscriber Servers

Server Platform	Number of Users	Synchronization Time
Cisco MCS 7816 or OVA equivalent	500	5 minutes
Cisco MCS 7825 or OVA equivalent	1,000	10 minutes
Cisco MCS 7835 or OVA equivalent	1,000	10 minutes
	10,000	50 minutes
Cisco MCS 7845 or OVA equivalent	1,000	10 minutes
	10,000	40 minutes
	45,000	140 minutes

Note When the Cisco IM and Presence Service is performing the initial database synchronization from Unified CM, do not perform any administrative activities on Unified CM while the synchronization agent is active.

If the database entries are not updating or if the Sync Agent service is stopped, you can check for broken connections with the synchronization agent by using the Real-Time Monitoring Tool (RTMT) to monitor the Critical Alarm **Cisco Unified Presence ServerSyncAgentAXLConnectionFailed**.

Cisco IM and Presence Service High Availability

The Cisco IM and Presence cluster consists of up to six servers, which can be configured into multiple subclusters, with a maximum of three subclusters for high availability. A subcluster contains a maximum of two servers and allows for users associated with one server of the subcluster to use the other server in the subcluster automatically if a failover event occurs. Cisco IM and Presence does not provide failover between subclusters.

When deploying a Cisco IM and Presence cluster for high availability, you must take into consideration the maximum number of users per server to avoid oversubscribing any one server within the subcluster in the event of a failover. When deploying a Cisco IM and Presence cluster, use equivalent hardware for all servers within the cluster.

Cisco IM and Presence Deployment Models

Unified CM provides a choice of the following deployment models:

- Single site
- Multisite WAN with centralized call processing
- Multisite WAN with distributed call processing
- Clustering over the WAN

Cisco IM and Presence is supported with all the Unified CM deployment models. However, Cisco recommends co-locating the Cisco IM and Presence publisher with the Unified CM publisher due to the initial user database synchronization. All Cisco IM and Presence Services should be co-located within the Cisco IM and Presence cluster, with the exception of geographic datacenter redundancy and clustering over the WAN (for details, see Clustering Over the WAN, page 23-21).

For more information on Unified CM deployment models, see the chapter on Unified Communications Deployment Models, page 5-1.

Cisco IM and Presence deployment depends on high-availability requirements, the total number of users, and the server hardware being used. Cisco recommends using similar hardware for each server in the Cisco IM and Presence cluster. Detailed configuration and deployment steps can be found in the *Deployment Guide for Cisco IM and Presence*, available at

> http://www.cisco.com/en/US/products/ps6837/products_installation_and_configuration_guides_list.html

A highly available Cisco IM and Presence cluster requires two servers per subcluster. This allows for users to fail-over between the servers within the subcluster; however, the total number of users supported and the time to failover vary based on which features are enabled, the average size of contact lists, and the rate of traffic placed on the servers. Once a Cisco IM and Presence subcluster is configured for two servers, it always operates as highly available. High availability can be deployed using an Active/Standby model or an Active/Active model, and these modes are controlled by the Sync Agent service parameter User Assignment Mode. By default all users are balanced across all servers in the cluster, and Cisco recommends leaving this parameter set to its default value.

Cisco IM and Presence Active/Standby mode (setting User Assignment Mode to **None**) is attained by manually assigning users to the first server in the subcluster, leaving the second server with no users assigned but all processes synchronized and ready for a failover if the first server in the subcluster fails. For example, in Figure 23-7 the first user would be assigned to server 1A, the second user to server 2A, the third user to server 3A, the fourth user to server 1A, the fifth user to server 2A, the sixth user to server 3A, and so forth. The users should be assigned equally across all the 'A' servers in the cluster.

Cisco IM and Presence Active/Active mode (setting User Assignment Mode to **balanced**) will automatically assign users equally across all servers in the subclusters. Each server is synchronized and ready for a failover if the other server in the subcluster fails. For example, in Figure 23-7 the first user would be assigned to server 1A, the second user to server 2A, the third user to server 3A, the fourth user to server 1B, the fifth user to server 2B, the sixth user to server 3B, and so forth. The users are assigned equally across all the servers in the cluster.

Cisco IM and Presence Active/Active deployments with a balanced User Assignment Mode allows for redundancy flexibility based on the features being used, the size of user contact lists, and the traffic (user data profiles) being generated. A Cisco IM and Presence Active/Active deployment with a fully redundant mode, regardless of features, requires the total number of supported users to be reduced in half (for example, Cisco MCS 7845 servers in a balanced high-availability redundant deployment support up to 15,000 users per subcluster). A Cisco IM and Presence Active/Active deployment with a non-redundant mode requires a more detailed look at the Cisco IM and Presence features being utilized, the average size of the users contact lists, as well as the traffic being generated. For example, for a deployment with presence and instant messaging enabled and calendaring and mobility integration disabled, with an average contact list of 30 users and a user data profile of a few presence and instant message updates, it is possible to support more than 15,000 users per subcluster (if using Cisco MCS 7845 servers).

A Cisco IM and Presence cluster deployment that is not highly available allows for each server in the subcluster to support up to the maximum number of users. Once a second server is added in a subcluster, the subcluster will still act as if in a high-available deployment; however, if a server failure occurs, an attempt to fail-over might not result in success if the online server reaches its capacity limit based on the Cisco IM and Presence features enabled, the average user contact list size, and the traffic being generated by the users.

Cisco IM and Presence Deployment Examples

Example 23-1 Single Unified CM Cluster with Cisco IM and Presence

Deployment requirements:

- 4,000 users that could scale to 13,000 users
- Single Cisco Unified Communications Manager cluster
- Instant message logging and compliance are not needed
- High availability is not needed

Hardware:

- Cisco MCS 7845 servers

Deployment:

- One single-server subcluster using User Assignment Mode = balanced

Example 23-2 Two Unified CM Clusters with Cisco IM and Presence

Deployment requirements:

- 11,000 users that could scale to 24,000 users
- Two Cisco Unified Communications Manager clusters
- Instant message logging and compliance are not needed
- High availability is not needed

Hardware:

- Cisco MCS 7845 servers

Deployment:

- Two Cisco IM and Presence clusters (one per Cisco Unified Communications Manager cluster), each with one server using User Assignment Mode = balanced

Example 23-3 Single Unified CM Cluster with Cisco IM and Presence

Deployment requirements:

- 500 users that could scale to 2500 users
- Single Cisco Unified Communications Manager cluster
- Instant message archiving is required
- High availability is required

Hardware:

- Cisco MCS 7835 servers

Deployment:

- One two-server subcluster using User Assignment Mode set to **balanced**, with a PostgreSQL database instance for the cluster

Example 23-4 Single Cisco Business Edition Cluster with Cisco IM and Presence

Deployment requirements:

- 100 users that could scale to 500 users
- Single Cisco Business Edition
- Instant message archiving and persistent chat are required
- High availability is required

Hardware:

- Cisco MCS 7825 servers

Deployment:

- One two-server subcluster using User Assignment Mode set to **balanced**, with a unique PostgreSQL database instance per server in the cluster for persistent chat functionality

Example 23-5 Multiple Unified CM Clusters with Cisco IM and Presence

Deployment requirements:

- 5,000 users that could scale to 40,000 users
- Multiple Cisco Unified Communications Manager clusters
- Instant message compliance is required
- High availability is required

Hardware:

- Cisco MCS 7845 servers

Deployment:

- Multiple Cisco IM and Presence clusters must be set up with intercluster peers between each. Start with a single two-server subcluster, with up to 15,000 users for each Cisco IM and Presence cluster, before adding more subclusters within existing Cisco IM and Presence clusters. With a large number of users within a single Cisco IM and Presence cluster, set the User Assignment Mode service parameter to **balanced**. Set up a third-party compliance server for instant messaging compliance for each server in each Cisco IM and Presence cluster.

Cisco IM and Presence Deployment for Instant Messaging Only

A Cisco IM and Presence cluster (or clusters) can be deployed to provide enterprise-class presence and instant messaging in an environment where Unified CM is not deployed for call control for specific users. A deployment of IM and Presence only is also referred to as Jabber for Everyone. Unified CM is still required to establish user accounts entered either manually or through LDAP synchronization. A Cisco IM and Presence instant messaging only deployment synchronizes user information from Unified CM in the same way as is done with a full Unified Communications deployment. If Unified CM is not deployed or if the existing deployed Unified CM will not be used for instant messaging only, a Cisco MCS 7816 Media Convergence Server with preloaded Unified CM software is provided as an option.

For existing Cisco IM and Presence deployments where a Unified CM cluster is already deployed, users can also be added for use with the instant messaging only mode. This allows for a mix of full Unified Communications users in addition to instant messaging only users, in accordance with the end-user license agreement.

Cisco IM and Presence Service Performance

Cisco IM and Presence Service clusters support single-server as well as multi-server configurations. However, if multiple servers are used, each server must be on the same type of server platform as the publisher server.

Table 23-4 lists the hardware platform requirements for the Cisco IM and Presence Service as well as the maximum number of users supported per platform. The maximum number of users supported for a Cisco IM and Presence cluster is based on the hardware being used in the deployment. For example, if a Cisco IM and Presence cluster is deployed with three Cisco MCS 7825 servers, each forming their own subcluster, a total of 6,000 users would be supported. The maximum number supported for a Cisco IM and Presence cluster is 45,000 users.

Table 23-4 Cisco IM and Presence Service Platforms and Number of Users Supported

Server Platform	Users Supported Per Platform in Full Unified Communications Mode	Users Supported Per Platform in Instant Messaging Only Mode
Cisco MCS 7816	3,000	7,500
Cisco MCS 7825	6,000	15,000
Cisco MCS 7835 or OVA equivalent	15,000	37,500
Cisco MCS 7845 or OVA equivalent	45,000	75,000

For additional hardware specifications, refer to the Media Convergence Server documentation available at

http://www.cisco.com/en/US/products/hw/voiceapp/ps378/prod_models_home.html

Cisco IM and Presence Deployment

Cisco IM and Presence can be deployed in any of the following configurations:

- Single-Cluster Deployment, page 23-17
- Multi-Cluster Deployment, page 23-20
- Clustering Over the WAN, page 23-21
- Federated Deployment, page 23-22
- Instant Messaging Only Deployment, page 23-26

Single-Cluster Deployment

Figure 23-9 represents the communication protocols between Cisco IM and Presence, the LDAP server, and Cisco Unified Communications Manager for basic functionality. For complete information on Cisco IM and Presence administration and configuration, refer to the Cisco IM and Presence installation, administration, and configuration guides, available at

http://www.cisco.com/en/US/products/ps6837/tsd_products_support_series_home.html

Figure 23-9 Interactions Between Cisco IM and Presence Components

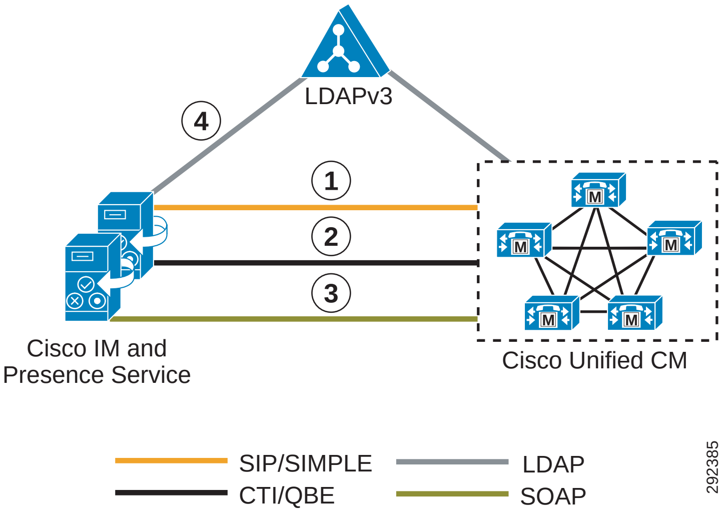


Figure 23-9 depicts the following interactions between Cisco IM and Presence components:

1. The SIP connection between the Cisco IM and Presence Service and Unified CM handles all the phone state presence information exchange.
   a. Unified CM configuration requires the Cisco IM and Presence Services to be added as application servers on Unified CM and also requires a SIP trunk pointing to the Cisco IM and Presence Service. The address configured on the SIP trunk could be a Domain Name System (DNS) server (SRV) fully qualified domain name (FQDN) that resolves to the Cisco IM and Presence Services, or it could simply be an IP address of an individual Cisco IM and Presence Service. The Cisco IM and Presence Service handles the configuration of the Cisco Unified Communications Manager application server entry automatically through AXL/SOAP once the administrator adds a node in the system topology page through Cisco IM and Presence administration.
   b. Configuration of Cisco IM and Presence occurs through the Unified CM Presence Gateway for presence information exchange with Unified CM. The following information is configured:

      Presence Gateway: *server_fqdn*:5070

      Note The *server_fqdn* could be the FQDN of the Unified CM publisher, a DNS SRV FQDN that resolves to the Unified CM subscriber servers, or an IP address.

   If DNS is highly available within your network and DNS SRV is an option, configure the SIP trunk on Unified CM with a DNS SRV FQDN of the Cisco IM and Presence publisher and subscriber. Also configure the Presence Gateway on the Cisco IM and Presence Service with a DNS SRV FQDN of the Unified CM subscribers, equally weighted. This configuration will allow for presence messaging to be shared equally among all the servers used for presence information exchange.

   If DNS is not highly available or not a viable option within your network, use IP addressing. When using an IP address, presence messaging traffic cannot be equally shared across multiple Unified CM subscribers because it points to a single subscriber.

Unified CM provides the ability to further streamline communications and reduce bandwidth utilization by means of the service parameter CUP PUBLISH Trunk, which allows for the PUBLISH method (rather than SUBSCRIBE/NOTIFY) to be configured and used on the SIP trunk interface to Cisco IM and Presence. Once the CUP PUBLISH Trunk service parameter has been enabled, the users must be associated with a line appearance and not just a primary extension.

2. The Computer Telephony Integration Quick Buffer Encoding (CTI-QBE) connection between Cisco IM and Presence and Unified CM is the protocol used by presence-enabled users in Cisco IM and Presence to control their associated phones registered to Unified CM. This CTI communication occurs when Cisco Jabber is using Desk Phone mode to do Click to Call or when Microsoft Office Communicator is doing Click to Call through Microsoft Live Communications Server 2005 or Office Communications Server 2007 or Microsoft Lync.

   a. Unified CM configuration requires the user to be associated with a CTI Enabled Group, and the primary extension assigned to that user must be enabled for CTI control (checkbox on the Directory Number page). The CTI Manager Service must also be activated on each of the Unified CM subscribers used for communication with the Cisco IM and Presence publisher and subscriber. Integration with Microsoft Live Communications Server 2005 or Office Communications Server 2007 or Microsoft Lync requires that you configure an Application User, with CTI Enabled Group and Role, on Unified CM.

   b. Cisco IM and Presence CTI configuration (CTI Server and Profile) for use with Cisco Jabber is automatically created during the database synchronization with Unified CM. All Cisco Jabber CTI communication occurs directly with Unified CM and not through the Cisco IM and Presence Service.

      Cisco IM and Presence CTI configuration (Desktop Control Gateway) for use with Microsoft Live Communications Server 2005 or Office Communications Server 2007 or Microsoft Lync requires you to set the Desktop Control Gateway address (Cisco Unified Communications Manager Address) and a provider, which is the application user configured previously in Unified CM. Up to eight Cisco Unified Communications Manager Addresses can be provisioned for increased scalability. Only IP addresses can be used for Desktop Control Gateway configuration in the Cisco IM and Presence Service. Administrators should ensure that any configuration and assignment of Cisco Unified Communications Manager addresses is evenly distributed for the purpose of load balancing.

3. The AXL/SOAP interface handles the database synchronization from Unified CM to populate the Cisco IM and Presence database.

   a. No additional configuration is required on Unified CM.

   b. Cisco IM and Presence security configuration requires you to set a user and password for the Unified CM AXL account in the AXL configuration.

      The Sync Agent Service Parameter, User Assignment, set to **balanced** by default, will load-balance all users equally across all servers within the Cisco IM and Presence cluster. The administrator can also manually assign users to a particular server in the Cisco IM and Presence cluster by changing the User Assignment service parameter to **None**.

4. The LDAP interface is used for LDAP authentication of Cisco Jabber users during login. For more information regarding LDAP synchronization and authentication, see the chapter on LDAP Directory Integration, page 16-1.

   Unified CM is responsible for all user entries via manual configuration or synchronization directly from LDAP, and Cisco IM and Presence then synchronizes all the user information from Unified CM. If a Cisco Jabber user logs into the Cisco IM and Presence Service and LDAP authentication is enabled on Unified CM, Cisco IM and Presence will go directly to LDAP for the Cisco Jabber user authentication using the Bind operation. Once Cisco Jabber is authenticated, Cisco IM and Presence forwards the information to Cisco Jabber to continue login.

When using Microsoft Active Directory, consider the choice of parameters carefully. Performance of Cisco IM and Presence might be unacceptable when a large Active Directory implementation exists and the configuration uses a Domain Controller. To improve the response time of Active Directory, it might be necessary to promote the Domain Controller to a Global Catalog and configure the LDAP port as 3268.

Multi-Cluster Deployment

The deployment topology in previous sections is for a single Cisco IM and Presence cluster communicating with a single Unified CM cluster. Presence and instant messaging functionality is limited by having communications within a single cluster only. Therefore, to extend presence and instant messaging capability and functionality, these standalone clusters can be configured for peer relationships for communication between clusters within the same domain. This functionality provides the ability for users in one cluster to communicate and subscribe to the presence of users in a different cluster within the same domain.

To create a fully meshed presence topology, each Cisco IM and Presence cluster requires a separate peer relationship for each of the other Cisco IM and Presence clusters within the same domain. The address configured in this intercluster peer could be a DNS SRV FQDN that resolves to the remote Cisco IM and Presence cluster servers, or it could also simply be the IP address of the Cisco IM and Presence cluster servers.

The interface between each Cisco IM and Presence cluster is two-fold, an AXL/SOAP interface and a signaling protocol interface (SIP or XMPP). The AXL/SOAP interface handles the synchronization of user information for home cluster association, but it is not a full user synchronization. The signaling protocol interface (SIP or XMPP) handles the subscription and notification traffic, and it rewrites the host portion of the URI before forwarding if the user is detected to be on a remote Cisco IM and Presence cluster within the same domain.

When Cisco IM and Presence is deployed in a multi-cluster environment, a presence user profile should be determined. The presence user profile helps determine the scale and performance of a multi-cluster presence deployment and the number of users that can be supported. The presence user profile helps establish the number of contacts (or buddies) a typical user has, as well as whether those contacts are mostly local cluster users or users of remote clusters.

The traffic generated between Cisco IM and Presence clusters is directly proportional to the characteristics of the presence user profile. For example, assume presence user profile A has 30 contacts with 20% of the users on a local Cisco IM and Presence cluster and 80% of the users on a remote Cisco IM and Presence cluster, while presence user profile B has 30 contacts with 50% of the users on a local Cisco IM and Presence cluster and 50% of the users on a remote Cisco IM and Presence cluster. In this case, presence user profile B will provide for slightly better network performance and less bandwidth utilization due to a smaller amount of remote cluster traffic.

Clustering Over the WAN

A Cisco IM and Presence cluster can be deployed with one of the nodes of a subcluster deployed across the Wide Area Network (WAN). This allows for geographic redundancy of a subcluster and high availability for the users between the nodes across the sites. The following guidelines must be used when planning for a Cisco IM and Presence deployment with clustering over the WAN.

- Geographic datacenter redundancy and remote failover

  A Cisco IM and Presence cluster can be deployed between two sites with a single subcluster topology, where one server of the subcluster is in one geographic site and the other server of the subcluster is in another site. This deployment must have a minimum bandwidth of 5 Mbps, a maximum latency of 80 ms round-trip time (RTT), and TCP method event routing.

- High availability and scale

  Cisco IM and Presence high availability allows for users on one node within a subcluster to automatically fail-over to the other node within the subcluster. With a Cisco IM and Presence subcluster containing a maximum of two nodes, remote failover is essentially between two sites, one site for each node. A scalable highly available capacity for a Cisco IM and Presence cluster is up to three subclusters; therefore, a scalable highly available remote failover topology would consist of the following two sites:

  – Site A: Subcluster 1 node A, subcluster 2 node A, and subcluster 3 node A

  – Site B: Subcluster 1 node B, subcluster 2 node B, and subcluster 3 node B

  This deployment must have a minimum bandwidth of 5 Mbps per subcluster, a maximum latency of 80 ms round-trip time (RTT), and TCP method event routing. Each new subcluster added to the deployment requires an additional 5 Mbps of dedicated bandwidth to handle the database and state replication.

- Local Failover

  A Cisco IM and Presence cluster deployment between two sites may also contain a subcluster topology per site (single node or dual node for high availability), where one subcluster is in one geographic site and the other subcluster is in another geographic site. This topology allows for the users to remain at their local site (highly available or not) without the requirement or need to fail-over to a different site or location. This deployment must have a minimum bandwidth of 5 Mbps dedicated bandwidth between each subcluster in the respective sites, a maximum latency of 80 ms round-trip time (RTT), and TCP method event routing.

- Bandwidth and latency considerations

  With a Cisco IM and Presence cluster that has a topology of nodes split across a WAN, the number of contacts within a user's client can impact the bandwidth needs and criteria for the deployment. The traffic generated within and between Cisco IM and Presence clusters is directly proportional to the characteristics of the presence user profile, and thus the amount of bandwidth required for deployment. Cisco recommends 25% or fewer remote contacts for a client in environments where the bandwidth is low (10 Mbps or less), and at all times the maximum round-trip latency must be 80 ms or less.

- Persistent Chat and Compliance logging considerations

  When Cisco IM and Presence is enabled for persistent chat, message archiving, or compliance logging and a sublcuster is split across a WAN, the external database server(s) must reside on the same side of the WAN as the Cisco IM and Presence Services that use them. With the ability to support multiple database instances on a single server and the requirement for an external database server to reside on the same side of the WAN, if a Cisco IM and Presence cluster is split across a WAN, then two external database servers will be required.

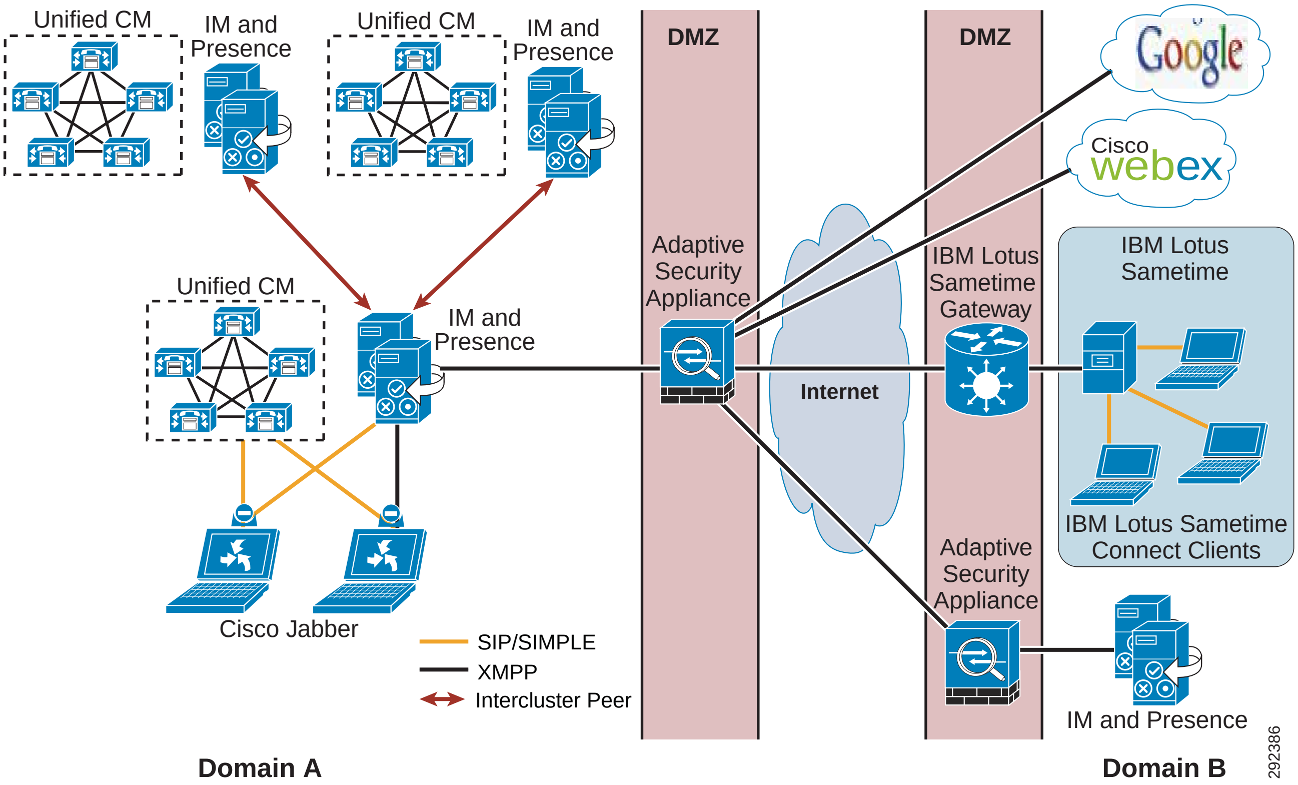
Federated Deployment

Cisco IM and Presence allows for business-to-business communications by enabling inter-domain federation, which provides the ability to share presence and instant messaging communications between different domains. Inter-domain federation requires two explicit DNS domains to be configured, as well as a security appliance (Cisco Adaptive Security Appliance) in the DMZ to terminate federated connections with the enterprise. If all the federated domains are within the same trust boundary, where a deployment has all components within a single datacenter, then the use of the Adaptive Security Appliance is not required. For information on inter-domain federation, refer to the *Integration Guide for Configuring Cisco IM and Presence Interdomain Federation*, available at

http://www.cisco.com/en/US/products/ps6837/products_installation_and_configuration_guides_list.html

Figure 23-10 shows the basic inter-domain federation deployment between two different domains, indicated by Domain A and Domain B. The Adaptive Security Appliance (ASA) in the DMZ is used as a demarcation into the enterprise. XMPP traffic is passed through, whereas SIP traffic is inspected. All federated incoming and outgoing traffic is routed through the Cisco IM and Presence Service that is enabled as a federation node, and is routed internally to the appropriate server in the cluster where the user resides. For multi-cluster deployments, intercluster peers propagate the traffic to the appropriate home cluster within the domain. Multiple nodes can be enabled as federation nodes within large enterprise deployments, where each request is routed based on a round-robin implementation of the data returned from the DNS SRV lookup.

Figure 23-10 Cisco IM and Presence XMPP Federation (Inter-Domain)



Cisco IM and Presence also provides configuration through SIP to allow for inter-domain federation with Microsoft and AOL, as depicted in Figure 23-11. Cisco IM and Presence inter-domain federation with Microsoft Lync Server, Office Communications Server (OCS), and Live Communications Server (LCS) provides basic presence (available, away, busy, offline) and point-to-point instant messaging. Rich presence capability (On the Phone, In a Meeting, On Vacation, and so forth), as well as advanced instant messaging features, are not supported. Cisco IM and Presence inter-domain federation with AOL allows federation with users of AOL public communities (aim.com, aol.com), with users of domains hosted by AOL, and with users of a far-end enterprise that federates with AOL (that is, AOL is being used as a clearing house).

Note A SIP federation (inter-domain to AOL) on Cisco IM and Presence must be configured for each domain of the AOL network, which can consist of both hosted networks and public communities. Each unique hosted domain must be configured; however, only a single aol.com public community needs to be configured because the AOL network allows a user to be addressed as user@aol.com or user@aim.com

The inter-domain federation configuration also allows for a specific federation between Cisco IM and Presence and Microsoft Lync Server or Microsoft Office Communications Server (OCS), as depicted in Figure 23-11. Cisco IM and Presence provides inter-domain federation with Microsoft Lync Server, Microsoft Office Communications Server (OCS), or Live Communications Server (LCS) to provide basic presence (available, away, busy, offline) and point-to-point instant messaging. Rich presence capability (On the Phone, In a Meeting, On Vacation, and so forth), as well as advanced instant messaging features, are not supported.

Figure 23-11 Cisco IM and Presence SIP Federation (Inter-Domain)

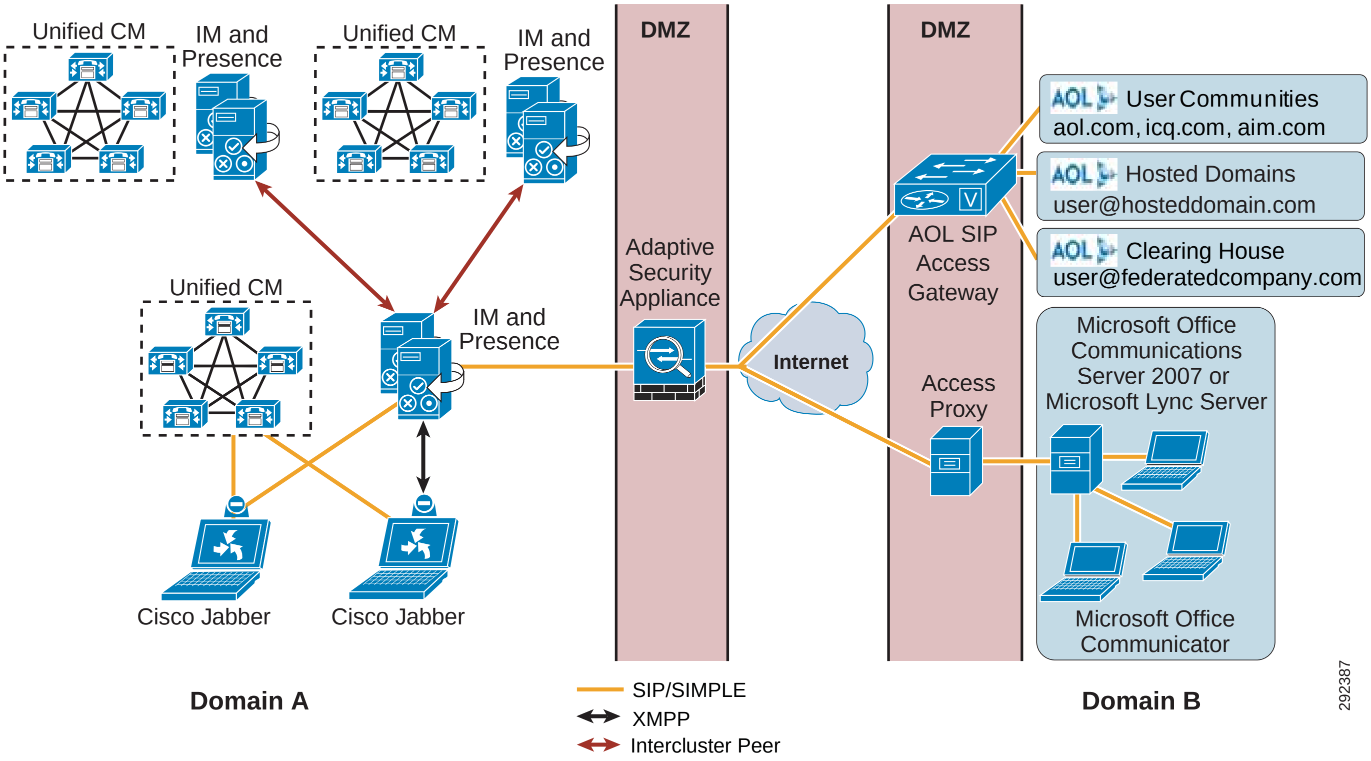


Table 23-5 lists the state mappings between Cisco IM and Presence and Microsoft Office Communications Server.

Table 23-5 Mapping of Presence States

Cisco Status	Cisco Color	Status to Microsoft Office Communications Server	Status to AOL
Out of office	RED	Away	Away
Do not disturb	RED	Busy	Away
Busy	RED	Busy	Away
On the phone	YELLOW	Busy	Away
In a meeting	YELLOW	Busy	Away
Idle on all clients	YELLOW	Away	Away
Available	GREEN	Available	Available
Unavailable/offline	GREY	Offline	Offline

Note Cisco IM and Presence must publish a DNS SRV record (SIP, XMPP, and each text conferencing node) for the domain to allow for other domains to discover the Cisco IM and Presence Services through DNS SRV. With a Microsoft Office Communications Server or Live Communications Server deployment, this is required because Cisco IM and Presence is configured as a Public IM Provider on the Access Edge server. If the Cisco IM and Presence Service cannot discover the Microsoft domain using DNS SRV, you must configure a static route on Cisco IM and Presence for the external domain.

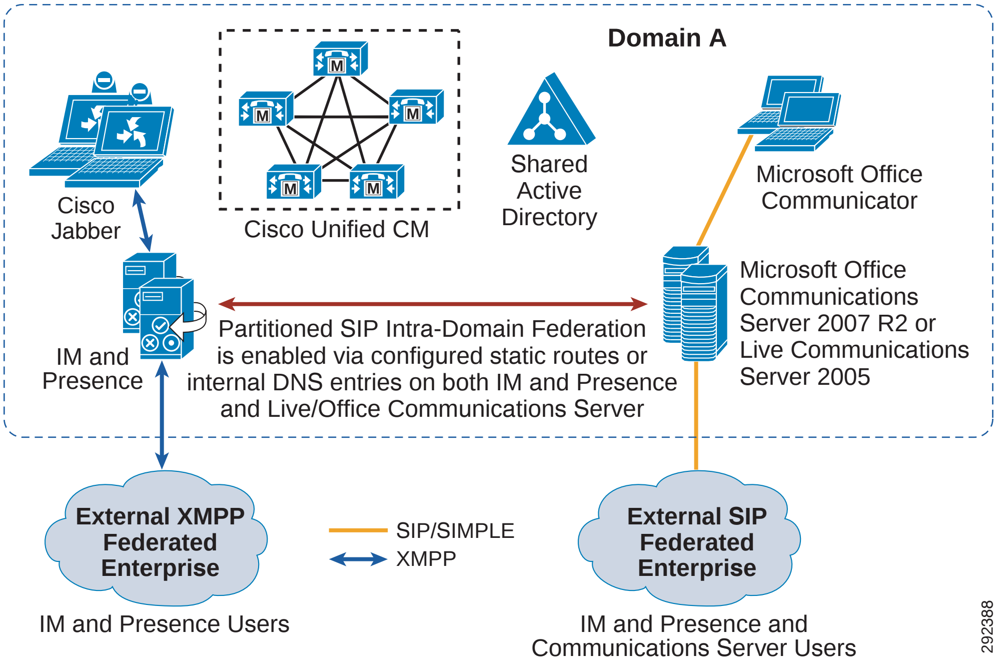
The Cisco IM and Presence federation deployment can be configured with redundancy using a load balancer between the Adaptive Security Appliance and the Cisco IM and Presence Service, or redundancy can also be achieved with a redundant Adaptive Security Appliance configuration.

Additional configuration and deployment considerations regarding a federated deployment can be found in the latest version of the *Integration Guide for Configuring Cisco IM and Presence for Interdomain Federation*, available at

http://www.cisco.com/en/US/products/ps6837/products_installation_and_configuration_guides_list.html

An intra-domain partitioned federated deployment, shown in Figure 23-12, is a secondary option that allows for Cisco IM and Presence and Microsoft Live Communications Server 2005 or Office Communications Server 2007 R2 to federate presence and instant messaging within the same presence domain. The users are partitioned across both deployments, within the single presence domain, and are licensed either on Cisco IM and Presence or on the Microsoft Live Communications Server or Office Communications Server. The user cannot be licensed on both the Cisco and Microsoft platforms at the same time.

Figure 23-12 Cisco IM and Presence Intra-Domain Federation


Domain A
Cisco Jabber
Cisco Unified CM
Shared Active Directory
Microsoft Office Communicator
Microsoft Office Communications Server 2007 R2 or Live Communications Server 2005
IM and Presence
Partitioned SIP Intra-Domain Federation is enabled via configured static routes or internal DNS entries on both IM and Presence and Live/Office Communications Server
External XMPP Federated Enterprise
SIP/SIMPLE
XMPP
External SIP Federated Enterprise
IM and Presence Users
IM and Presence and Communications Server Users
292388

Note Microsoft Office Communications Server 2007 R1 and Microsoft Lync are not supported in an intra-domain partitioned federated deployment.

The partitioned intra-domain federation between the Cisco and Microsoft platforms is based on the SIP/SIMPLE protocol and allows for basic presence and instant messaging exchange, as supported with the Cisco IM and Presence inter-domain federation support for Microsoft Office Communications Server. Rich presence and group chat functionality are not supported with the partitioned intra-domain presence federation.

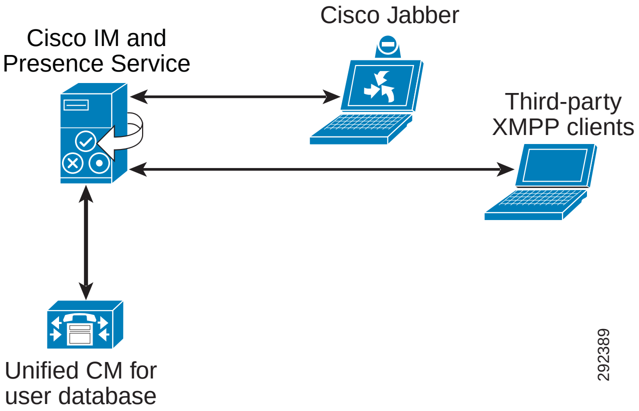
Inter-domain federation and partitioned intra-domain federation can be supported simultaneously with the following qualifications:

- XMPP federation may be enabled on the Cisco IM and Presence deployment but is available only to Cisco IM and Presence licensed users.
- SIP federation may be enabled either on Cisco IM and Presence or on Microsoft Live Communications Server 2005 or Office Communications Server 2007 R2; however, for SIP Federation to be available to both Cisco and Microsoft users, it must be enabled on Microsoft Live Communications Server 2005 or Office Communications Server 2007 R2.
- SIP intercluster trunking with Cisco Unified Presence 7.*x* is not available with partitioned intra-domain federation.
- If SIP/SIMPLE inter-domain federation with Microsoft Office Communications Server is required in parallel with the partitioned intra-domain federation, then the Microsoft Live Communications Server or Office Communications Server will manage that external federation. Cisco IM and Presence administration must be configured with static routes to the Microsoft environment for the external domain.

Instant Messaging Only Deployment

Cisco IM and Presence allows for an enterprise-class instant messaging only solution, which provides full presence and instant messaging support as defined in the section on Cisco IM and Presence Enterprise Instant Messaging, page 23-28, to be deployed in cases where a full Unified Communications deployment is not yet provided. A Cisco IM and Presence cluster deployed in an instant messaging only environment supports up to three servers in a cluster (see Figure 23-13). Instant messaging only users on Cisco IM and Presence are still provisioned from Unified CM through the AXL/SOAP interface by means of LDAP synchronization or manual provisioning. Cisco Jabber and third-party XMPP clients are the supported clients in a Cisco IM and Presence instant messaging only deployment, and all other design guidelines for Cisco IM and Presence apply.

Figure 23-13 Instant Messaging Only User Mode Deployment



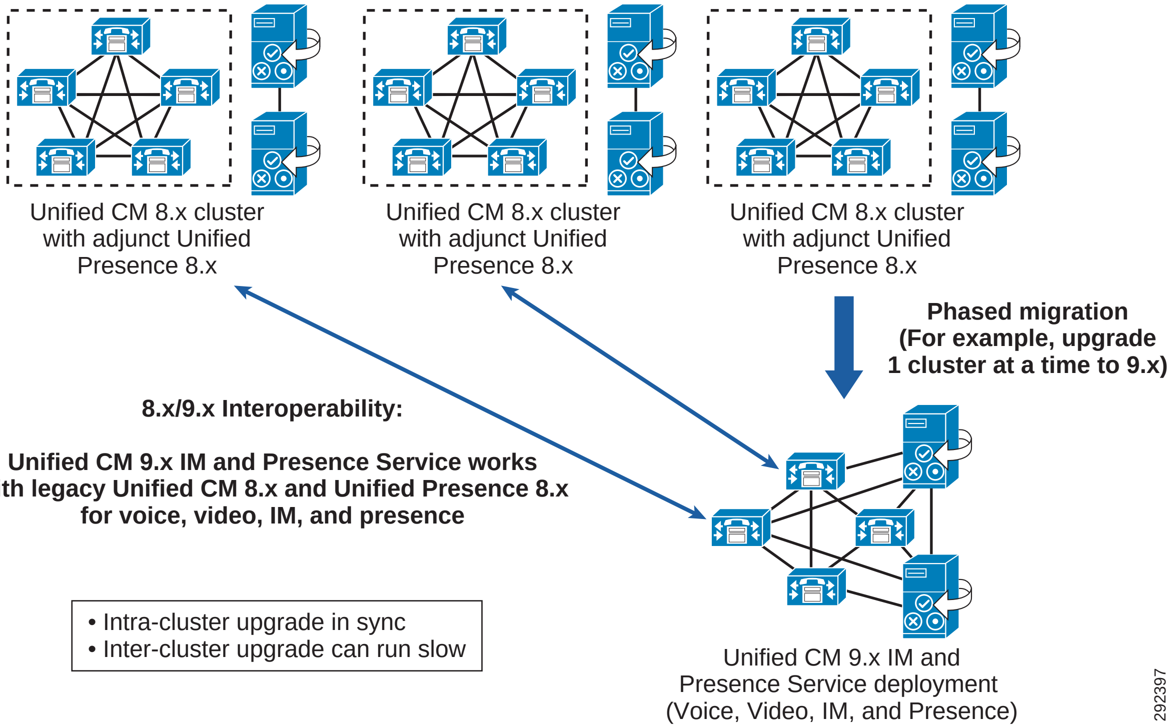
Cisco IM and Presence Migration

Migration from earlier releases of Cisco Unified Presence to Cisco IM and Presence is supported in the following cases (see Figure 23-14):

- Direct migration from Cisco Unified Communications Manager (Unified CM) 7.*x* or 8.*x* to Cisco Unified Communications 9.*x* IM and Presence — provides the ability to deploy Unified CM with voice, video, IM, and presence.
- Direct migration from Cisco Unified Presence 8.*x* IM-only to Cisco Unified Communications 9.*x* IM and Presence — provides the ability to deploy Unified CM with voice, video, IM, and presence.
- Direct migration from Unified CM 7.*x* or 8.*x* and Cisco Unified Presence 8.*x* to Cisco Unified Communications 9.*x* IM and Presence — provides the ability to deploy Unified CM with voice, video, IM, and presence.

Earlier versions of Unified CM and Cisco Unified Presence require a multi-step upgrade migration.

Figure 23-14 Large Enterprise Migration with Backward Compatibility


Unified CM 8.x cluster with adjunct Unified Presence 8.x
Unified CM 8.x cluster with adjunct Unified Presence 8.x
Unified CM 8.x cluster with adjunct Unified Presence 8.x
Phased migration (For example, upgrade 1 cluster at a time to 9.x)
8.x/9.x Interoperability:
Unified CM 9.x IM and Presence Service works with legacy Unified CM 8.x and Unified Presence 8.x for voice, video, IM, and presence
• Intra-cluster upgrade in sync
• Inter-cluster upgrade can run slow
Unified CM 9.x IM and Presence Service deployment (Voice, Video, IM, and Presence)
292397

Cisco IM and Presence Service Policy

Cisco IM and Presence Service policy is set by the user and not the administrator. A default set of rules, with everything open and available, applies if the user does not make any modifications to the policy rules. All policy configuration control is provided in the User Options area of the Cisco IM and Presence user pages at https://<*cup_server_address*>/cupuser/.

The user can configure rule sets that contain access control lists (ACLs) of watchers for which these rules apply. There are three types of rules in each rule set:

- Visibility Rules
  - Reachability Only — Watchers see only the overall reachability of the user, with no device detail information.
  - All State (default) — Watchers see all unfiltered device state information in addition to the overall reachability.
- Reachability Rules
  - Based on precedence rules (first match) for determining reachability (away, available, busy, unavailable, do not disturb, unknown)
  - Based on device type, media type, and calendar (For example, if my cell phone is busy or my calendar is busy, mark me as busy. If any IM device is do-not-disturb, mark me as do-not-disturb.) Do-not-disturb set from the phone will not be propagated to the client device.

- Filtering Rules
  - Exclude presence status for specific device types, media types, or calendar

The filtering rules are applied prior to determining reachability; therefore, a device's filtered status does not affect the reachability status of the user. The user may also define device types (for example, mobile phone, office phone, and so forth) for use with the reachability and filtering rules.

Privacy lists are based on subscription, such that users in your contact list will always be allowed to see your presence. In order to block a user that is in a user's contact list, the blocked user must explicitly be added to the blocked list.

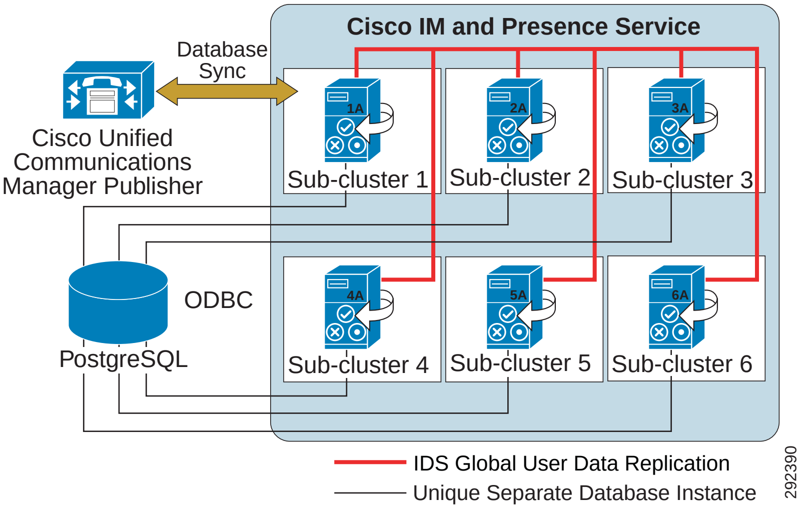
Cisco IM and Presence Enterprise Instant Messaging

Cisco IM and Presence incorporates the supported enterprise instant messaging features of the Jabber Extensible Communications Platform (XCP), while allowing for some modifications to enhance support for multi-device user experience. Cisco IM and Presence changes the Jabber XCP instant messaging routing architecture to allow for initial instant messages to be routed to all of the user's non-negative priority logged-in devices, rather than routing to the highest priority device as is done with existing Jabber XCP installations. Backward compatibility support for point-to-point instant messaging between Cisco IM and Presence SIP clients and XMPP clients is provided by an IM gateway.

Text conferencing, sometimes referred to as multi-user chat, is defined as ad-hoc group chat and persistent group chat and is supported as part of the Jabber XCP feature set. In addition, offline instant messaging (storing instant messages for users who are currently offline) is also supported as part of the Jabber XCP feature set. Cisco IM and Presence handles storage for each of these instant messaging features in different locations. Offline instant messaging is stored locally in the Cisco IM and Presence IDS database. Ad-hoc group chat is stored locally in memory on Cisco IM and Presence. Persistent group chat requires an external database to store chat rooms and conversations. The only external database supported is PostgreSQL (see http://www.postgresql.org/).

Cisco IM and Presence uses the basic interfaces of the external database and does not provide any administration, interface hooks, or configuration of the database. Cisco requires a separate database instance for each server in the cluster when Cisco IM and Presence is deployed with persistent group chat. (See Figure 23-15.) The database instances can share the same hardware but are not required to do so.

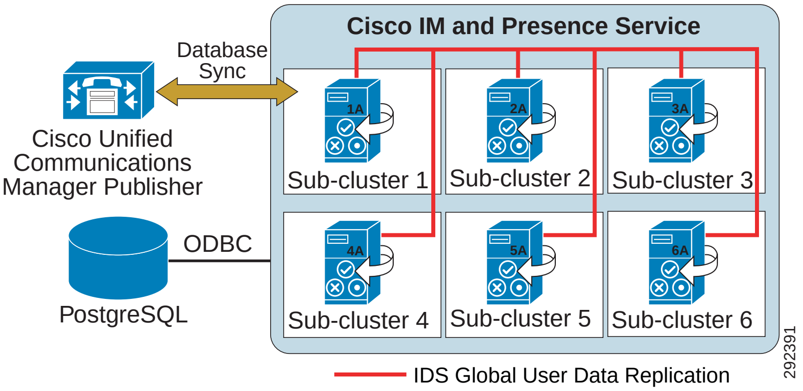
Figure 23-15 Cisco IM and Presence Persistent Chat


Cisco IM and Presence Service
Database Sync
Cisco Unified Communications Manager Publisher
1A
2A
3A
Sub-cluster 1
Sub-cluster 2
Sub-cluster 3
ODBC
PostgreSQL
4A
5A
6A
Sub-cluster 4
Sub-cluster 5
Sub-cluster 6
IDS Global User Data Replication
Unique Separate Database Instance
292390

Cisco IM and Presence Message Archiving and Compliance

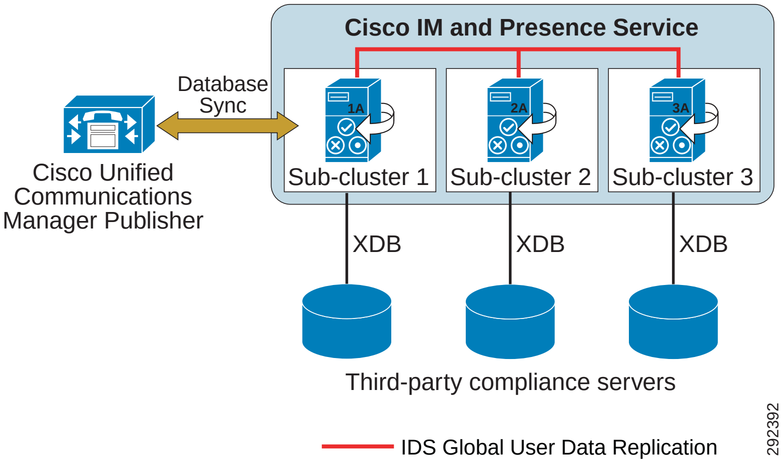
As part of the Jabber XCP architecture, Cisco IM and Presence contains a Message Archiver component that allows for logging of text conferencing, federated, and intercluster messages into an external database as part of a non-blocking native compliance. Cisco IM and Presence native compliance and message archival requires a PostgreSQL database instance per cluster, as shown in Figure 23-16. The same database can be shared with multiple clusters; however, a large number of users in a multi-cluster deployment might require multiple database servers.

Figure 23-16 Cisco IM and Presence Native Compliance and Message Archiving



A blocking third-party compliance solution, which not only allows logging of messages but also applies policy to message delivery and message content, is provided through a third-party compliance server solution. Cisco IM and Presence third-party compliance requires a compliance server for each server in the cluster, as shown in Figure 23-17.

Figure 23-17 Cisco IM and Presence Third-Party Compliance


Cisco IM and Presence Service
Database Sync
Cisco Unified Communications Manager Publisher
1A
2A
3A
Sub-cluster 1
Sub-cluster 2
Sub-cluster 3
XDB
XDB
XDB
Third-party compliance servers
IDS Global User Data Replication
292392

Instant Messaging Storage Requirements

The message archiving and Persistent Chat functionality use an external database to store messages offline. There are a number of factors to consider for the storage requirements of a deployment, such as the customer topology, how the database is tuned, and how messaging is used within the organization. The following calculations provide guidelines for these inputs to be used in estimating the raw database storage requirements of a deployment for external database storage. These calculations presume single-byte character data encoding; therefore, additional storage may be needed if internationalized character sets are used.

Cisco IM and Presence supports both SIP and XMPP clients, and there are slightly different amounts of overhead per message based on the protocol. The overhead per message for message archiving could actually be larger or smaller depending on deployment, Jabber Identifier/UserID size, client type, and thread ID; therefore, an average overhead amount is used. For SIP-based messages the average overhead is 800 bytes and for XMPP messages the average overhead is 600 bytes.

The minimum storage requirements (in bytes) for message archiving per month for Cisco Jabber users can be calculated as follows:

(Number of users) ∗ (Number of messages/hour) ∗ (Number of busy hours/month) ∗ (600 + (3 ∗ Number of characters/message))

The message archiving requirements above must be doubled if **Enable Outbound Message Logging** is enabled on Cisco IM and Presence compliance configuration.

The minimum storage requirements (in bytes) for persistent chat per month for Cisco Jabber users can be calculated as follows:

(Number of users) ∗ (Number of Persistent Chat messages/hour) ∗ (Number of busy hours/month) ∗ (700 + (3 ∗ Number of characters/message))

Note Persistent Chat is supported only with XMPP clients and uses an average overhead of 700 bytes.

These message archive and Persistent Chat numbers are the minimum storage requirements based on an average over time; therefore, a buffer multiplier of 1.5 (150%) should be used to account for very large UserIDs, larger than expected instant message lengths, and other factors that tend to increase the storage requirements. Table 23-6 and Table 23-7 list some examples of storage requirements for Cisco Unified Personal Communicator 8.*x* and 7.*x*, respectively.

Table 23-6 Examples of Unified Personal Communicator 8.x Message Logging Storage Requirements

Profile	Number of Users	Number of Messages per Hours	Number of Busy Hours per Month	Average Size of Message	Message Archive Storage Requirement	Persistent Chat Storage Requirement
Light	1500	10	200	100	2.7 GB	3.0 GB
Medium	2500	15	200	250	10.2 GB	10.9 GB
High	2500	25	200	500	26.3 GB	27.5 GB

Table 23-7 Examples of Unified Personal Communicator 7.x Message Logging Storage Requirements

Profile	Number of Users	Number of Messages per Hour	Number of Busy Hours per Month	Average Size of Message	Message Archive Storage Requirement
Light	1500	10	200	100	3.3 GB
Medium	2500	15	200	250	11.7 GB
High	2500	25	200	500	28.8 GB

Cisco IM and Presence Calendar Integration

Cisco IM and Presence has the ability to retrieve calendar state and aggregate it into a presence status via the calendar module interface with Microsoft Exchange 2003, 2007, or 2010. Microsoft Exchange integration is supported with Microsoft Active Directory 2003 and Active Directory 2008 as well as Windows Server 2003 and Windows Server 2008. Microsoft Exchange 2003 or 2007 makes the calendar data available from the server through Outlook Web Access (OWA), which is built upon extensions to the WebDAV protocol (RFC 2518). Microsoft Exchange 2007 or 2010 makes the calendar data available from the server through Exchange Web Services (EWS), which allows submitting requests and receiving notifications from Microsoft Exchange. The integration with Microsoft Exchange is done through a separate Presence Gateway configuration for calendar applications. Once the administrator configures a Presence Gateway for Outlook, the user has the ability to enable or disable the aggregation of calendar information into their presence status (see Table 23-8).

Note EWS and WebDAV cannot be configured on the same server.

Table 23-8 Aggregated Presence State Based on Calendar State

Cisco IM and Presence State	Calendar State
Available	Free / Tentative
Idle/Busy	Busy
Away	Out of Office

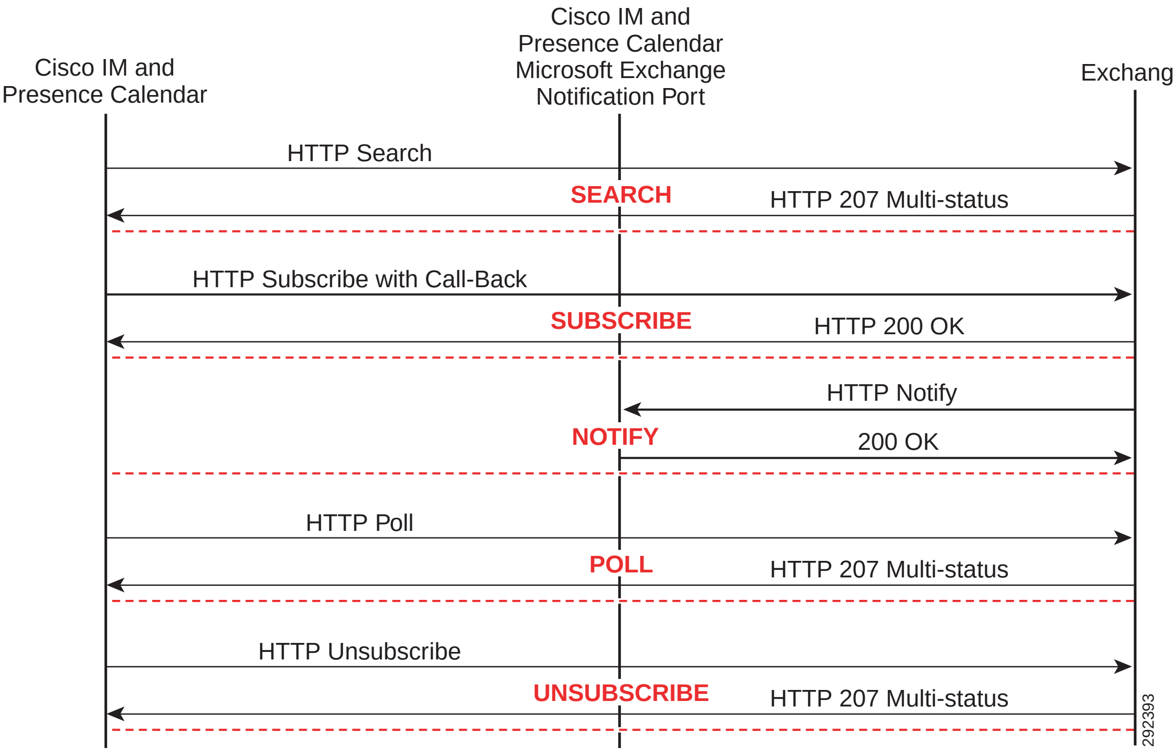
The exchange ID that is used to retrieve calendar information is taken from the email ID of the LDAP structure for that user. If the email ID does not exist or if LDAP is not being used, then the Cisco IM and Presence user ID is mapped as the exchange ID.

Information is gathered via a subscription for calendar state from the Cisco IM and Presence Service to the Microsoft Exchange server. Figure 23-18 depicts this communication.

Outlook Web Access Calendar Integration

This feature requires a service parameter that is the port address for the UDP HTTP (Microsoft Exchange Notification Port) listen port. This port is where Microsoft Exchange sends any notifications (indicated by the NOTIFY message) indicating a change to a particular subscription identifier for calendar events. (See Figure 23-18.)

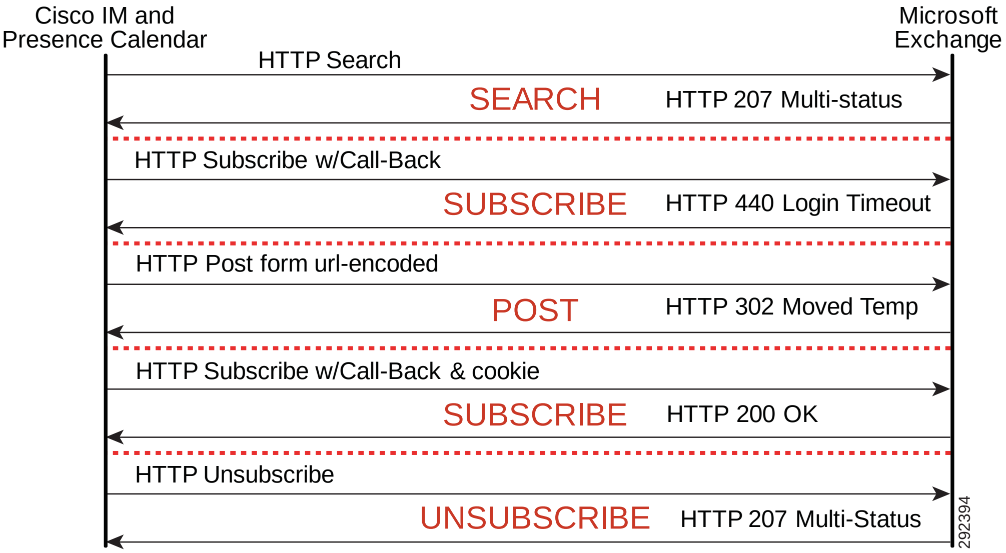
Figure 23-18 Outlook Web Access Communication Between Cisco IM and Presence and Microsoft Exchange



The SEARCH transaction is used to search a user's calendar relevant to a given interval, and is invoked when the user has set a preference to include the calendar information in the presence status. The results of the search are converted into a list of free/busy state transitions. The SUBSCRIBE message indicates the subscription for notifications to changes in the free/busy state of the user in the folder /exchange/user*X*/Calendar. The POLL method is used to acknowledge that the client has either received or responded to a particular event, while the UNSUBSCRIBE message is used to terminate a previous subscription or subscriptions.

Cisco IM and Presence Outlook Web Access integration supports enabling of Forms Based Authentication, which performs an additional HTTP POST transaction request containing the actual URL of the Exchange Server encoded as part of the header, as shown in Figure 23-19.

Figure 23-19 Forms Based Authentication with Cisco IM and Presence Calendar



Note Cisco IM and Presence can be deployed with a single Microsoft Exchange Server or with multiple Microsoft Exchange Servers, in a single forest only. Microsoft Exchange deployment allows for clustering of multiple Exchange servers; therefore, Cisco IM and Presence will honor the REDIRECT message to the exchange server that is hosting the user for which Cisco IM and Presence is requesting status.

Multi-Language Calendar Support

In cases where the requirements for a calendar integration deployment specify more than one language, use the following design guidelines:

- Cisco IM and Presence, as well as Cisco Unified Communications Manager, must have the appropriate locales installed for the users to select their locale.
- Cisco IM and Presence supports all the standard Unified Communications locales for calendar integration.
- Users must be configured for the locale that is desired, either through the end user pages or administratively through the Bulk Administration Tool.
- Cisco IM and Presence sends the appropriate locale folder with the initial query. Queries are redirected, if required, through the response of the initial Front-End or Client Access Microsoft Exchange server.

Exchange Web Services Calendar Integration

Cisco IM and Presence can be configured to allow for Microsoft Exchange Web Services to collect calendar state information to be aggregated into an overall presence view of the user. If the users mailbox is located on the configured Exchange server, Cisco IM and Presence will communicate directly with the Exchange server; whereas, if the users mailbox is located on a different Exchange server than the one configured, Cisco IM and Presence will follow the Exchange server redirection to find the server where the users mailbox is located. Only Exchange Servers from the server farm can serve as the configured Exchange server, and you are required to specify only one of these servers from the server farm.

Microsoft Exchange Web Services specifies the protocol used to transact with the Exchange Client Access Servers independent of the language that the end-user uses; therefore, there is no need to utilize the locale to determine the language of the end-user. Cisco IM and Presence calendar integration is supported with a single Microsoft Exchange forest only.

Cisco IM and Presence Exchange Web Services calendar integration supports both a polling of calendar information as shown in Figure 23-20 as well as a subscription/notification for calendar information as shown in Figure 23-21. Various configuration parameters control the rate of polling intervals, the frequency of subscriptions, and the fault tolerance of timers. For additional configuration details, refer to the *Integration Note for Configuring Cisco IM and Presence with Microsoft Exchange*, available at

http://www.cisco.com/en/US/products/ps6837/products_installation_and_configuration_guides_list.html

Figure 23-20 Exchange Web Services Polling with Cisco IM and Presence Calendar

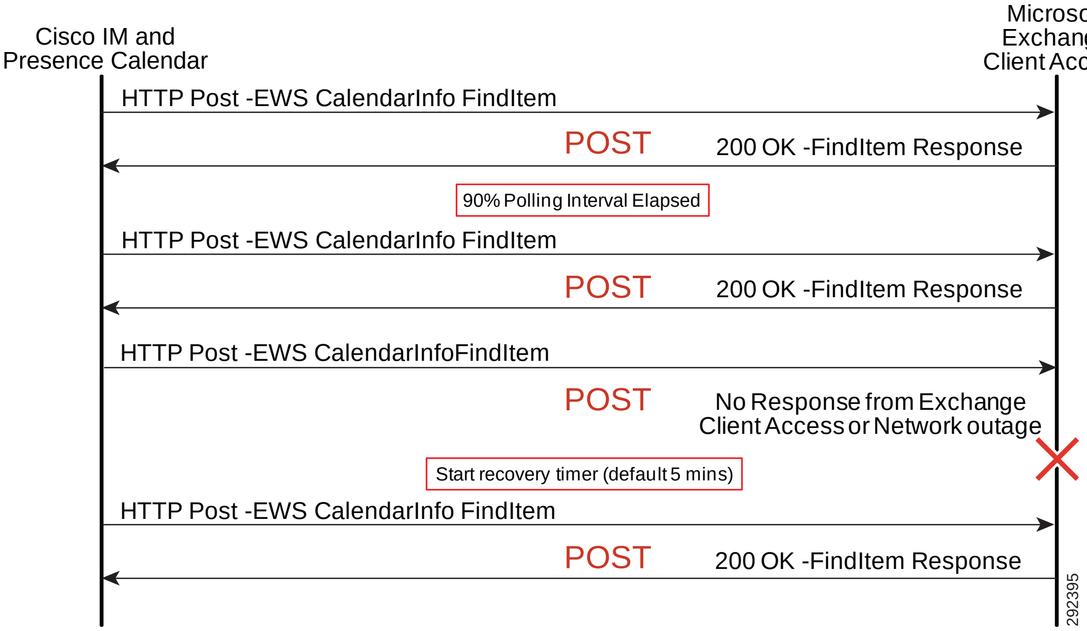
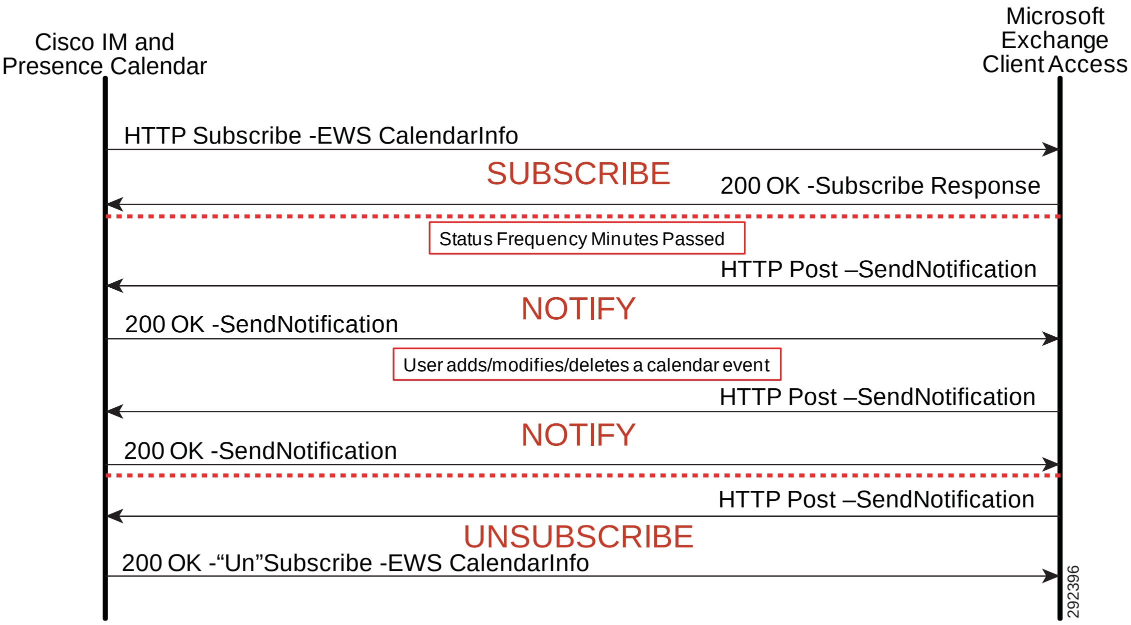

Cisco IM and Presence Calendar
Microsoft Exchange Client Access
HTTP Post -EWS CalendarInfo FindItem
POST
200 OK -FindItem Response
90% Polling Interval Elapsed
HTTP Post -EWS CalendarInfo FindItem
POST
200 OK -FindItem Response
HTTP Post -EWS CalendarInfoFindItem
POST
No Response from Exchange Client Access or Network outage
Start recovery timer (default 5 mins)
HTTP Post -EWS CalendarInfo FindItem
POST
200 OK -FindItem Response
292395

Figure 23-21 Exchange Web Services Subscription/Notification with Cisco IM and Presence Calendar


Cisco IM and Presence Calendar
Microsoft Exchange Client Access
HTTP Subscribe -EWS CalendarInfo
SUBSCRIBE
200 OK -Subscribe Response
Status Frequency Minutes Passed
HTTP Post –SendNotification
NOTIFY
200 OK -SendNotification
User adds/modifies/deletes a calendar event
HTTP Post –SendNotification
NOTIFY
200 OK -SendNotification
HTTP Post –SendNotification
UNSUBSCRIBE
200 OK -“Un”Subscribe -EWS CalendarInfo
292396

Exchange Web Services Auto Discover is also supported by Cisco IM and Presence if a service connection point (SCP) Active Directory object has been created for each server where the Client Access Server (CAS) role is installed. The calendar gateway is configured with Auto Discover using the domain and optionally the site instead of a host and port. Cisco IM and Presence uses the auto-discover algorithm to determine which Exchange Web Services URL to use in contacting the correct Client Access Server Exchange Server.

Cisco IM and Presence Mobility Integration

Cisco IM and Presence has the ability to integrate contact lists and presence state with Cisco Jabber Mobile IM. Jabber Mobile IM continues to communicate directly with Cisco Unified CM, while Cisco Unified CM communicates with Cisco IM and Presence via AXL/SOAP and SIP.

An application user must be configured on Cisco IM and Presence and Cisco Unified CM before Cisco Unified CM can establish an administrative session with Cisco IM and Presence. Cisco Jabber Mobile IM end-user logins will generate a Cisco Unified CM SOAP request to Cisco IM and Presence for system configuration, user configuration, contact list, presence rules, and application dial rules, followed by Unified Communicator Change Notifier (UCCN) configuration and Presence SIP subscriptions.

Cisco IM and Presence Third-Party Open API

Cisco IM and Presence has the ability to integrate with third-party applications through HTTP in addition to SIP/SIMPLE and XMPP. The HTTP interface has a configuration interface as well as a presence interface via Representational State Transfer (REST). The Third-Party Open API provides two mechanisms to access presence: a real-time eventing model and a polling model.

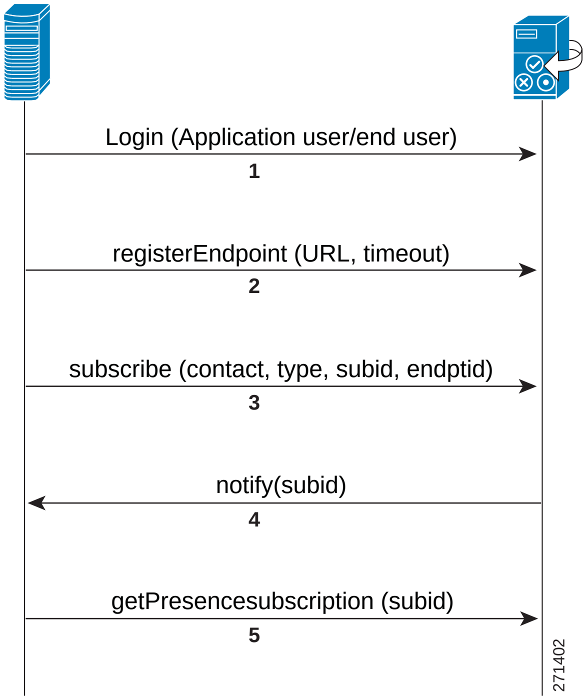
For more information on the Third-Party Open API, refer to the Cisco Developer Community at

http://developer.cisco.com/web/cdc

Real-Time Eventing Model

The real-time eventing model uses an application user on Cisco IM and Presence to establish an administrative session, which allows for end users to log in with that session key. Once the end user has logged in, the user registers and subscribes for presence updates using Representational State Transfer (REST). Figure 23-22 highlights the Third-Party Open API real-time eventing model interaction with Cisco IM and Presence.

Figure 23-22 Third-Party Open API Real-Time Eventing Model


Login (Application user/end user)
1
registerEndpoint (URL, timeout)
2
subscribe (contact, type, subid, endptid)
3
notify(subid)
4
getPresencesubscription (subid)
5
271402

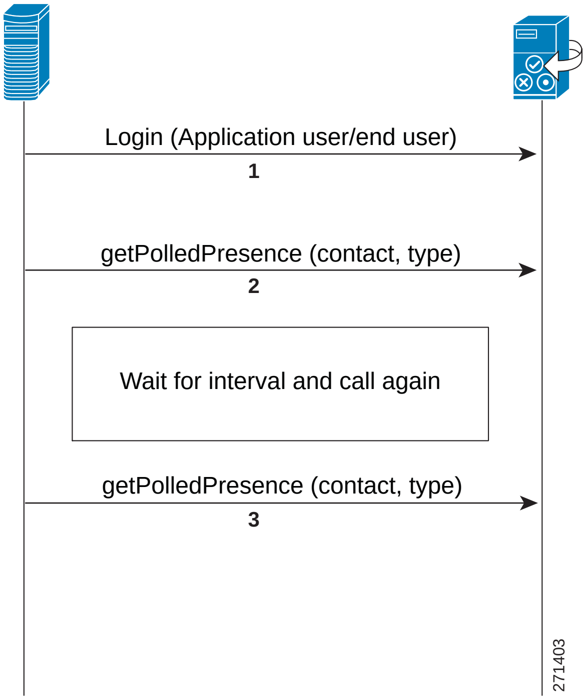
The call flow in Figure 23-22 illustrates the following sequence of events:

1. The application initiates a SOAP login request to Cisco IM and Presence via the super-user application user (APIUser), and Cisco IM and Presence returns a session key. The application can then log in the end-user with this session key (essentially, the end-user logs in via the application).
2. The end user registers the endpoint using the application-user session key.
3. The application initiates a subscribe request (using the session key) on behalf of the end user to retrieve user information, contact list, and presence rules.
4. Cisco IM and Presence sends a notification – unsecured.
5. The application requests the user’s presence status.

Polling Model

The polling model uses an application user on Cisco IM and Presence to establish an administrative session, which allows for end users to log in with that session key. Once the end user has logged in, the application requests presence updates periodically, also using Representational State Transfer (REST). Figure 23-23 highlights the Third-Party Open API polling model interaction with Cisco IM and Presence.

Figure 23-23 Third-Party Open API Polling Model


Login (Application user/end user)
1
getPolledPresence (contact, type)
2
Wait for interval and call again
getPolledPresence (contact, type)
3
271403

The call flow in Figure 23-23 illustrates the following sequence of events:

1. The application initiates a SOAP login request to Cisco IM and Presence via the super-user application user (APIUser), and Cisco IM and Presence returns a session key. The application can then log in the end-user with this session key (essentially, the end-user logs in via the application).
2. The application requests presence state and bypasses the eventing model.
3. The application requests presence state and bypasses the eventing model.

Note Both Basic presence and Rich presence can be retrieved; however, the polling model puts an additional load on the presence server.

Extensible Messaging and Presence Protocol Interfaces

The Jabber XCP architecture allows for two additional open interfaces for presence, instant messaging, and roster management: a client XMPP interface and a Cisco AJAX XMPP Library interface. The client XMPP functionality enables third-party XMPP clients to integrate presence, instant messaging, and roster management, and it is a complementary interface to the SIP/SIMPLE interface on Cisco IM and Presence. The client XMPP interface is treated as a normal XMPP client within Cisco IM and Presence; therefore, sizing of the interface should be treated as a normal XMPP client.

The Cisco AJAX XMPP Library API provides a Web 2.0 style of interface to integrate Jabber XCP features into web applications and widgets, and it is made directly available from Cisco IM and Presence. The Cisco AJAX XMPP Library API is exclusively a client-side JavaScript library that communicates to the Bidirectional-streams Over Synchronous HTTP (BOSH) interface, which is essentially an XMPP over HTTP interface that allows the server to push data to a web browser through a long-polling technique.

Observe the following requirements when integrating either model of the Third-Party Open API with Cisco IM and Presence:

- Certificates are required for the presence interface (sipproxy.der) and the configuration interface (tomcat_cert.der).
- No more than 1000 Third-Party Open API users can be integrated per Cisco IM and Presence deployment.
- To improve performance, balance the Third-Party Open API users across all servers in the Cisco IM and Presence cluster.

You can obtain additional information and support for use of the Cisco IM and Presence Third-Party Open API through Cisco Developer Services, available at:

http://developer.cisco.com/web/cupapi

Information and assistance for developers is also available from the Cisco Developer Community, which is accessible through valid Cisco login authentication at:

http://developer.cisco.com/

Design Considerations for Cisco IM and Presence

- If LDAP integration is possible, LDAP synchronization with Unified CM should be used to pull all user information (number, ID, and so forth) from a single source. However, if the deployment includes both an LDAP server and Unified CM that does not have LDAP synchronization enabled, then the administrator should ensure consistent configuration across Unified CM and LDAP when configuring user directory number associations.
- Cisco IM and Presence marks Layer 3 IP packets via Differentiated Services Code Point (DSCP). Cisco IM and Presence marks all call signaling traffic based on the Differential Service Value service parameter under SIP Proxy, which defaults to a value of DSCP 24 (PHB CS3).
- Presence Policy for Cisco IM and Presence is controlled strictly by a defined set of rules created by the user.
- The Cisco IM and Presence publisher must be co-located with the Unified CM publisher.
- Use the service parameter CUP PUBLISH Trunk to streamline SIP communication traffic with the Cisco IM and Presence Service.
- Associate presence users in Unified CM with a line appearance, rather than just a primary extension, to allow for increased granularity of device and user presence status. When using the service parameter CUP PUBLISH Trunk, you must associate presence users in Unified CM with a line appearance.
- A Presence User Profile (the user activity and contact list contacts and size) must be taken into consideration for determining the server hardware and cluster topology characteristics.
- Use the User Assignment Mode sync agent parameter default of **balanced** for best overall cluster performance.

- Cisco IM and Presence requires an external database instance for each server within in the cluster for persistent chat, and one database instance per cluster for message archiving and native compliance. The database instances can share the same hardware; however, the only external database supported is PostgreSQL.
- Cisco IM and Presence supports a total of 45,000 users per cluster for full Unified Communications mode or 75,000 users for IM-only mode. The sizing for users must take into account the number of SIP/SIMPLE users and the number of XMPP users. XMPP users have slightly better performance because SIP/SIMPLE users employ the IM Gateway functionality into the Jabber XCP architecture.
- When migrating a Cisco IM and Presence deployment from version 7.*x* to 8.*x*, you must deactivate the presence engine service prior to the upgrade and re-enable it after all servers have been upgraded to 8.*x*.
- All eXtensible Communications Platform (XCP) communications and logging are stored in GMT and not localized to the installed location.
- Cisco IM and Presence 9.*x* is compatible with Unified CM 9.*x* only.
- Cisco Business Edition 7.*x* and later releases support LDAP synchronization, which should be enabled when integrating Cisco Business Edition with Cisco IM and Presence.
- For ease of user migration and contact list migration, Cisco IM and Presence Bulk Administration Tool supports bulk contact list importation using a comma-separated value (csv) file as input for this bulk importation.

For a complete listing of ports used by Cisco IM and Presence, refer to *Port Usage Information for Cisco IM and Presence*, available at

http://www.cisco.com/en/US/products/ps6837/products_device_support_tables_list.html

Third-Party Presence Server Integration

Cisco IM and Presence provides an interface based on SIP and SIP for Instant Messaging and Presence Leveraging Extensions (SIMPLE) for integrating SIP and SIMPLE applications into the Cisco Unified Communications solution. You can configure and integrate a third-party presence server or application with this SIP/SIMPLE interface to provide presence aggregation and federation.

Microsoft Communications Server

For all setup, configuration, and deployment of Microsoft Live Communications Server 2005 or Office Communications Server 2007, Microsoft Lync, and Microsoft Office Communicator, refer to the documentation at:

http://www.microsoft.com/

Cisco does not provide configuration, deployment, or best practice procedures for Microsoft Communications products, but Cisco does provide the guidelines listed below for integrating Cisco IM and Presence with Microsoft Live Communications Server 2005 or Office Communications Server 2007 or Microsoft Lync.

Cisco Systems has developed an application note to show feature interoperability and configuration steps for integrating Cisco IM and Presence with Microsoft Live Communications Server 2005. You can access this application note at:

http://www.cisco.com/en/US/docs/voice_ip_comm/cucme/pbx/interop/notes/602270nt.pdf

Cisco Systems has also developed application notes to show feature interoperability and configuration steps for integrating Cisco IM and Presence with Microsoft Office Communications Server 2007 or Microsoft Lync. You can access the application notes at the following locations:

http://www.cisco.com/en/US/docs/voice_ip_comm/cucme/pbx/interop/notes/617030nt.pdf

http://www.cisco.com/en/US/solutions/collateral/ns340/ns414/ns728/ns784/712410.pdf

Cisco Systems has also developed a guide for integrating Cisco IM and Presence with Microsoft Office Communications Server 2007 or Microsoft Lync. This *Integration Note for Configuring Cisco IM and Presence with Microsoft LCS/OCS for MOC Call Control* is available at:

http://www.cisco.com/en/US/products/ps6837/products_installation_and_configuration_guides_list.html

Guidelines for Integrating Cisco IM and Presence with Microsoft Live Communications Server 2005 or Office Communications Server 2007

The following guidelines apply when integrating the Cisco IM and Presence Service and Microsoft Live Communications Server 2005 or Office Communications Server 2007 or Microsoft Lync:

- Communications between Cisco IM and Presence and Microsoft Live Communications Server 2005 or Office Communications Server 2007 or Microsoft Lync uses the SIP/SIMPLE interface. However, Microsoft Live Communications Server 2005 or Office Communications Server 2007 or Microsoft Lync tunnels Computer-Supported Telecommunications Applications (CSTA) traffic over SIP. Therefore, the CTI gateway on the Cisco IM and Presence Service must be configured to handle the CSTA-to-CTI conversion for Click to Call phone control.
- Cisco IM and Presence deployment with Microsoft Office Communications Server 2007 or Live Communications Server 2005 or Microsoft Lync for Remote Call Control, should consist of a single subcluster pair of servers that make up the Cisco IM and Presence cluster.
- The following table lists the number of users supported per platform.

Cisco IM and Presence Platform	Cisco Unified Communications Manager Platform	Number of Microsoft Office Communicator Users Supported per Server[1]	Number of Microsoft Office Communicator Users Supported per Cluster[1]
MCS 7825, 7835, or 7845	MCS 7825-H4/I4 or earlier	900	3,600
MCS 7825, 7835, or 7845	MCS 7825-H5/I5 or OVA equivalent servers	1,000	4,000
MCS 7825, 7835, or 7845	MCS 7835-H2/I2	2,000	8,000
MCS 7825, 7835, or 7845	MCS 7835-H3/I3 or OVA equivalent servers	2,500	10,000
MCS 7825, 7835, or 7845	MCS 7845-H2/I2 or OVA equivalent servers	5,000	20,000
MCS 7825, 7835, or 7845	MCS 7845-I3 or OVA equivalent servers	10,000	40,000

[1] These numbers are based on Cisco Unified CM 7.1(3) and later releases.

- You must configure the same end-user ID in LDAP, Unified CM, and Microsoft Live Communications Server 2005 or Office Communications Server 2007 or Microsoft Lync. This practice avoids any conflicts between Microsoft Live Communications Server 2005 or Office Communications Server 2007 or Microsoft Lync authentication with Active Directory (AD) and the end-user configuration on Unified CM, as well as conflicts with user phone control on Unified CM.

  For Active Directory, Cisco recommends that the user properties of General, Account, and Live Communications all have the same ID. To ensure the Cisco IM and Presence users are consistent, LDAP Synchronization and Authentication should be enabled with Unified CM.
- You must configure Microsoft Live Communications Server 2005 or Office Communications Server 2007 or Microsoft Lync Host Authentication to contain the Cisco IM and Presence publisher and subscriber.
- You can configure routing of the SIP messages to Cisco IM and Presence by means of Static Routes in the Microsoft Live Communications Server 2005 or Office Communications Server 2007 or Microsoft Lync properties.
- You must configure an incoming and outgoing access control list (ACL) on the Cisco IM and Presence Service to allow for communications with Microsoft Live Communications Server 2005 or Office Communications Server 2007 or Microsoft Lync.
- You must enable each user for use of Microsoft Office Communicator in the Cisco IM and Presence Service configuration, in addition to enabling each user for presence in Unified CM.
- Take into account bandwidth considerations for Microsoft Office Communicator login due to the exchange of configuration information between Microsoft Office Communicator and the Microsoft Communications Server, and due to initial communication with the Cisco IM and Presence Service CTI gateway.
- The parameters that are required for Microsoft Office Communications Server 2007 or Microsoft Lync have changed names from Live Communications Server 2005. The TEL URI parameter defined in Live Communications Server 2005 is the same as the Line URI parameter in Office Communications Server 2007 or Microsoft Lync, and the Remote Call Control SIP URI parameter defined in Live Communications Server 2005 is the same as the Server URI parameter in Office Communications Server 2007 or Microsoft Lync.
- To address the issue of a reverse look-up of a directory number that corresponds to a user, use the guidelines documented in the *Release Notes for Cisco IM and Presence*, available at

  http://www.cisco.com/en/US/products/ps6837/prod_release_notes_list.html

IBM Lotus Sametime

Cisco Systems, Inc. provides the following guidelines around how best to integrate IBM Lotus Sametime Server with Cisco Unified Communications, but does not contend configuration, deployment, or best practice procedures for IBM Communications products.

For all setup, configuration, and deployment of IBM Lotus Sametime Server, refer to the documentation at

http://www.ibm.com/

Cisco does not provide configuration, deployment, or best practice procedures for IBM communications products, but Cisco does provide the guidelines listed below for integrating IBM Lotus Sametime Server with a Cisco Unified Communications system.

Guidelines for Integrating Cisco IM and Presence with IBM Lotus Sametime Server (Version 7.5.1 and Later)

Click-to-call and click-to-conference functionality integrated within the IBM Lotus Sametime client is handled via a Cisco Call Control plugin resident on IBM Lotus Sametime Server. The integration into Cisco Unified Communications for click-to-call and click-to-conference functionality is handled via the SIP trunk interface with Unified CM. The integration into Cisco Unified Communications for presence functionality is handled via the SIP/SIMPLE interface with Cisco IM and Presence.

- The Unified CM SIP trunk for click-to-call and click-to-conference functionality must be configured for out-of-dialog REFER processing. Enable the Accept Out-of-Dialog REFER checkbox in the SIP Trunk Security Profile associated with the SIP trunk communicating with IBM Lotus Sametime Server.
- The Cisco Call Control plugin, resident on IBM Lotus Sametime Server, maintains a configured list of Unified CMs utilized in a round-robin manner. This list should be populated with the IP address of Unified CM subscribers that have been configured with the out-of-dialog REFER SIP trunks.

  The Unified CM list can also be configured with DNS SRV; however, currently this SRV logic is used for redundancy only and not for load balancing, therefore it is not a recommended setting.
- Deployment topologies using IBM Lotus Sametime Server typically will integrate with multiple Unified CM clusters due to the capacity differences between the two systems. With the Cisco Click-to-Call plugin utilizing a list of Unified CMs in a round-robin fashion, a call initiation can result in a REFER being sent to a cluster different from the user's home cluster. The Unified CM receiving this call setup will process the REFER and generate an INVITE to the appropriate destination to complete the call setup.
- Traffic marking has not been implemented fully with the Cisco Call Control plugin. Therefore, follow the traffic marking guidelines in the *Enterprise QoS Solution Reference Network Design (SRND)*, available at

  http://www.cisco.com/go/designzone

CHAPTER 24

Cisco Collaboration Clients and Applications

Revised: August 31, 2012; OL-27282-05

Note This chapter has been revised significantly for the current release of this document. Cisco recommends that you read this entire chapter before attempting to deploy collaboration clients and applications in your Cisco Unified Communications System.

Cisco Collaboration Clients and Applications provide an integrated user experience and extend the capabilities and operations of the Cisco Unified Communications System. These clients and applications enable collaboration both inside and outside the company boundaries by bringing together, in a single easy to use collaboration client, applications such as online meetings, presence notification, instant messaging, audio, video, voicemail, and many more.

Several Cisco collaboration clients and applications are available. Third-party XMPP clients and applications are also supported. Cisco clients use the Cisco Unified Client Services Framework to integrate with underlying Unified Communication services through a common set of interfaces. In general, each client provides support for a specific operating system or device type. Use this chapter to determine which collaboration clients and applications are best suited for your deployment. The client-specific sections of this chapter also provide relevant deployment considerations, planning, and design guidance around integration into the Cisco Unified Communications System.

The following collaboration clients and applications are supported by the Cisco Unified Communications System:

- Cisco Jabber for Windows and Mac

  Cisco Jabber for Windows and Cisco Jabber for Mac are Unified Communications clients that provide robust and feature-rich collaboration capabilities including standards-based IM and presence, audio and video, visual voicemail, desktop sharing, deskphone control, Microsoft Office integration and directory integration.

  Cisco Jabber for Windows and Cisco Jabber for Mac can be deployed to use on-premises services in which Cisco IM and Presence (formerly Cisco Unified Presence) and Cisco Unified Communications Manager provide client configuration, instant messaging and presence, and user and device management. Cisco Jabber for Windows and Cisco Jabber for Mac can also be deployed to use cloud-based services through integration with Cisco WebEx Messenger service.

  Cisco Jabber forms the basis of the next generation of Cisco collaboration clients, which will supersede Cisco Unified Personal Communicator and Cisco Unified Integration for WebEx Connect in future Cisco Unified Communications System releases. Therefore, only Cisco Jabber for Windows and Cisco Jabber for Mac features and functionality are discussed in this release of the *Cisco Unified Communications System SRND*. Cisco Unified Personal Communicator and Cisco

Unified Integration for WebEx Connect clients are still available and supported, but their features and functionality have not changed from Cisco Unified Communications System release 8.*x*. For design guidance on Unified Personal Communicator and WebEx Connect clients, refer to the clients information in the *Cisco Unified Communications System 8.x SRND*, available at

http://www.cisco.com/go/ucsrnd

- Cisco Jabber for Everyone

  Cisco Jabber for Everyone makes Cisco Jabber presence and instant messaging (IM) available at zero cost. Jabber IM client applications and Cisco IM and Presence, zero-cost licenses are available to Cisco Unified Communications Manager customers on the following platforms: Windows, Mac, Android, BlackBerry, iPhone, iPad, and Cisco Jabber Web SDK. For more information on Jabber for Everyone, refer to the Jabber for Everyone Solution Overview, available at

  http://www.cisco.com/en/US/docs/voice_ip_comm/cups/8_6/english/jabber_for_everyone/CUP0_BK_JE526021_00_jabber-for-everyone-solution-overview.html

- Cisco Jabber for mobile devices

  Cisco provides collaboration clients for the following mobile devices: Android, BlackBerry, and Apple iOS devices such as iPhone and iPad. For more information on Cisco Jabber for mobile devices, see the chapter on Mobile Unified Communications, page 25-1.

- Cisco Jabber Video for TelePresence (Movi)

  Cisco Jabber Video for TelePresence (Jabber Video) extends the reach of telepresence. Jabber Video works with a compatible PC or Mac and a webcam or Cisco TelePresence PrecisionHD camera to provide high-definition video communications to mobile workers, allowing them to connect to telepresence systems. Cisco Jabber Video for TelePresence is a video-only client that is used with the Cisco TelePresence Video Communication Server (Cisco VCS). For more information on Cisco Jabber Video for TelePresence (Movi), refer to the documentation at

  http://www.cisco.com/en/US/products/ps11328/tsd_products_support_series_home.html

- Cisco Virtual Experience Clients

  The Cisco Virtualization Experience Clients (VXC) are the integral collaboration components of the Cisco Virtualization Experience Infrastructure (VXI). The VXCs provide user access to data, applications, and services across various network environments, as well as user preferences and device form factors for a fully integrated voice, video, and virtual desktop environment.

- Cisco UC Integration^{TM} for Microsoft Lync

  Cisco UC Integration^{TM} for Microsoft Lync allows for integrated Cisco Unified Communications services with Microsoft Lync and Microsoft Office Communications Server (OCS) R2 using the Cisco Unified Client Services Framework, while delivering a consistent user experience. The solution extends the presence and instant messaging capabilities of Microsoft Lync by providing access to a broad set of Cisco Unified Communications services, including standards-based audio and video, unified messaging, web conferencing, deskphone control, and telephony presence.

- Third-party XMPP clients and applications

  Cisco IM and Presence, with support for SIP/SIMPLE and Extensible Messaging and Presence Protocol (XMPP), provides support of third-party clients and applications to communicate presence and instant messaging updates between multiple clients. Third-party XMPP clients, MomentIM, Adium, Spark, Pidgin, and others, allow for enhanced interoperability across various desktop operating systems. In addition, web-based applications can obtain presence updates, instant messaging, and roster updates using the HTTP interface with SOAP, REST, or BOSH (based on the Cisco AJAX XMPP Library API). For additional information on the third-party open interfaces, see the chapter on Cisco IM and Presence, page 23-1.

What's New in This Chapter

Table 24-1 lists the topics that are new in this chapter or that have changed significantly from previous releases of this document.

Table 24-1 New or Changed Information Since the Previous Release of This Document

New or Revised Topic	Described in:	Revision Date
Product name change from Cisco WebEx Connect service to Cisco WebEx Messenger service	Various sections throughout this chapter	August 31, 2012
Numerous updates for Cisco Unified Communications System Release 9.0	All sections of this chapter	June 28, 2012

Cisco Unified Client Services Framework Architecture

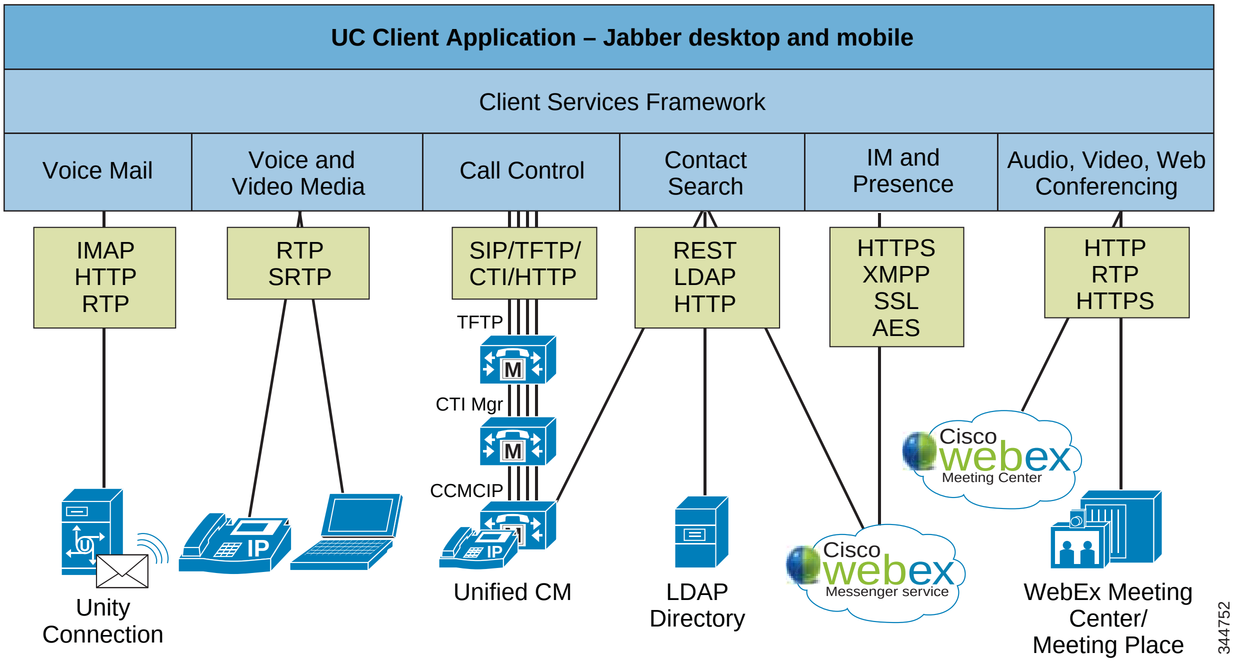
Cisco Jabber for Windows, Cisco Jabber for Mac, and Cisco UC Integration™ for Microsoft Lync all use the Client Services Framework as a base building block for the client application. Cisco Unified Client Services Framework is a software application that combines a number of services into an integrated client. An underlying framework is provided for integration of Unified Communications services, including audio, video, web collaboration, visual voicemail, and so forth, into a presence and instant messaging application.

These Cisco Jabber client applications reside on top of the Clients Services Framework, which provides a simplified client interface and an abstraction layer that allows access to the following underlying communications services:

- SIP-based call control for voice and video softphone clients from Unified CM
- Deskphone call control and "Click to Dial" services from Unified CM's CTI interface
- Voice and video media termination for softphone clients
- Instant messaging and presence services using XMPP, from either the Cisco IM and Presence Service or Cisco WebEx. Cisco WebEx Meeting Center also offers hosted collaboration services such as online meetings and events
- Scheduled audio, video and web conferencing services
- Desktop sharing using either, video desktop sharing (BFCP) or WebEx desktop sharing
- Visual voicemail services from Cisco Unity Connection using IMAP
- Contact management using:
  - Unified CM User Data Service (UDS) as a contact source (LDAP directory synchronization supported)
  - Directory access using Microsoft Active Directory or supported LDAP directories as a contact source
  - WebEx Messenger service
  - Client Services Framework cache and contact list
- Microsoft Office Integration, which provides user availability status and messaging capabilities directly through the user interface of Microsoft Office applications such as Microsoft Outlook

The ability to communicate and abstract services and APIs, as shown in Figure 24-1, allows the Client Services Framework to coordinate the management of protocols to these services and APIs, handle event notifications, and control the low-level connection logic for local system resources.

Figure 24-1 Cisco Unified Client Services Framework



Client Services Framework Services

The following sections discuss the services provided by the Client Services Framework in more detail.

Client Services Framework – Call Control

Cisco Unified Client Services Framework can operate in one of two modes for call control:

- Softphone Mode — Using audio and video on a computer

  The Client Services Framework in softphone mode is directly registered to Unified CM as a SIP endpoint for audio and video call control functionality, and it is configured on Unified CM as device type Client Services Framework.

- Deskphone Control Mode — Using a Cisco IP Phone for audio (and video, if supported)

  The Client Services Framework in deskphone control mode does not register with Unified CM using SIP, but instead uses CTI/JTAPI to initiate, monitor, and terminate calls, monitor line state, and provide call history, while controlling a Cisco Unified IP Phone. The Cisco CallManager Cisco IP Phone (CCMCIP) service on Unified CM is used by the Client Services Framework to retrieve a list of devices associated with each user. This list of devices is used by a client in deskphone mode to choose which Cisco IP Phone it wishes to control.

Softphone Mode

When operating in softphone mode, the Client Services Framework is a SIP line-side registered device on Unified CM, utilizing all the call control capabilities and functionality of a Cisco Unified IP Phone, including configuration of registration, redundancy, regions, locations, dial plan management, authentication, encryption, user association, and so forth. The Client Services Framework supports a single line appearance for the user.

The SIP registered device of the Client Services Framework must be factored in as a regular SIP endpoint, as any other SIP registered endpoint, for purposes of sizing calculations for a Unified CM cluster. The Client Services Framework in softphone mode uses the CCMCIP service to discover its device name for registration with Unified CM.

Deskphone Control Mode

When operating in deskphone control mode, the Client Services Framework uses CTI/JTAPI to provide the ability to place, monitor, and receive calls using Cisco Unified IP Phones. When audio calls are received or placed in this mode, the audio path is through the Cisco Unified IP Phone. For video calls, the video stream can originate and terminate either on the Cisco IP Phone (if it has a camera) or on the computer using a Cisco Unified Video Advantage camera. The Client Services Framework uses the CCMCIP service on Unified CM to discover the associated devices of the user.

When using deskphone control mode for the Client Services Framework, factor the CTI scaling numbers into the Unified CM deployment calculations. For additional information around capacity planning, see the chapter on Unified Communications Design and Deployment Sizing Considerations, page 29-1.

Client Services Framework – Audio and Video Media

A number of standard audio and video codecs for use in low bandwidth or high fidelity deployments are supported with the Client Services Framework. Audio codecs include G.729a, G.711, and G.722.1, while video codecs include H.264 AVC (Advanced Video Coding) with support for H.264 baseline profile levels 1 through 3.1. Video formats supported include QCIF, CIF, VGA, and 720p HD at a rate of up to 30 frames per second.

The Client Services Framework always attempts to transmit and receive high definition video; however, there are a number of throttling factors that need to be considered when deploying video. These throttling considerations include the capability of the device communicating with, the local processing capability of the PC, administrative or user settings, local camera capabilities, and any call admission control policies in place.

There are a number of references the Client Services Framework can use to determine the video frame rate for a call. The processing power and CPU used by the client play an important role in determining the video frame rate used. Another decision point is based on the Windows Experience Index (WEI) for the personal computer being used (see http://technet.microsoft.com/en-us/library/cc507870.aspx). The minimum values for encoding and decoding high definition video require a processor WEI encode value of 5.9 and a bandwidth requirement of 1 Mbps for 720p at 15 frames per second or 2 Mbps for 720p at 30 frames per second.

For a listing of client system requirements, video frame rates based on H.264 Level, and WEI encode and decode values, refer to the Client Application Release Notes, page 24-6.

Bandwidth utilization for audio and video calls from the Client Services Framework can be maintained using the Unified CM regions and locations call admission control mechanisms. Administratively placing the Client Services Framework in a Region provides the ability to control the per-call voice and video bandwidth usage and the preferred audio codecs to be used for calls within and between regions. Unified CM locations-based call admission control, and/or the use of RSVP, provides intra-location and inter-location audio and video bandwidth control. The Client Services Framework requires the Unified CM region per-call bandwidth settings to be sufficient to cover both the audio and video portions of the call. For example, to have a video call at a frame size of 720p and a frame rate of 30 frames per second, the session bit rate needs to be 2,000 kbps just for video; therefore, the region bandwidth for a call must account for the audio portion at 64 kbps (assuming a G.711 or G.722 codec) as well as the video portion at 2,000 kbps (assuming 720p at 30 fps). For more information on Unified CM support for regions and locations for call admission control, see the chapter on Call Processing, page 8-1.

Quality of Service for Audio and Video Media from Softphones

An integral part of the Cisco Unified Communications network design recommendations is to classify or mark voice and video traffic so that it can be prioritized and appropriately queued as it traverses the Unified Communications network. A number of options exist to set the DSCP values of audio and video traffic generated by clients. For example:

- Using a Unified CM Trusted Relay Point to enforce DSCP marking for QOS on behalf of a softphone client registered with Unified CM.
- Using network-based access control lists (ACLs) to mark DSCP values for voice and video traffic.
- Using Active Directory Group Policy to mark DSCP values for voice and video traffic. Note that many operating systems limit the ability of applications to mark traffic with DSCP values for QoS treatment.

QoS Enforcement Using a Trusted Relay Point (TRP)

A Trusted Relay Point (TRP) can be used in conjunction with the device mobility feature to enforce and/or re-mark the DSCP values of media flows from endpoints. This feature allows QoS to be enforced for media from endpoints such as softphones, where the media QoS values might have been modified locally.

A TRP is a media resource based upon the existing Cisco IOS media termination point (MTP) function.

Endpoints can be configured to **Use Trusted Relay Point**, which will invoke a TRP for all calls.

For QoS enforcement, the TRP uses the configured QoS values for media in Unified CM's Service Parameters to re-mark and enforce the QoS values in media streams from the endpoint. If no TRP is available, the call will proceed without modification of the DSCP value of the traffic generated by the endpoint. TRP functionality is supported by Cisco IOS MTPs and transcoding resources. (Use Unified CM to check **Enable TRP** on the MTP or transcoding resource to activate TRP functionality.)

Client Application Release Notes

- Cisco Jabber for Windows

  http://www.cisco.com/en/US/products/ps12511/prod_release_notes_list.html
- Cisco Jabber for Mac

  http://www.cisco.com/en/US/products/ps11764/prod_release_notes_list.html
- Cisco UC Integration^{TM} for Microsoft Lync

  http://www.cisco.com/en/US/products/ps10317/prod_release_notes_list.html

Client Services Framework – Instant Messaging and Presence Services

Instant messaging and presence services for Jabber clients can be provided through the Cisco Client Services Framework XMPP interface. Cisco offers instant messaging and presence services with the following products:

- Cisco IM and Presence
- WebEx Messenger service

Note With Cisco UC Integration^{TM} for Microsoft Lync, instant messaging and presence services are provided by Microsoft.

The choice between Cisco IM and Presence or WebEx Messenger service for instant messaging and presence services can depend on a number of factors. WebEx Messenger service deployments use WebEx as a cloud-based service that is accessible from the internet. On-premises deployments based on Cisco IM and Presence provide the administrator with direct control over their IM and presence platform and also allow presence federation using SIP/SIMPLE to Microsoft IM and presence services.

For information on the full set of features supported by each IM and Presence platform, refer to the following documentation:

- Cisco IM and Presence

  http://www.cisco.com/en/US/products/ps6837/products_data_sheets_list.html

- WebEx Messenger service

  http://www.cisco.com/en/US/products/ps10528/index.html

  http://www.cisco.com/en/US/products/ps10528/prod_literature.html

Client Services Framework – Audio, Video and Web Conferencing Services

Access to scheduled conferencing services for clients can be provided through a Cisco Client Services Framework HTTP interface. Cisco audio, video and web-based scheduled conferencing services can be provided by using the cloud-based WebEx Meeting Center service or a combination of on-premises MeetingPlace audio and video conferencing services and WebEx cloud-based web conferencing services. For more information, refer to the following documentation:

- Cisco Unified MeetingPlace

  http://www.cisco.com/en/US/products/sw/ps5664/ps5669/products_data_sheets_list.html

- WebEx Meeting Center

  http://www.psimeeting.com/pdf/WebEx_Meeting_Center.pdf

Client Services Framework – Contact Management

The Client Services Framework can handle the management of contacts through a number of sources, including the following:

- Cisco Unified CM User database via the User Data Service (UDS)
- LDAP directory integration
- WebEx Messenger service

Contacts can also be stored and retrieved locally using either of the following:

- Client Services Framework Cache
- Local address books and contact lists

The Client Services Framework uses reverse number lookup to map an incoming telephone number to a contact, in addition to photo retrieval. The Client Services Framework contact management allows for up to five search bases to be defined for LDAP queries.

Cisco Unified CM User Data Service (UDS)

UDS provides clients with a contact search service on Cisco Unified Communications Manager. You can synchronize contact data into the Cisco Unified CM User database from Microsoft Active Directory or other LDAP directory sources. Clients can then automatically retrieve that contact data directly from Unified CM using the UDS REST interface.

LDAP Directory

You can configure a corporate LDAP directory to satisfy a number of different requirements, including the following:

- User provisioning — You can provision users automatically from the LDAP directory into the Cisco Unified Communications Manager database using directory integration. Cisco Unified CM synchronizes with the LDAP directory content so that you avoid having to add, remove, or modify user information manually each time a change occurs in the LDAP directory.
- User authentication — You can authenticate users using the LDAP directory credentials. Cisco IM and Presence synchronizes all the user information from Cisco Unified Communications Manager to provide authentication for client users.
- User lookup — You can enable LDAP directory lookups to allow Cisco clients or third-party XMPP clients to search for contacts in the LDAP directory.

WebEx Directory Integration

WebEx Directory Integration is achieved through the WebEx Administration Tool. WebEx imports a comma-separated value (CSV) file of your enterprise directory information into its WebEx Messenger service. For more information, refer to the documentation at

http://www.webex.com/webexconnect/orgadmin/help/index.htm?toc.htm?17444.htm

Client Services Framework Cache

The Client Services Framework maintains a local cache of contact information derived from previous directory queries and contacts already listed, as well as the local address book or contact list. If a contact for a call already exists in the cache, the Client Services Framework does not search the directory. If a contact does not exist in the cache, the Client Services Framework performs a directory search.

Directory Search

When a contact cannot be found in the local Client Services Framework cache or contact list, a search for contacts can be made. The WebEx Messenger user can utilize a predictive search whereby the cache, contact list, and local Outlook contact list are queried as the contact name is being entered. If no matches are found, the search continues to query the corporate directory (WebEx Messenger database).

Client Services Framework – Dial Plan Considerations

Dial plan and number normalization considerations must be taken into account when deploying the Client Services Framework as part of any Unified Communications endpoint strategy. The Client Services Framework, as part of a Unified Communications collaboration client, will typically use the directory for searching, resolving, and adding contacts. The number that is associated with those contacts must be in a form that the client can recognize, resolve, and dial.

Deployments may vary, depending on the configuration of the directory and Unified CM. In the case where the directory contains E.164 numbering (for example, +18005551212) for business, mobile, and home telephone numbers and Unified CM also contains an E.164 dial plan, the need for additional dial rules is minimized because every lookup, resolution, and dialed event results in an E.164 formatted dial string.

If a deployment of Unified CM has implemented a private dial plan (for example, 5551212), then translation of the E.164 number to a private directory number needs to occur on Unified CM. Outbound calls can be translated by Unified CM translation patterns that allow the number being dialed (for example, +18005551212) to be presented to the endpoint as the private number (5551212 in this example). Inbound calls can be translated by means of directory lookup rules. This allows an incoming number of 5551212 to be presented for reverse number lookup caller identification as +18005551212.

Private numbering plan deployments may arise, where the dial plan used for your company and the telephone number information stored in the LDAP directory may require the configuration of translation patterns and directory lookup rules in Cisco Unified Communications Manager to manage number format differences. Directory lookup rules define how to reformat the inbound call ID to be used as a directory lookup key. Translation patterns define how to transform a phone number retrieved from the LDAP directory for outbound dialing.

Translation Patterns

Translation patterns are used by Unified CM to manipulate the dialed digits before a call is routed, and they are strictly handled by Unified CM. Translation patterns are the recommended method for manipulating dialed numbers. For additional guidelines on translation pattern usage and dial plan management, see the chapter on Dial Plan, page 9-1.

Application Dialing Rules

Application dialing rules can be used as an alternative to translation patterns to manipulate numbers that are dialed. Application dialing rules can automatically strip numbers from, or add numbers to, phone numbers that the user dials. Application dial rules are configured in Unified CM and are downloaded through TFTP to the client from Unified CM. Translation patterns are the recommended method for manipulating dialed numbers.

Directory Lookup Rules

Directory lookup rules transform caller identification numbers into numbers that can be looked up in the directory. A directory lookup rule specifies which numbers to transform based on the initial digits and the length of the number. Directory lookup rules are configured in Unified CM and are downloaded through TFTP to the client from Unified CM.

Client Transformation

Before a call is placed through contact information, the client application removes everything from the phone number to be dialed, except for letters and digits. The application transforms the letters to digits and applies the dialing rules. The letter-to-digit mapping is locale-specific and corresponds to the letters found on a standard telephone keypad for that locale. For example, for a US English locale, 1-800-4UCSRND transforms to 18004827763. Users cannot view or modify the client transformed numbers before the application places the call.

Deploying Client Services Framework

Because the Client Services Framework is a fundamental building block for Unified Communications client integration and communication, it is necessary to deploy these devices to a number of users. Cisco recommends using the Bulk Administration Tool for the Client Services Framework deployment. The administrator can create a phone template for device pool, device security profile, and phone buttons, and can create a CSV data file for the mapping of device name to directory number. The administrator can also create a User template to include user groups and CTI, if enabled, as well as a CSV data file to map users to the appropriate controlled device.

Capacity Planning for Client Services Framework

Cisco Unified Client Services Framework operates as either a SIP endpoint registered to Unified CM or as a deskphone controller of a Cisco Unified IP Phone using a CTI connection to Unified CM. When planning a deployment using the Client Services Framework, Cisco partners and employees can use the Cisco Unified Communications Sizing Tool (available at http://tools.cisco.com/cucst) to assist in the appropriate sizing of SIP registered endpoints and CTI controlled devices. The following additional items must be considered for a Client Services Framework deployment:

- TFTP — When configured in softphone mode, a Client Services Framework device configuration file is downloaded through TFTP to the client for Unified CM call control configuration information. In addition, any application dial rules or directory lookup rules are also downloaded through TFTP to Client Services Framework devices.
- CTI — When configured in deskphone mode, the Client Services Framework establishes a CTI connection to Unified CM upon login and registration to allow for control of the IP phone. Unified CM supports up to 40,000 CTI connections. If you have a large number of clients operating in deskphone mode, make sure that you evenly distribute those CTI connections across all Unified CM subscribers running the CTIManager service. This can be achieved by creating multiple CTI Gateway profiles, each with a different pair of CTIManager addresses, and distributing the CTI Gateway profile assignments across all clients using deskphone mode.
- CCMCIP — The Client Services Framework uses the Cisco CallManager Cisco IP Phone (CCMCIP) service to gather information about the devices associated with a user, and it uses this information to provide a list of IP phones available for control by the client in deskphone control mode. The Client Services Framework in softphone mode uses the CCMCIP service to discover its device name for registration with Unified CM.
- IMAP — When configured for voicemail, the Client Services Framework updates and retrieves voicemail through an IMAP connection to the mailstore.
- LDAP — Client login and authentication, contact profile information, and incoming caller identification are all handled through a query to the LDAP directory, unless stored in the local Client Services Framework cache.
- UDS — The UDS service can be used by clients to search for contacts in the Unified CM User database. Like LDAP directory searches, UDS contact searches take place if the requested contact cannot be found in the local Client Services Framework cache.

High Availability for Client Services Framework

Cisco Unified Client Services Framework provides primary and secondary servers for each of the following configuration components: TFTP server, CTIManager, CCMCIP server, voicemail server, UDS server, and LDAP server. When operating in softphone mode, the Client Services Framework is registered with Cisco Unified CM as a SIP endpoint, and it supports all of the registration and redundancy capabilities of a registered endpoint of Unified CM. When operating in deskphone mode, the Client Services Framework is controlling a Cisco Unified IP Phone using CTI, and it supports configuration of a primary and secondary CTIManager in the CTIManager Profile. For additional details on CTI deployment, see the chapter on Call Processing, page 8-1.

Design Considerations for Client Services Framework

Observe the following design considerations when deploying the Cisco Unified Client Services Framework:

- The administrator must determine how to install, deploy, and configure the Unified Client Services Framework in their organization. Cisco recommends using a well known installation package such as Altiris to install the application.
- The userid and password configuration of the Cisco Unified Client Services Framework user must match the userid and password of the user stored in the LDAP server to allow for proper integration of the Unified Communications and back-end directory components.
- The directory number configuration on Cisco Unified CM and the telephoneNumber attribute in LDAP should be configured with a full E.164 number. A private enterprise dial plan can be used, but it might involve the need to use translation patterns or application dialing rules and directory lookup rules.
- The use of deskphone mode for control of a Cisco Unified IP Phone uses CTI; therefore, when sizing a Unified CM deployment, you must also account for other applications that require CTI usage.
- For firewall and security considerations, the port usage required for the Client Services Framework and corresponding applications being integrated can be found in the product release notes for each application.
- To reduce the impact on the amount of traffic (queries and lookups) to the back-end LDAP servers, configure concise LDAP search bases for the Client Services Framework rather than a top-level search base for the entire deployment.

Common Deployment Models for Jabber Clients

Cisco Jabber for Windows and Jabber for Mac clients support the following deployment models:

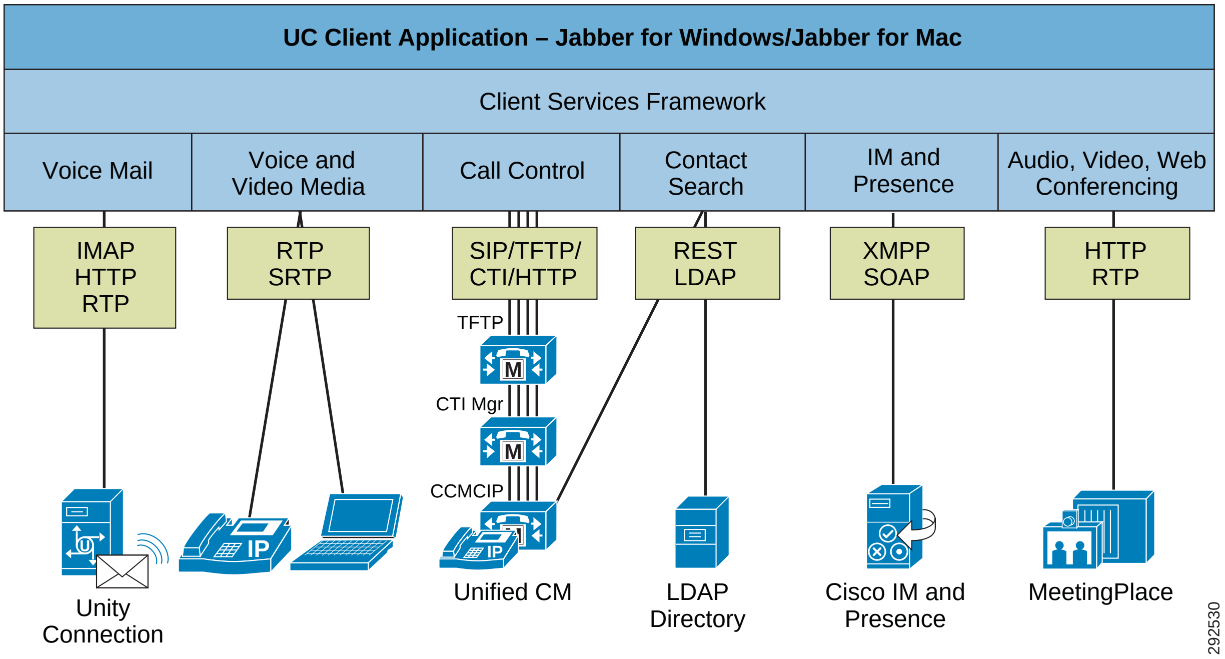
- On-Premises Deployment Model, page 24-12
- Cloud-Based Deployment Model, page 24-13
- WHybrid Cloud-Based and On-Premises Deployment Model, page 24-14

Your choice of deployment will depend primarily upon your product choice for IM and presence and the requirement for additional services such as voice and video, voicemail, and deskphone control.

On-Premises Deployment Model

The on-premises deployment model is one in which all services are set up and configured on an enterprise network that you manage and maintain. (See Figure 24-2.)

Figure 24-2 Jabber On-Premises Deployment Model



The on-premises deployment model for Cisco Jabber for Windows relies on the following components:

- Cisco Unified Communications Manager provides user and device configuration capabilities.
- Cisco IM and Presence provides instant messaging and presence services.
- Microsoft Active Directory or another supported LDAP directory provides contact sources.

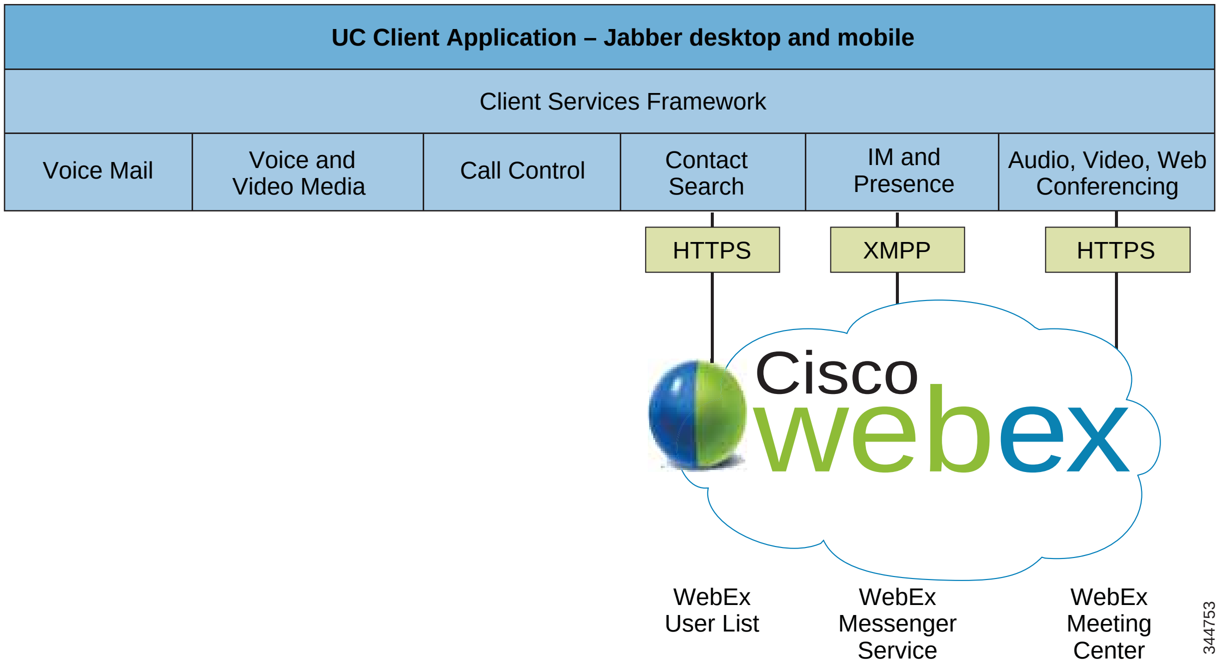
These components are the essential requirements to achieve a base deployment of Cisco Jabber clients. After you set up and configure a base deployment, you can set up and configure additional deployment options such as:

- Voice — Provides audio call capabilities.
- Video — Provides capabilities to enable users to transmit and receive video calls.
- Voicemail — Provides voicemail capabilities that users can retrieve directly in the Cisco Jabber client user interface or when users dial their voicemail number.
- Desktop sharing — Enables users to share their desktops.
- Microsoft Office integration — Provides user availability status and messaging capabilities directly through the user interface of Microsoft Office applications such as Microsoft Outlook.

Cloud-Based Deployment Model

The cloud-based deployment model is one in which all, or most, services are hosted on Cisco WebEx Messenger service. You manage and monitor your cloud-based deployment with the Cisco WebEx Administration Tool. (See Figure 24-3.)

Figure 24-3 Jabber Cloud-Based Deployment Model



The cloud-based deployment model for Cisco Jabber for Windows relies on Cisco WebEx Messenger service for the following services:

- Instant messaging and chat capabilities
- Presence capabilities for users
- User configuration and contact sources

These services are the essential components required to achieve a base deployment of Cisco Jabber for Windows. After you set up and configure a base deployment, you can set up and configure additional deployment options such as:

- Cisco WebEx Meeting Center — Offers hosted collaboration features such as online meetings and events.
- Microsoft Office integration — Provides user availability status and messaging capabilities directly through the user interface of Microsoft Office applications such as Microsoft Outlook.

For information on WebEx Messenger service configuration for Jabber Clients, refer to the WebEx administration guide available at

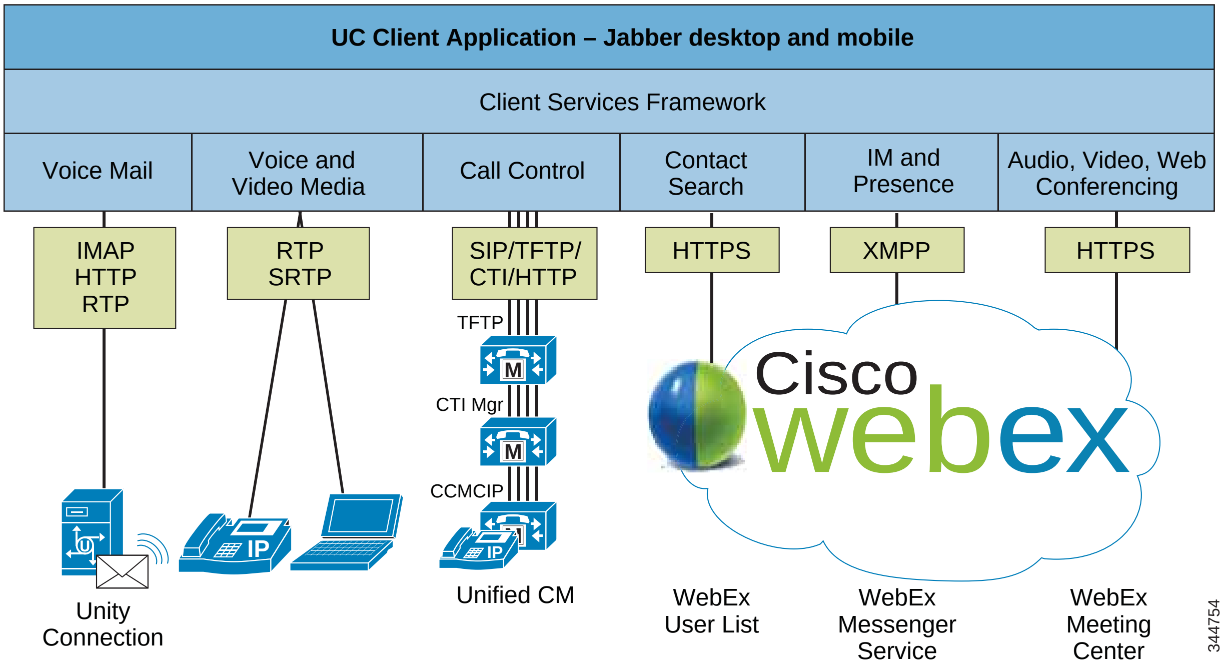
http://www.webex.com/webexconnect/orgadmin/help/index.htm

WHybrid Cloud-Based and On-Premises Deployment Model

A hybrid deployment is one in which the cloud-base services hosted on Cisco WebEx Messenger service are combined with the following components of an on-premises deployment (see Figure 24-4):

- Cisco Unified Communications Manager provides user and device services.
- Cisco Unity Connection provides voicemail services.

Figure 24-4 Jabber Hybrid Cloud -Based and On-Premises Deployment Model



Integration with Cisco WebEx Messenger service, Cisco Unified Communications Manager, and Cisco Unity Connection lets you extend your cloud-based deployment and enable the following deployment options:

- Voice — Provides audio calls managed through Cisco Unified Communications Manager.
- Video — Provides capabilities to enable users to transmit and receive video calls.
- Voicemail — Provides voicemail capabilities that users can retrieve directly in the Cisco Jabber for Windows user interface or when users dial their voice mailbox number.
- Desktop Sharing — Enables users to share their desktops.

Client-Specific Design Considerations

The following sections discuss design considerations that are specific to Cisco Jabber for Windows and Jabber for Mac. For common design considerations for these client types, use the design guidance provided in the section on Cisco Unified Client Services Framework Architecture, page 24-3.

Cisco Jabber for Windows

Cisco Jabber for Windows is a Unified Communications client that provides robust and feature-rich collaboration capabilities that include the following:

- Chat over XMPP, including:
  - Rich text formatting
  - File transfer
  - Screen capture
  - Group chat
  - Emoticons
- Desk phone control
- Software phone calling
- High definition video
- Video desktop sharing
- Visual voicemail
- Microsoft Office integration
- Directory integration
- Support for custom embedded tabs to render HTML content

Jabber for Windows clients support on-premises, cloud-based, and hybrid deployment models.

Client Launch Sequence

The following steps describe the initial Cisco Jabber for Windows launch sequence from a high level.

1. Retrieve the presence server type (WebEx or Unified IM and Presence) from jabber-bootstrap.properties in the installation directory.
2. Authenticate with the presence server.
3. Retrieve profile details and connect to available services such as:
   - TFTP servers
   - CTI gateway servers
   - Cisco CallManager Cisco IP Phone (CCMCIP) servers
   - Voicemail servers
   - Directory servers

4. Retrieve Cisco Jabber for Windows configuration files. These XML files are loaded from the TFTP server and can contain additional configuration information such as:
   - Client configuration parameters for automatic updates, password reset URL, and so forth
   - Client policy parameters to allow/disallow screen captures, files transfers, and so forth
   - Directory service information such as directory type and directory attribute mappings
   - Application dial rules and directory look up rules

Directory Integration

Cisco Jabber for Windows defaults to using Enhanced Directory Integration (EDI), which uses preconfigured directory attribute mappings for integration with Microsoft Active Directory. For integration with an LDAP directory that requires custom attribute mapping, these attribute mappings can be created in a configuration file that can be downloaded to the client from a Unified CM TFTP server. Cisco Jabber for Windows does *not* use directory settings that are specified in the Cisco IM and Presence Service configuration.

Jabber for Windows also supports the Unified CM User Data Service (UDS), which allows a client to search for contacts using the Unified CM user database (which may be synchronized with an LDAP directory).

Video Rate Adaptation and Resolution

Cisco Jabber for Windows uses the Cisco Precision Video Engine and ClearPath technology to optimize video media. The Cisco Precision Video Engine uses fast video rate adaptation to negotiate optimum video quality based on network conditions. Video rate adaptation dynamically scales video quality upward when video transmission begins. Cisco Jabber for Windows also saves history so that subsequent video calls begin at the optimal resolution. ClearPath technology improves resolution on sub-optimal networks by reducing the effects of packet loss.

Jabber for Windows supports desktop sharing using either WebEx Desktop Share or Video Desktop Share (using BFCP).

For more information on the configuration options and administration of a Jabber for Windows client, refer to the *Cisco Jabber for Windows Administration Guide*, available at

http://www.cisco.com/en/US/products/ps12511/prod_installation_guides_list.html

Also refer to the Jabber for Windows Release Notes, available at

http://www.cisco.com/en/US/products/ps12511/prod_release_notes_list.html

For specific details about Jabber for Windows client features, refer to the Jabber for Windows data sheet, available at

http://www.cisco.com/en/US/products/ps12511/products_data_sheets_list.html

Cisco Jabber for Mac

Cisco Jabber for Mac is a Unified Communications client that provides robust and feature-rich collaboration capabilities that include the following:

- Chat over XMPP, including:
  - Rich text formatting
  - File transfer
  - Screen capture
  - Group chat
  - Emoticons
- Desk phone control
- Software phone calling
- Visual voicemail
- Directory integration

Cisco Jabber for Mac clients support on-premises, cloud-based, and hybrid deployment models.

Client Launch Sequence

The following steps describe the initial Cisco Jabber for Mac launch sequence from a high level.

1. Retrieve the presence server type (WebEx or Cisco IM and Presence) from jabber-bootstrap.properties in the installation directory.
2. Authenticate with the presence server.
3. Retrieve profile details and connect to available services such as:
   - TFTP servers
   - CTI gateway servers
   - Cisco CallManager Cisco IP Phone (CCMCIP) servers
   - Voicemail servers
   - Directory servers
4. Retrieve Cisco Jabber for Mac configuration files. These XML files are loaded from the TFTP server and can contain additional configuration information such as application dial rules and directory look up rules.

Directory Integration

Cisco Jabber for Mac supports Microsoft Active Directory and LDAP directory integration. With Cisco IM and Presence, Jabber for Mac supports the following directory server types: Microsoft Active Directory, iPlanet, Sun ONE, and OpenLDAP. When one of these directory server types is selected, the presence server populates the LDAP attribute map with Cisco Jabber user fields and LDAP user fields. These default mappings can be modified to support the attribute mappings of other LDAP directory servers. For Cisco Jabber for Mac clients using the Cisco IM and Presence Service, you must configure the LDAP server and attribute mappings for your environment.

Jabber for Mac does *not* support Unified CM UDS contact searches.

Jabber for Mac supports desktop sharing using WebEx Desktop Share.

For more information on the configuration options and administration of a Jabber for Mac client, refer to the Cisco Jabber for Mac Installation and Configuration Guide, available at

http://www.cisco.com/en/US/products/ps11764/prod_maintenance_guides_list.html

Also refer to the Jabber for Mac Release Notes, available at

http://www.cisco.com/en/US/products/ps11764/prod_release_notes_list.html

For specific details about Jabber for Mac client features, refer to the Jabber for Mac data sheet, available at

http://www.cisco.com/en/US/products/ps11764/products_data_sheets_list.html

Cisco Jabber Instant Messaging and Presence Deployments

Instant messaging and presence services for Jabber clients can be provided through the Cisco Client Services Framework XMPP interface. Cisco offers instant messaging (IM) and presence services with the following products:

- Cisco IM and Presence, page 24-18
- Cisco WebEx Messenger, page 24-20

The following sections discuss the architecture and design considerations for Cisco IM and Presence and WebEx Messenger service.

Cisco IM and Presence

The main component of the Cisco IM and Presence solution is the Cisco IM and Presence Service, which incorporates the Jabber Extensible Communications Platform and supports SIP/SIMPLE and Extensible Messaging and Presence Protocol (XMPP) for collecting information regarding a user's availability status and communications capabilities. The user's availability status indicates whether or not the user is actively using a particular communications device. The user's communications capabilities indicate the types of communications that user is capable of using, such as video conferencing, web collaboration, instant messaging, or basic audio. The architecture of Cisco IM and Presence and the design guidance for deployments are discussed in detail in the chapter on Cisco IM and Presence, page 23-1.

The following sections discuss aspects of Cisco IM and Presence design that are relevant to Jabber clients using a Cisco IM and Presence cluster.

Client Scalability

The Cisco IM and Presence Service hardware deployment determines the number of users a cluster can support. Cisco Jabber client deployments must balance all users equally across all servers in the cluster. This can be done automatically by setting the User Assignment Mode Sync Agent service parameter to **balanced**.

High Availability for Jabber Clients

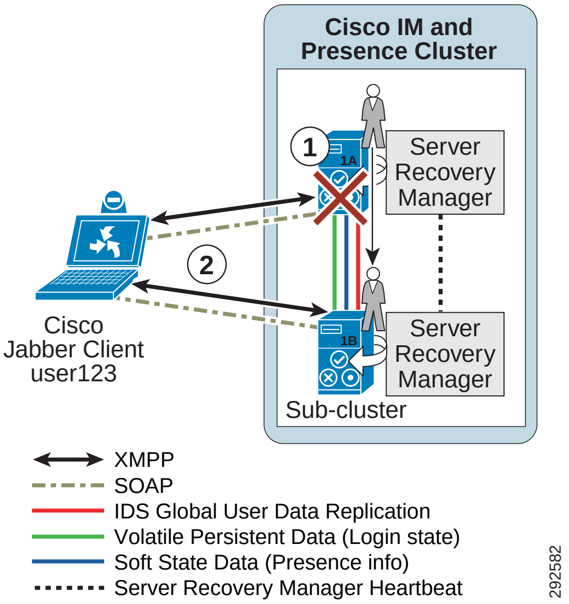
All users in the Cisco IM and Presence cluster must be assigned to a server prior to any exchange of information. By default, Cisco IM and Presence allows for automatic user assignment that is equally balanced across all servers in the cluster. If desired, the administrator can control where users are assigned by setting the User Assignment Mode Sync Agent service parameter to **None** instead of the default **balanced**. If this parameter is set to **None**, user assignment is done from the **System** > **Topology** menu.

Cisco Jabber clients can be provisioned with a basic deployment, a highly available deployment for automatic redundancy, and an IM and presence only deployment. In a Cisco IM and Presence two-server subcluster, users associated with one server are known by the other server in the subcluster, thus allowing for automatic failover when service communication with the configured server is interrupted. Cisco Jabber client high availability is supported only within a Cisco IM and Presence subcluster.

As illustrated in Figure 24-5, the server recovery manager monitors the various services on Cisco IM and Presence to determine if a service has failed and then to initiate an XMPP failover event. The following sequence of events occurs during an XMPP failover:

1. When the server recovery manager determines that a service is no longer communicating, a failover user move operation from server 1A to server 1B is initiated. User123 is moved from home server 1A and is now homed to server 1B.
2. The Cisco Jabber client determines that connectivity with server 1A is lost through time-out, connection loss, or XMPP protocol update, and it initiates a new connection to server 1B.

Figure 24-5 Cisco Jabber Client XMPP Failover


Cisco IM and Presence Cluster
1
1A
Server Recovery Manager
2
1B
Server Recovery Manager
Sub-cluster
Cisco Jabber Client user123
XMPP
SOAP
IDS Global User Data Replication
Volatile Persistent Data (Login state)
Soft State Data (Presence info)
Server Recovery Manager Heartbeat
292582

Cisco WebEx Messenger

Cisco WebEx Messenger is a multi-tenant software-as-a-service (SaaS) platform for synchronous and asynchronous collaboration. The WebEx Messenger platform is hosted inside the Cisco WebEx Collaboration Cloud and it enables collaborative applications and integrations, which allows for organizations and end users to customize their work environments. For additional information on the Cisco WebEx Messenger platform, refer to the documentation available at

http://developer.cisco.com/web/webex-developer

For more information on the Cisco Collaboration Cloud, refer to the documentation available at

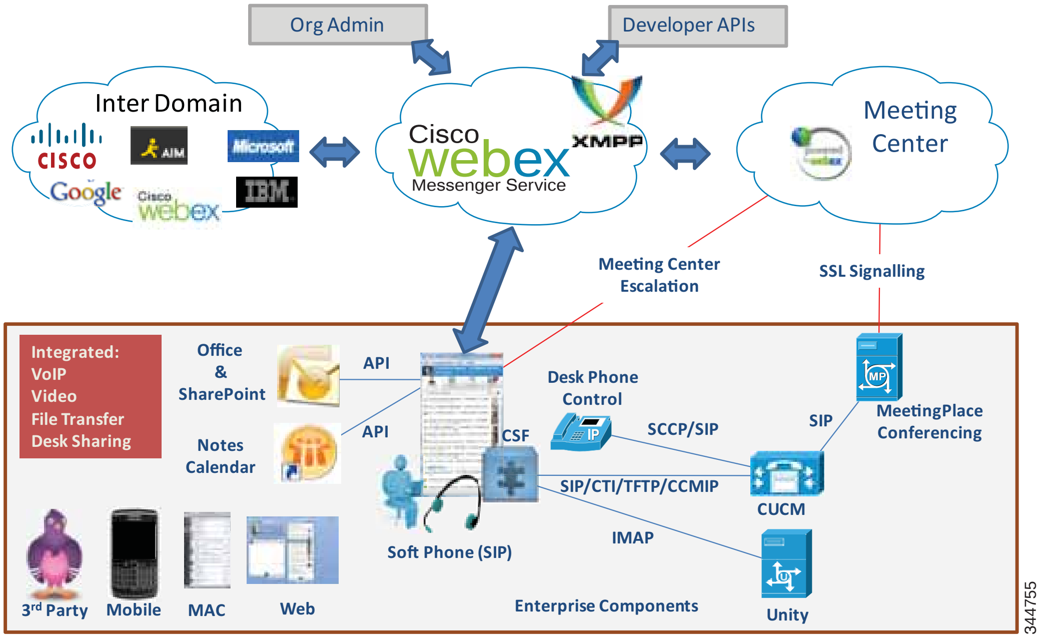
http://www.cisco.com/en/US/solutions/ns1007/collaboration_cloud.html

Deploying Cisco WebEx Messenger Service

A Cisco WebEx Messenger solution deployment consists of the following components, as depicted in Figure 24-6:

- A secure connection (SSL and AES) to the Cisco WebEx Messenger XMPP cloud platform for presence, instant messaging, VoIP, PC-to-PC video, media transfer (screen capture and file transfer), and desktop sharing
- Cisco WebEx Meetings
- XMPP federation with other WebEx Messenger organizations and third-party XMPP clients and XMPP instant messaging (IM) networks
- Cisco Unified Communications integration for call control, voice messaging, and call history
- Microsoft Outlook and IBM Lotus Notes calendar integration
- Integration to Microsoft Outlook for presence and click-to-communicate functionality

Figure 24-6 Deploying Cisco WebEx Messenger Service



Centralized Management

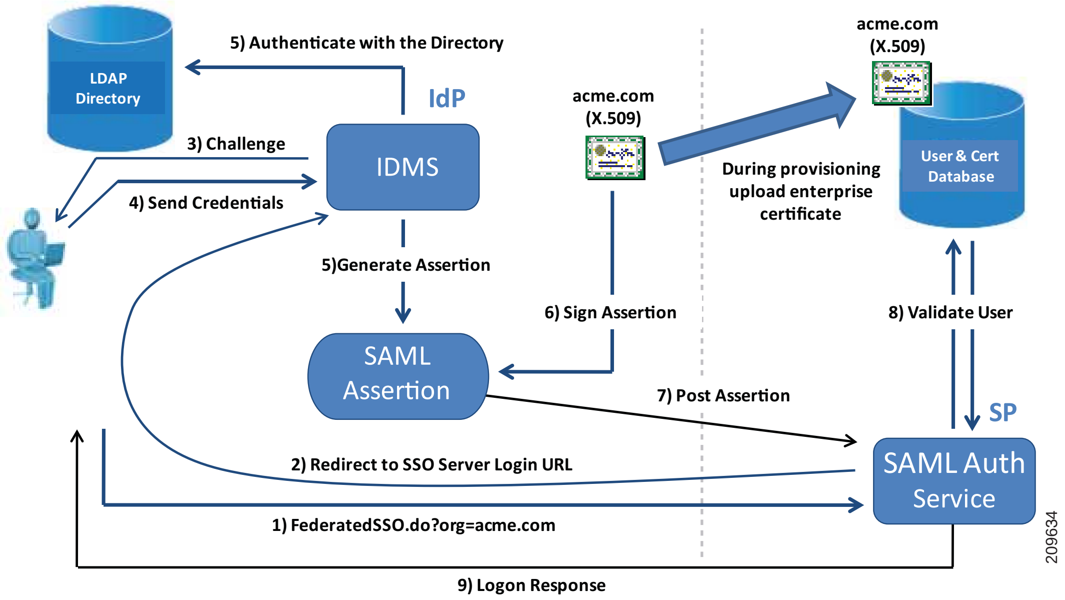
Cisco WebEx Messenger service provides a web-based administrative tool to manage the solution across the organization. Cisco WebEx Messenger service users are configured and managed through the Cisco WebEx Administration Tool, which enables administrators to set up basic security and policy controls for features and services. These policies can be applied enterprise-wide, by group, or individually. There are various methods to provision the user database that are further described in the Cisco WebEx administrator's guide available at

http://www.webex.com/webexconnect/orgadmin/help/index.htm

Single Sign On

Single Sign On (SSO) enables companies to use their on-premises SSO system, including Security Assertion Markup Language (SAML) support, to simplify the management of Cisco WebEx Messenger by allowing users to securely log into Cisco WebEx Messenger service using their corporate login credentials. The user's login credentials are not sent to Cisco, thus protecting the user's corporate login information. Figure 24-7 shows the credential handshake that occurs on user login to Cisco WebEx Connect.

Figure 24-7 User Login Authentication Process for Cisco WebEx Messenger Service



A user account can be configured to be created automatically the first time a user logs into Cisco IM client. Users are prevented from accessing the Cisco WebEx Messenger service if their corporate login account is deactivated.

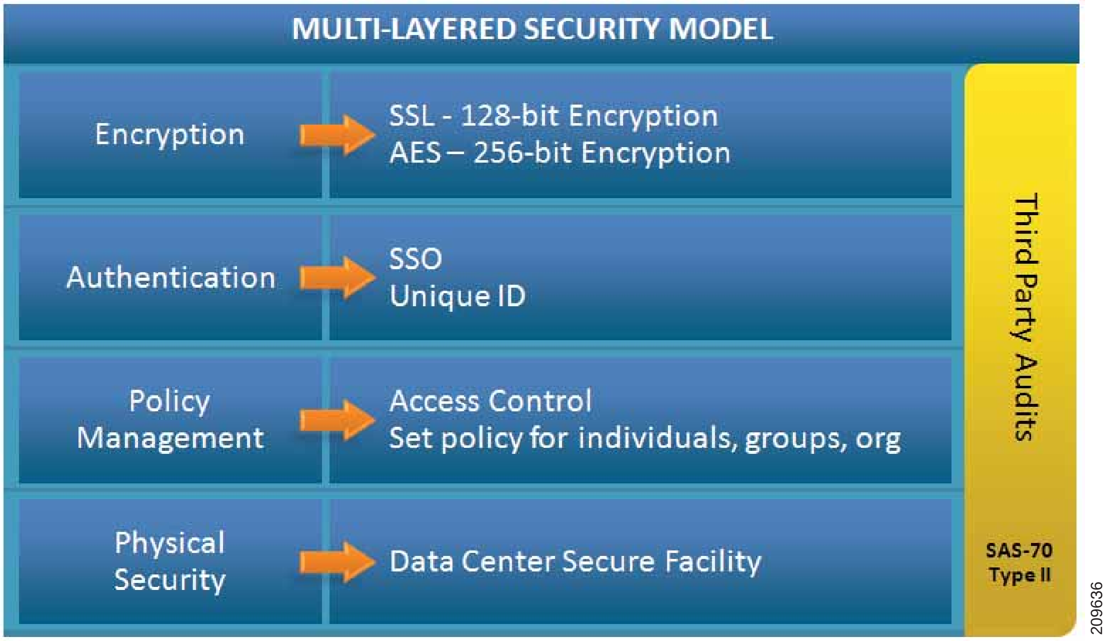
For more information on Single Sign On with WebEx Messenger service, refer to the documentation available at

http://developer.cisco.com/web/webex-developer/sso-reference

Security

The Cisco WebEx security model consists of functional layers of security. Figure 24-8 illustrates the separate but interrelated elements that compose each layer.

Figure 24-8 WebEx Security Model



The bottom layer represents the physical security in the Cisco WebEx data centers. All employees go through an extensive background check and must provide dual-factor authentication to enter the datacenter.

The next level is policy management, where the WebEx Messenger organization administrator can set and manage access control levels by setting different policies for individual users, groups, or the entire Cisco WebEx Messenger organization. White-list policies, specific to external users or domains, can be created to allow instant messaging exchanges. The Cisco WebEx Messenger organizational model also allows for the creation of specific roles and groups across the entire user base, which allows the administrator to assign certain privileges to roles or groups as well as to set policies, including access control, for the entire organization.

Access to the Cisco WebEx Messenger service is controlled at the authentication layer. Every user has a unique login and password. Passwords are never stored or sent over email in clear text. Passwords can be changed only by the end-users themselves. The administrator can choose to reset a password, forcing the end-user to change his or her password upon the next login. Alternatively, an administrator may choose to use the Single Sign On (SSO) integration between Cisco WebEx Messenger service and the company's directory to simplify end-user access management. The Single Sign On integration is achieved through the use of an Identity Management System (IDMS).

The encryption layer ensures that all instant messaging communications between Cisco WebEx Messenger users is encrypted. All instant messaging communication between Cisco WebEx Messenger users and the server in the Messenger Collaboration cloud is encrypted by default using SSL encryption. An additional level of security is available whereby IM communication can be encrypted end-to-end using 256-bit AES level encryption.

The Cisco WebEx Messenger platform uses third-party audits such as the SAS70 Type II audit to provide customers with an independent semi-annual security report. This report can be reviewed by any customer upon request with the Cisco Security organization. For additional Cisco WebEx Messenger service security, refer to the C*isco WebEx Connect Security White Paper*, available at

http://www.cisco.com/en/US/products/ps10528/prod_white_papers_list.html

Firewall Domain White List

Access control lists should be set specifically to allow all communications from the webex.com and webexconnect.com domains and all sub-domains for both webex.com and webexconnect.com. The WebEx Messenger platform sends email to end-users for username and password communications. These email messages come from the mda.webex.com domain.

Logging Instant Messages

Cisco WebEx Messenger service instant messaging communications are logged on the local hard drive of the personal computer where the user is logged in. Instant message logging is a capability in Cisco WebEx Messenger service that can be enabled by means of policy through the Org Admin tool.

The end-user can set logging specifics, whether to enable or disable logging, and how long the logs are kept. These message history settings are located under General in the IM client preferences.

Customers looking for advanced auditing and e-discovery capabilities should consider third-party solutions. Currently Cisco does not provide support for advanced auditing of instant messaging communications. Cisco WebEx Messenger service, however, does allow for logging and archiving of instant messages exchanged between users. Archiving of the logs is possible though the use of third-party SaaS archiving services, or the logs can be delivered securely to an on-premises SMTP server.

For additional information on instant message archiving, refer to the Cisco WebEx administrator's guide available at

http://www.webex.com/webexconnect/orgadmin/help/index.htm

Capacity Planning for Cisco WebEx Messenger Service

A single end-user requires only a 56 kbps dial-up Internet connection to be able to log in to WebEx Messenger service and get the basic capabilities such as presence, instant messaging, and VoIP calling. However, for a small office or branch office, a broadband connection with a minimum of 512 kbps is required in order to use the advanced features such as file transfer, screen capture, PC-to-PC video calling, and team spaces. For higher quality video such as High Definition 720p, the minimum bandwidth connection recommendation is 2 Mbps.

For additional information on network and desktop requirements, refer to the Cisco WebEx administrator's guide available at

http://www.webex.com/webexconnect/orgadmin/help/index.htm

Cisco Webex Messenger deployment network requirements are available at

http://www.webex.com/webexconnect/orgadmin/help/17161.htm

High Availability for Cisco WebEx Messenger Service

WebEx Messenger is a Software-as-a-Service (SaaS) application. The end-user device must be connected to the Internet for the end user to log in to the IM client. A standard Internet connection is all that is required. If an end user is remote, it is not necessary for that user to be connected through the company VPN in order to log in to the WebEx Messenger service. Cisco WebEx Messenger service IM clients can be deployed in a highly available redundant topology. Deployment of the Cisco WebEx Messenger Software-as-a-Service architecture consists of various network and desktop requirements described in this section.

High Availability

With the use of the multi-tenant Software-as-a-Service architecture, if any individual server in a group fails for any reason, requests can be rerouted to another available server in the Cisco WebEx Messenger Platform.

The Cisco WebEx Network Operations Team provides 24x7 active monitoring of the Cisco WebEx Collaboration Cloud from the Cisco WebEx Network Operations Center (NOC). For a comprehensive overview of the Cisco WebEx technology, refer to the information at

http://www.cisco.com/en/US/solutions/ns1007/collaboration_cloud.html

Redundancy, Failover, and Disaster Recovery

The Cisco WebEx Global Site Backup architecture handles power outages, natural disaster outages, service capacity overload, network capacity overload, and other types of service interruptions. Global Site Backup supports both manual and automatic failover. The manual failover mode is typically used during maintenance windows. The automatic failover mode is used in case of real-time failover due to a service interruption.

Global Site Backup is automatic and transparent to the end users, it is available for all users, and it imposes no limits on the number of users that can fail-over.

Global Site Backup consists of the following main components:

- Global Site Service — Is responsible for monitoring and switching traffic at the network level.
- Database Replication — Ensures that the data transactions occurring on the primary site are transferred to the backup site.
- File Replication — Ensures that any file changes are maintained in synchronization between the primary and the backup site.

Design Considerations for Cisco WebEx Messenger Service

Cisco WebEx Messenger is deployed as a Software-as-a-Service model, therefore design and deployment considerations are minimal. The Cisco WebEx Messenger solution has client options available for the Windows and Mac desktop as well as the popular mobile devices.

Third-Party XMPP Clients Connecting to Cisco WebEx Messenger Platform

Although Cisco does not officially support any other XMPP clients to connect to the Cisco WebEx Messenger Platform, the nature of the XMPP protocol is to allow end users to connect to presence clouds with various XMPP clients using their WebEx Messenger service credentials. A list of XMPP software clients is available at

http://xmpp.org/software/clients.shtml

Organization policies cannot be enforced on third-party XMPP clients, and features such as end-to-end encryption, desktop share, video calls, PC-to-PC calls, and teleconferences are not supported with third-party clients. To allow non-WebEx Messenger service XMPP IM clients to authenticate to your WebEx Messenger service domain(s), DNS SRV records must be updated. The specific DNS SRV entry can be found in Cisco WebEx administration, under Configuration and IM Federation.

The use of non-Messenger service XMPP clients in Cisco WebEx administration, under Configuration and XMPP IM Clients, must be explicitly allowed.

For additional information on enabling third-party XMPP clients to connect to the WebEx Messenger platform, refer to the Cisco WebEx administrator's guide available at

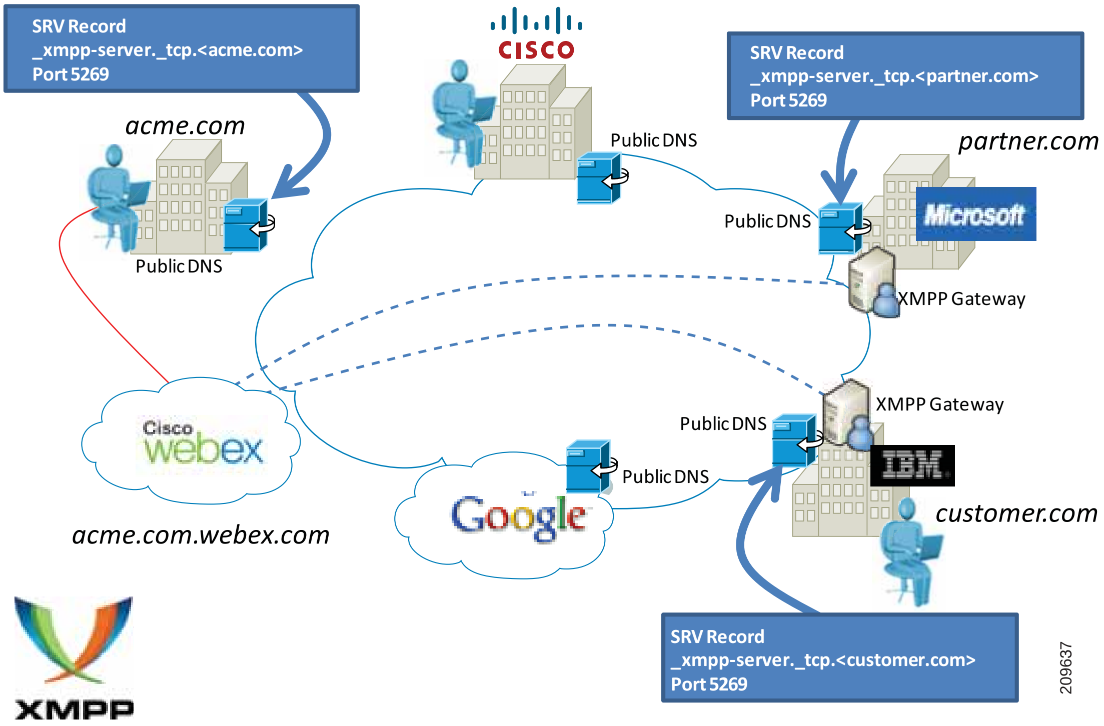
http://www.webex.com/webexconnect/orgadmin/help/index.htm

Instant Messaging and Presence Federation Using Third-Party XMPP Clients

The Cisco WebEx Messenger service network can federate with XMPP-based instant messaging networks such as GoogleTalk and Jabber.org. (See Figure 24-9.) A list of public instant messaging networks based on XMPP is available at

http://xmpp.org/

Figure 24-9 Inter-Domain Federation



Currently the WebEx Messenger service does not interoperate with Yahoo! Messenger and Windows Live Messenger, but it can federate with AIM through a federation gateway.

Other Resources and Documentation

The Cisco WebEx administrator's guide is available at

http://www.webex.com/webexconnect/orgadmin/help/index.htm

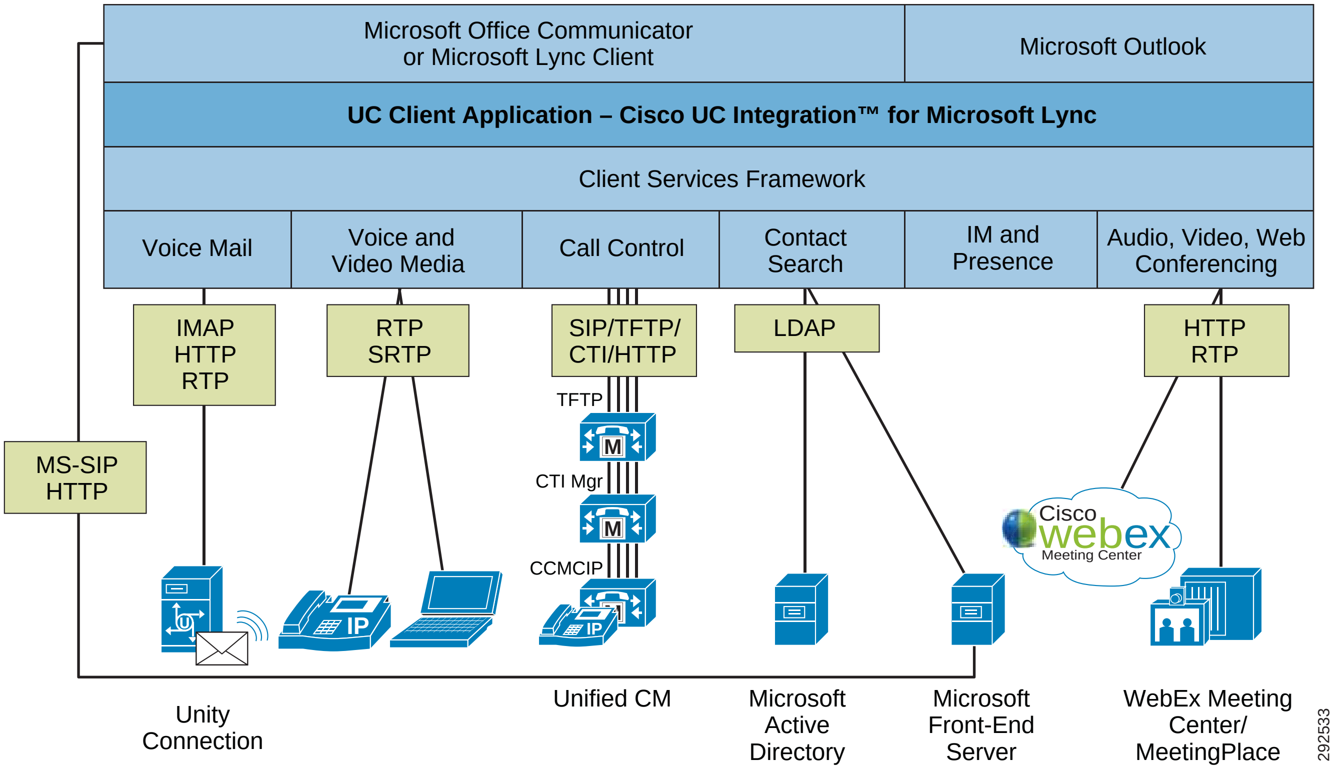
Cisco UC Integration^{TM} for Microsoft Lync Architecture

Cisco UC Integration^{TM} for Microsoft Lync clients support a variation of the on-premises deployment models, where IM and presence services are provided by Microsoft Applications instead of Cisco IM and Presence.

Cisco UC Integration^{TM} for Microsoft Lync allows for tightly integrated Cisco Unified Communications services for Microsoft Lync using the Cisco Unified Client Services Framework, while delivering a consistent user experience. The solution extends the presence and instant messaging capabilities of Microsoft Lync by providing access to a broad set of Cisco Unified Communications services, including standards-based audio and video, unified messaging, web conferencing, deskphone control, and telephony presence.

The solution architecture for a Cisco UC Integration^{TM} for Microsoft Lync deployment, shown in Figure 24-10, includes Cisco Unified Communications Manager for audio and video services, Microsoft Office Communications Server 2007 for presence and instant messaging services, Microsoft Active Directory for user account information, Cisco Unified Client Services Framework for PC audio or deskphone control, and Microsoft Lync.

Figure 24-10 Cisco UC Integration^{TM} for Microsoft Lync


Microsoft Office Communicator
or Microsoft Lync Client
Microsoft Outlook
UC Client Application – Cisco UC Integration™ for Microsoft Lync
Client Services Framework
Voice Mail
Voice and Video Media
Call Control
Contact Search
IM and Presence
Audio, Video, Web Conferencing
IMAP HTTP RTP
RTP SRTP
SIP/TFTP/ CTI/HTTP
LDAP
HTTP RTP
MS-SIP HTTP
TFTP
CTI Mgr
CCMCIP
M
M
IP
IP
Cisco webex Meeting Center
Unity Connection
Unified CM
Microsoft Active Directory
Microsoft Front-End Server
WebEx Meeting Center/ MeetingPlace
292533

With a deployment of Cisco UC Integration^{TM} for Microsoft Lync, the client utilizes user information from the Office Communications Server Address Book that gets downloaded to the client. The address book is generated and delivered to the clients from the Office Communications Server once the user is enabled for presence and instant messaging. Cisco recommends that administrators populate the user directory number information with an E.164 value (for example, +18005551212) and enable LDAP synchronization and authentication on Unified CM for user account consistency. Cisco UC Integration^{TM} for Microsoft Lync connects to both Cisco Unified CM and Microsoft Active Directory and provides for account credential synchronization rules.

Deploying Cisco UC Integration^{TM} for Microsoft Lync

When deploying Cisco UC Integration^{TM} for Microsoft Lync, observe the guidelines presented in this section.

Configuration Settings

Cisco UC Integration^{TM} for Microsoft Lync reads its configuration settings from a series of registry entries that the administrator must configure. Cisco recommends pushing these registry configuration settings from Microsoft Active Directory by means of Group Policy to distribute the configuration settings automatically to the client computer. Although Group Policy is the recommended installation mechanism, there are other methods available as well, including third-party software deployment tools, batch files, Vbscript, or manual configuration.

Microsoft Active Directory group policies can be extended using administration templates, and Cisco UC Integration^{TM} for Microsoft Lync provides an administrative template that the administrator can add to provide the group policy support. After the administrative template is loaded, a Cisco UC Integration^{TM} for Microsoft Lync configuration policy can be created by the administrator for the registry configuration settings (TFTP servers, CTI servers, CCMCIP servers, voicemail, and LDAP servers). The registry location where these settings are stored is:

HKCU\Software\Policies\Cisco Systems, inc.\Client Services Framework\AdminData

The Group Policy Management Console can be used to control how and where these group policies are applied to different organizational units. From a client policy perspective, when you deploy Cisco UC Integration^{TM} for Microsoft Lync, Cisco recommends setting the Microsoft Telephony Mode Policy to **IM and Presence Only** and **DisableAVConferencing**. These client policy changes will allow for only a single set of call options to be displayed in the Microsoft Lync user experience.

A Cisco UC Integration^{TM} for Microsoft Lync deployment also allows for custom presence states to be defined and deployed in the cisco-presence-states-config.xml file that gets installed. However, Cisco recommends that administrators relocate this file to an HTTPs location, such as the Microsoft Office Communications Server, to allow Microsoft Lync to use this custom presence state file based on the following registry location:

HKLM\Software\Policies\Microsoft\Communicator\CustomStateURL

Software Installation

The software installation can be handled a number of different ways and is designed to be deployed using desktop management tools such as Microsoft Active Directory Group Policy, Systems Management Server (SMS), Altiris, or self-extracting executable with script/batch file. Because customer topologies vary, there is no recommendation about which method to use. For details on the software deployment method, refer to the Cisco UC Integration^{TM} for Microsoft Lync documentation available at

http://www.cisco.com/en/US/products/ps10317/index.html

Capacity Planning for Cisco UC Integration^{TM} for Microsoft Lync

Cisco UC Integration^{TM} for Microsoft Lync uses Unified CM CTIManager for click-to-dial applications, as well as deskphone control mode with the Cisco Unified Client Services Framework. Therefore, observe the CTI limits as defined in the chapter on Call Processing, page 8-1. When Cisco UC Integration^{TM} for Microsoft Lync is operating in softphone (audio on computer) mode, the Cisco Unified Client Services Framework is a SIP registered endpoint with Cisco Unified CM. When sizing a solution involving Cisco Unified Communications, you must include the CTI devices and the SIP endpoint devices utilizing resources on the Unified CM clusters.

High Availability for Cisco UC Integration^{TM} for Microsoft Lync

Cisco Unified Client Services Framework provides primary and secondary servers for each of the configuration components, TFTP server, CTIManager, CCMCIP server, voicemail server, and LDAP server. When operating in softphone (audio on computer) mode, the Client Services Framework is a SIP registered endpoint with Cisco Unified CM, and it supports all of the registration and redundancy capabilities of a registered endpoint of Unified CM. When operating in deskphone mode, the Client Services Framework is controlling a Cisco Unified IP Phone using CTI, and it supports configuration of a primary and secondary CTIManager. For additional details on CTI deployments, refer to the chapter on Call Processing, page 8-1. The Client Services Framework does not rely on Microsoft Lync being online to support high availability.

Microsoft Lync provides primary and secondary servers with the configuration of enterprise pools for an Office Communications Server deployment. For additional details, refer to the Microsoft Office Communications Server 2007 deployment documentation available at

http://technet.microsoft.com/en-us/library/dd425168%28office.13%29.aspx

Design Considerations for Cisco UC Integration™ for Microsoft Lync

Observe the following design considerations when deploying Cisco UC Integration™ for Microsoft Lync:

- The administrator must determine how to install, deploy, and configure Cisco UC Integration™ for Microsoft Lync in their organization. Cisco recommends using a well known installation package such as Altiris to install the application, and use Group Policies to configure the user registry settings for the required components of TFTP server, CTIManager, CCMCIP server, voicemail pilot, LDAP server, LDAP domain name, and LDAP search contexts.
- Cisco UC Integration™ for Microsoft Lync connects to both Cisco Unified CM and Microsoft Active Directory; therefore, Cisco recommends enabling LDAP synchronization and LDAP authentication on Unified CM to allow for integration of the Unified Communications and back-end directory components.
- The address book generated by the Microsoft Office Communications Server and distributed to the clients is used by Cisco UC Integration™ for Microsoft Lync to initiate voice and video calls. Before enabling the user for Microsoft Office Communications Server instant messaging and presence, Cisco recommends configuring the user with an E.164 telephone number in Microsoft Active Directory.

Cisco Virtualization Experience Client Architecture

Cisco Virtualization Experience Client (VXC) endpoints enable the end users to have secure real-time access to content and business applications as well as a rich collaborative user experience. These endpoints provide access to collaboration services that are part of the larger Cisco Virtualization Experience Infrastructure (VXI) solution. For information on a complete end-to-end VXI solution design, refer to the documentation available at

http://www.cisco.com/en/US/solutions/ns340/ns414/ns742/ns1100/landing_vxi.html

Deploying Cisco Virtualization Experience Clients

Cisco Virtualization Experience Clients (VXC) are endpoints in the Cisco Unified Communications portfolio; however, they are more than just simple endpoints because they interact with a user's working environment by providing voice, video, and virtual desktop capability and functionality. A user's work environment can assume various profiles (for example, task worker, knowledge worker, or mobile worker), and Cisco has various Virtualization Experience Clients to meet those different needs.

The Cisco VXC 2111 and VXC 2112 provide an integrated form factor that is paired with a Cisco Unified IP Phone 8961, 9951, or 9971 for a fully integrated voice, video, and virtual desktop environment. The Cisco VXC 2211 and VXC 2212 provide a standalone form factor that can be used as simply a virtual desktop (for a task worker), or they can be paired with an IP phone for a fully integrated user environment. The Cisco VXC 4000 provides a software-only solution by using a re-purposed PC to provide the user with voice and virtual desktop functionality, while the Cisco VXC 6215 provides a Linux-based thin client for fully integrated voice, video, and virtual desktop in a single device.

Cisco Virtualization Experience Client Manager

Cisco Virtualization Experience Client (VXC) Manager is a critical and mandatory component of any Virtualization Experience Client deployment. Cisco VXC Manager uses industry standard protocols to manage network intelligent devices simply, efficiently, remotely, and securely using a component-based architecture. As a required component of any VXC deployment, VXC Manager is used to easily manage, organize, upgrade, control, and support various Cisco VXC devices running Independent Computing Architecture (ICA) or PC over IP (PCoIP) protocol.

Note Cisco VXC 4000 is installed on Microsoft Windows only, thus is not managed by VXC Manager. The VXC 4000 Windows installer can be deployed using any common software deployment utility.

Power Over Ethernet

The Cisco Virtualization Experience Client 2111 and 2112 integrated form factor receives power from the spine connector on the unit, which attaches to the Key Expansion port on the Cisco Unified IP Phone 8961, 9951, and 9971. Power to the Cisco Unified IP Phone 8961, 9951, and 9971 is provided through a PWR-CUBE-4 or through 802.3at inline power.

The Cisco Virtualization Experience Client 2211 and 2212 standalone form factor receives power from one of three sources: the PWR-CUBE-4, 802.3at inline power, or 802.3af inline power.

The Cisco Virtualization Experience Client 4000 and 6215 do not support inline power over Ethernet.

Network Considerations (Call Admission Control, Quality of Service, and Bandwidth)

Cisco VXC zero clients (VXC 2111, 2112, 2211, and 2212) provide a virtual desktop environment through display protocol interaction between the zero client and the connection broker datacenter back end. Quality of Service (QoS) is best-effort, and the Cisco VXC zero clients should be placed in the data VLAN. Display protocols inherently use as much bandwidth as a link provides; therefore, bandwidth controls can be put in place at the network port level, or they can be configured through the back-end Citrix or VMware connection broker settings. When a Cisco Unified IP Phone is paired with a VXC zero client, follow existing Unified Communications call admission control, QoS, and bandwidth guidelines.

Cisco VXC 4000 is a software-only solution that uses applications running locally on the PC for a fully integrated solution that includes a thick Virtual Desktop Infrastructure (VDI) client (Citrix Receiver 3.0 or VMware View Client 5.0) and the VXC 4000 software application. With the VXC 4000, QoS is best-effort, and the VXC 4000 should be placed in the data VLAN because all the traffic (voice and virtual desktop) will originate from the local PC resource.

The Cisco VXC 6215 thin client provides a fully integrated software appliance running locally on the device, and it provides display protocol interaction through standard APIs with the hosted virtual desktop environment when used in a fully integrated Unified Communications deployment. The VXC 6215 can operate as a VDI-only endpoint, similar to a Cisco VXC zero client deployment, or it can operate in a fully integrated voice, video, virtual desktop deployment. In both deployments, QoS is best-effort, and the Cisco VXC 6215 should be placed in the data VLAN. Call admission control for voice and video follow existing Cisco Unified IP Phone guidelines, and bandwidth controls for the virtual desktop are provided through the connection broker settings.

Capacity Planning for Cisco Virtualization Experience Clients

All Cisco Virtualization Experience Clients are deployed with a Virtual Desktop Infrastructure (VDI) component, while some of the deployments may also contain a Unified Communications component. Capacity planning and datacenter resource utilization for VDI when using the Cisco Virtualization Experience Clients is covered as part of the Virtualization Experience Infrastructure (VXI) sizing. For details, refer to the VXI documentation available at

http://www.cisco.com/en/US/solutions/ns340/ns414/ns742/ns1100/landing_vxi.html

Capacity planning for the Unified Communications components depends on which Virtualization Experience Client is deployed:

- Cisco VXC 2111 and 2112 integrated form factor zero clients are paired with a Cisco Unified IP Phone 8961, 9951, or 9971. The Cisco client running in the user's virtual desktop uses the deskphone control mode of the Cisco Unified IP Phone; therefore, Computer Telephony Integration (CTI) planning guidelines must be followed for each client deployed.
- Cisco VXC 2211 and 2212 standalone form factor zero clients can be deployed as VDI-only or as a fully integrated voice, video, and virtual desktop with a number of different Cisco Unified IP Phones. When deployed in a Unified Communications environment, the Cisco client running in the user's virtual desktop uses the deskphone control mode of the Cisco Unified IP Phone; therefore, CTI planning guidelines must be followed for each client deployed.
- Cisco VXC 4000 software appliance is a software-only VXC deployment option. The Cisco client running in the user's virtual desktop uses the deskphone control mode of the VXC 4000; therefore, CTI planning guidelines must be followed for each VXC 4000 deployed.
- Cisco VXC 6215 thin client running in VDI-only mode follows VDI capacity planning; however, when the VXC 6215 is deployed as a fully integrated voice, video, and virtual desktop, additional Unified Communications capacity must be accounted for. The Cisco client running in the user's virtual desktop uses the deskphone control mode of the VXC software appliance running locally on the Linux thin client; therefore, CTI planning guidelines must be followed for each client deployed. The VXC software appliance is a SIP line-side registered device on Cisco Unified CM; therefore, for each VXC 6215 thin client running as a fully integrated voice, video, and virtual desktop, a SIP line device and CTI connection is used.

High Availability for Cisco Virtualization Experience Clients

A Cisco Virtualization Experience Client deployment has several aspects that involve high availability: the Virtual Desktop Infrastructure (VDI), the Cisco client running within the hosted virtual desktop (HVD), and the Unified Communications endpoint registered to Unified CM. A user's desktop virtualization environment can be deployed according to Citrix or VMware high availability guidelines. The Cisco client running within the user's virtual desktop supports high availability according to the guidelines listed for Cisco UC Integration^{TM} for Microsoft Lync (see High Availability for Cisco UC Integration^{TM} for Microsoft Lync, page 24-28). The Unified Communications endpoint registered to Unified CM can be either a Cisco Unified IP Phone when using the Cisco VXC 2111, 2112, 2211, and 2212 zero clients, or the VXC software appliance if using the Cisco VXC 4000 or 6215. These Unified CM registered endpoints support failover for the devices as part of their call control group assignment.

Note CTI failover is not supported with Cisco Virtualization Experience Clients. Survivable Remote Site Telephony (SRST) is supported with the Cisco Unified IP Phones, but SRST is not supported with the VXC software appliance.

Design Considerations for Cisco Virtualization Experience Clients

The following design considerations apply to the Cisco Virtualization Experience Clients:

- Cisco VXC Manager is a required component to manage, configure, and upgrade Cisco Virtualization Experience Clients.
- Cisco Virtualization Experience Clients provide end-user access as part of the larger Cisco Virtualization Experience Infrastructure deployment. Cisco VXI end-to-end solution deployment design guidance is tested and documented as a Cisco Validated Design.
- CTI guidelines must be observed when deploying Cisco Virtualization Experience Clients in a fully integrated voice, video, and virtual desktop environment.
- With the Cisco VXC Software Appliance, QoS is best-effort and the VXC 6215 should be placed in the data VLAN. For details on traffic marking, refer to the *Enterprise QoS Solution Reference Network Design Guide*, available at

  http://www.cisco.com/go/designzone

CHAPTER 25

Mobile Unified Communications

Revised: April 30, 2013; OL-27282-05

Mobile Unified Communications solutions and applications provide the ability to deliver features and functionality of the enterprise IP communications environment to mobile workers wherever they might be. With mobile Unified Communications solutions, mobile users can handle business calls on a multitude of devices and access enterprise applications whether moving around the office building, between office buildings, or between geographic locations outside the enterprise. Mobile Unified Communications solutions provide mobile workers with persistent reachability and improved productivity as they move between, and work at, a variety of locations.

Unified Communications mobility solutions can be divide into two main categories:

- Mobility within the enterprise

  This type of mobility is limited to movement of users within enterprise locations.

- Mobility beyond the enterprise

  This type of mobility refers to mobility beyond the enterprise infrastructure and typically involves some form of Internet, mobile voice network, and/or mobile data network traversal.

Mobility within the enterprise is limited to utilization within the network boundaries of the enterprise, whether those boundaries span only a single physical building, multiple physical buildings in close proximity or separated by long distances, or even home offices where network infrastructure is still controlled and managed by the enterprise when it is extended to the home office.

On the other hand, mobility beyond the enterprise involves a bridging of the enterprise infrastructure to the Internet or mobile provider infrastructures and finds users leveraging public and private networks for connectivity to enterprise services. In some cases the lines between these two types of mobility are somewhat blurred, especially in scenarios where mobile devices are connecting back to the enterprise for Unified Communications services over the Internet or mobile data and mobile voice networks.

Mobility within the enterprise can be divided into three main areas based on feature sets and solutions:

- Campus or single-site mobility

  With this type of enterprise mobility, users move around within a single physical location typically bounded by a single IP address space and PSTN egress/ingress boundary. This type of mobility involves operations and features such as phone movement from one physical network port to another, wireless LAN device roaming between wireless infrastructure access points, and even Cisco Extension Mobility (EM), where users temporarily apply their device profile including their enterprise number to a particular phone in a different area.

- Multisite mobility

  With this type of mobility, users move within the enterprise from one physical location to another, and this movement typically involves crossing IP address spaces as well as PSTN egress/ingress boundaries. This type of mobility involves the same types of operations and features as with campus mobility (physical hardware moves, WLAN roaming, and Cisco Extension Mobility) but replicated at each site within the enterprise. In addition, the Device Mobility feature can be leveraged to ensure that, as user's move devices between sites, phone calls are routed through the local site egress gateway, media codecs are negotiated appropriately, and call admission control mechanisms are aware of the device's location.

- Remote site mobility

  With this type of mobility, users move to a location outside the enterprise but still have some form of secure connection back to the enterprise, which virtually extends the enterprise network to the remote location. This type of mobility typically involves remote teleworker solutions such as Cisco Virtual Office as well as other remote connectivity methods such as VPN-based phones and the Office Extend Access Point feature.

Mobility beyond the enterprise can be divided into two high-level Cisco solution sets:

- Cisco Unified Mobility

  As part of Cisco Unified Communications Manager (Unified CM), the Cisco Unified Mobility feature suite offers the ability to associate a mobile user's enterprise number to their mobile or remote devices and provides connectivity between the user's fixed enterprise desk phone on the enterprise network and the user’s mobile device on the mobile voice provider network. This type of functionality is sometimes referred to as fixed mobile convergence.

- Cisco Mobile Client Solutions

  Cisco mobile client applications run on dual-mode smartphones and other mobile devices, and they provide access to enterprise voice and collaboration applications and services. Dual-mode phones provide dual radio antennas for connecting to both 802.11 wireless LAN networks and cellular voice and data networks. With a Cisco mobile client deployed on mobile devices, they can be registered to Cisco Unified CM through the enterprise wireless LAN or over the Internet through public or private Wi-Fi hot spots or the mobile data network, and they can in turn leverage the IP telephony infrastructure of the enterprise for making and receiving calls through voice over IP (VoIP). In the case of dual-mode phones, when mobile users are not associated to the enterprise WLAN or securely attached to the enterprise network with these devices, phone calls are made using the mobile voice provider network. In addition to enabling voice services for the mobile device, Cisco mobile clients also provide access to other collaboration services such as voice and instant messaging, presence, and enterprise directory access.

The various applications and features discussed in this chapter apply to all Cisco Unified Communications deployment models unless otherwise noted.

This chapter begins with a discussion of mobility features and solutions available within the enterprise infrastructure. It includes an examination of functionality and design considerations for campus or single-site deployments, multisite deployments, and even remote site deployments. This comprehensive set of solutions provides many benefits for mobile workers within the enterprise, including enterprise-class communications and improved productivity regardless of physical location. This discussion of mobility within the enterprise paves the way for examination of mobility solutions beyond the enterprise that leverage the mobile provider and Internet provider infrastructure and capabilities. These solutions enable a bridging of the enterprise network infrastructure and mobile functionality to the provider network infrastructure in order to leverage advanced mobile features and communication flows that can be built on the solid enterprise mobility infrastructure.

This chapter provides a comprehensive examination of mobility architectures, functionality, and design and deployment implications for enterprise Unified Communications mobility solutions. The analysis and discussions contained within this chapter are organized at a high level as follows:

- Mobility within the Enterprise
  - Campus Enterprise Mobility, page 25-4
  - Multisite Enterprise Mobility, page 25-11
  - Remote Enterprise Mobility, page 25-26
- Mobility beyond the Enterprise
  - Cisco Unified Mobility, page 25-32
  - Cisco Mobile Clients and Devices, page 25-60

What's New in This Chapter

Table 25-1 lists the topics that are new in this chapter or that have changed significantly from previous releases of this document.

Table 25-1 New or Changed Information Since the Previous Release of This Document

New or Revised Topic	Described in	Revision Date
A few minor updates	Various sections	April 30, 2013
Dial plan guidance for Device Mobility with local route group, and limitations of line/device approach	Dial Plan Design Considerations, page 25-21	September 28, 2012
Guidance and behavior regarding remote destination number configuration and caller ID matching	Remote Destination Configuration and Caller ID Matching, page 25-50	September 28, 2012
Cisco Unified Mobility dial plan guidance as it relates to Standard Local Route Group and Local Route Group	Dial Plan Considerations for Cisco Unified Mobility, page 25-53	September 28, 2012
Cisco mobile clients and devices architecture, including video capabilities and cloud-based integrations	Cisco Mobile Clients and Devices Architecture, page 25-61	September 28, 2012
Cisco Jabber for iPad mobile client	Cisco Jabber for iPad, page 25-76	September 28, 2012
Cisco Jabber mobile clients, including QoS and Bring Your Own Device (BYOD) considerations	Mobile Client and Device Quality of Service, page 25-63 Cisco Bring Your Own Device (BYOD) Infrastructure, page 25-70	September 28, 2012
Capacity information related to video calls per WLAN channel cell	Capacity Planning for Campus Enterprise Mobility, page 25-9	September 28, 2012
Detailed capacity planning information has been moved to the chapter on *Unified Communications Design and Deployment Sizing Considerations*.	Unified Communications Design and Deployment Sizing Considerations, page 29-1	June 28, 2012
Enhanced Single Enterprise Voicemail Box (mobile voicemail avoidance feature) — adding DTMF-based user answer confirmation to ensure that mobile voicemail is not connected	Mobile Voicemail Avoidance with Single Enterprise Voicemail Box, page 25-40	June 28, 2012

Table 25-1 New or Changed Information Since the Previous Release of This Document (continued)

New or Revised Topic	Described in	Revision Date
Ring All Shared Line, which provides enhancement to the Intelligent Session Control feature	Intelligent Session Control and Ring All Shared Lines, page 25-55	June 28, 2012
Nokia Call Connect, Cisco Unified Mobile Communicator, and Cisco Mobile 8.5 for Nokia have reached End-of-Sale (EoS) and are no longer covered in this chapter	Various sections removed from this chapter	June 28, 2012

Mobility Within the Enterprise

This section examines mobility features and solutions available within the enterprise. This examination includes discussions related to architecture, functionality, and design and deployment implications for the following types of enterprise mobility

- Campus Enterprise Mobility, page 25-4
- Multisite Enterprise Mobility, page 25-11
- Remote Enterprise Mobility, page 25-26

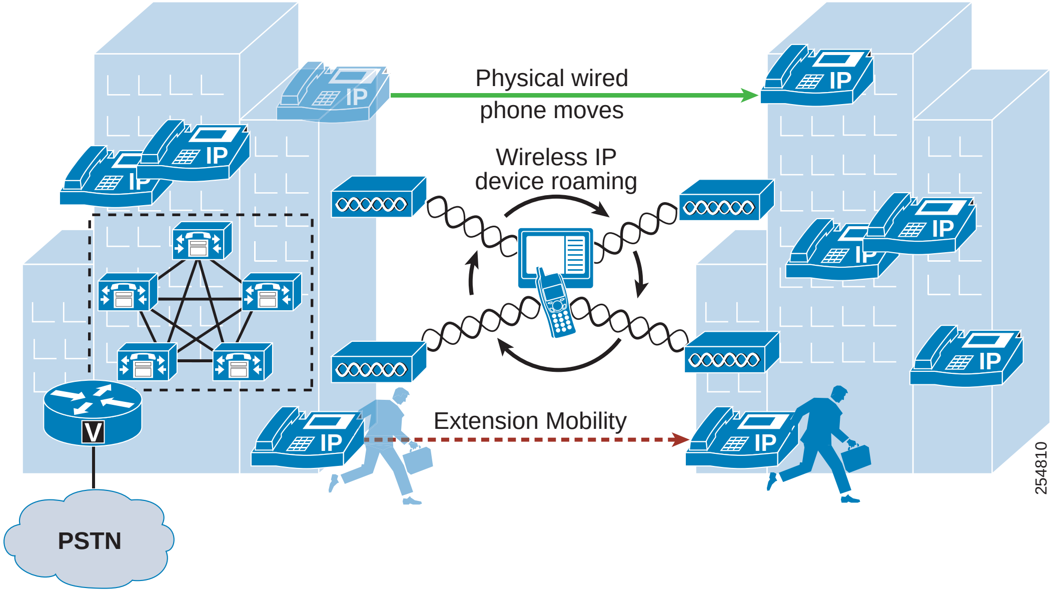
Campus Enterprise Mobility

Campus or single-site enterprise mobility refers to mobility within a single physical location typically bounded by a single IP address space and PSTN egress/ingress boundary. Mobility here not only includes the movement of users within this physical location but also the movement of endpoint devices.

Campus Enterprise Mobility Architecture

As illustrated in Figure 25-1, the enterprise campus mobility architecture is based on a single physical location that may include a single building or multiple buildings (as depicted) in close proximity, such that users are able to move freely within the campus and maintain IP and PSTN connectivity. Typically campus deployments involve a shared common connection or set of connections to the PSTN and Internet provider networks bound by a single IP address space and PSTN egress/ingress boundary. All users within this enterprise campus are connected to and reachable from a common network infrastructure.

Figure 25-1 Campus Enterprise Mobility Architecture



Types of Campus Mobility

Mobility within the campus enterprise typically involves the movement of devices, users, or both throughout the campus infrastructure. Campus enterprise mobility within Cisco Unified Communications deployments can be divided into three main categories: physical wired phone movement, wireless device movement, and user movement without phone hardware or software. Each of these types of movements are discussed below.

Physical Wired Device Moves

As shown in Figure 25-1, movement of physical wired phones is easily accommodated within the campus infrastructure. These types of phone movements can occur within a single floor of a building, across multiple floors of a building, or even between buildings within the campus. Unlike with traditional PBX deployments where physical phone ports are fixed to a particular office, cubicle, or other space within the building, in IP telephony deployments a phone can be plugged into any IP port within the network infrastructure in order to connect to the IP PBX.

In a Cisco environment, this means a user can simply unplug a Cisco Unified IP Phone from the network, pick it up and carry it to another location within the campus, and plug it into another wired network port. Once connected to the new network location, the phone simply re-registers to Unified CM and is able to make and receive calls just like in the previous location.

This same physical device movement also applies to software-based phones running on wired personal computers. For example, a user can move a laptop computer running Cisco IP Communicator or Cisco Jabber from one location to another within the campus, and after plugging the laptop into a network port in the new location, the software-based phone can re-register to Unified CM and begin to handle phone calls again.

To accommodate physical device mobility within the campus, care should be taken when physically moving phone devices or computers running software-based phones to ensure that the network connection used at a new location has the same type of IP connectivity, connection speed, quality of service, security, and network services such as in-line power and dynamic host control protocol (DHCP), as were provided by the previous location. Failure to replicate these connection parameters, services, and features will lead to reduced functionality or in some cases complete loss of functionality.

Wireless Device Roaming

Wireless devices can move or roam throughout the enterprise campus, as shown in Figure 25-1, provided a wireless LAN network has been deployed to provide wireless network connectivity to the campus edge.

Examples of wireless devices include Cisco Unified Wireless IP Phone 7925G, wirelessly attached Cisco Unified IP Phone 9971, Cisco Cius, and Cisco mobile clients such as Cisco Jabber (see Cisco Mobile Clients and Devices, page 25-60).

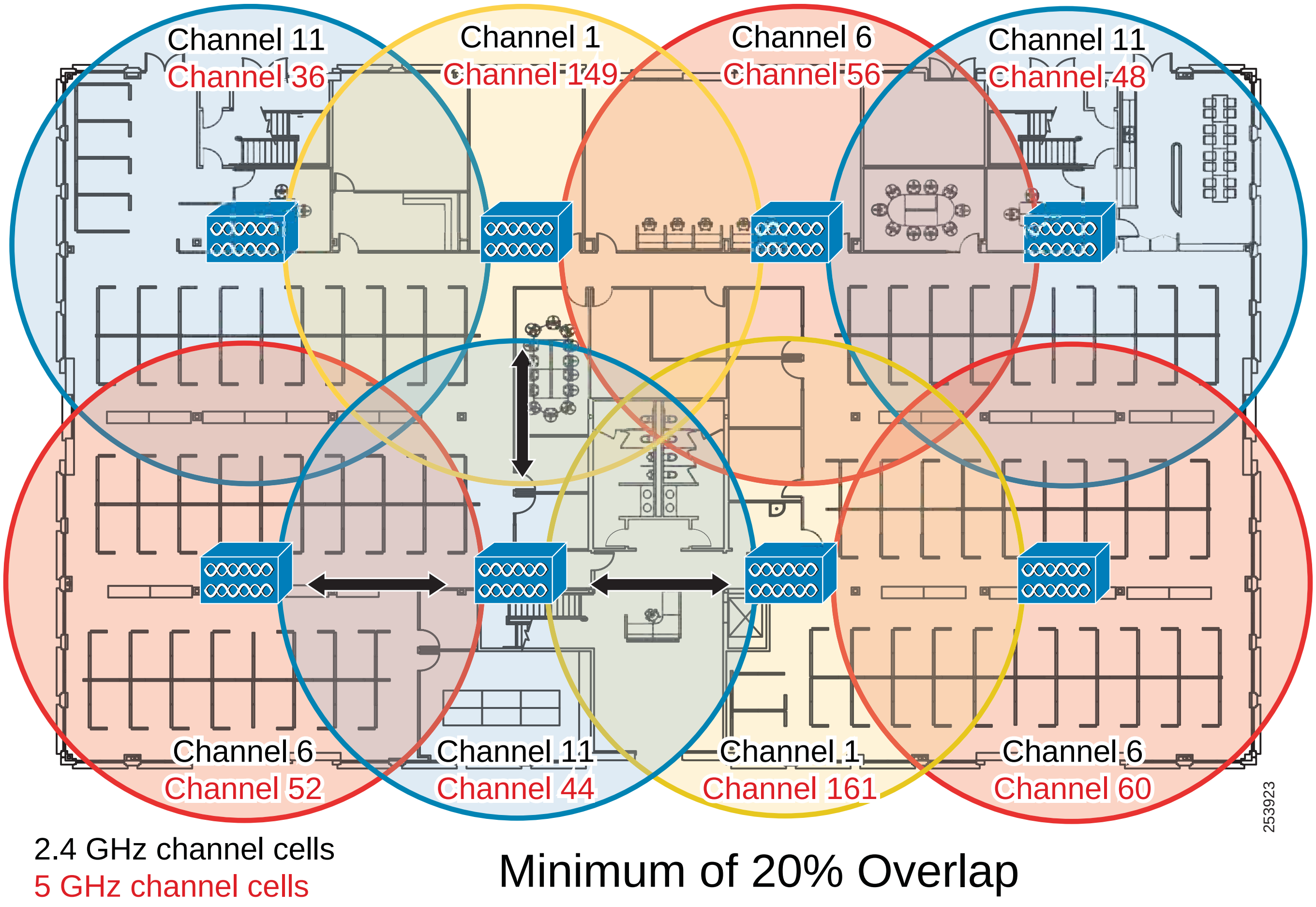
A WLAN network consists of one or more wireless access points (APs), which provide wireless network connectivity for wireless devices. Wireless APs are the demarcation point between the wireless network and the wired network. Multiple APs are deployed and distributed over a physical area of coverage in order to extend network coverage and capacity.

Because wireless devices and clients rely on the underlying WLAN infrastructure to carry both critical signaling and the real-time voice and video media traffic, it is necessary to deploy a WLAN network optimized for both data and real-time traffic. A poorly deployed WLAN network will be subjected to large amounts of interference and diminished capacity, leading not only to poor voice and video quality but in some cases dropped or missed calls. This will in turn render the WLAN deployment unusable for making and receiving voice calls. Therefore, when deploying wireless phones and clients, it is imperative to conduct a WLAN radio frequency (RF) site survey before, during, and after the deployment to determine appropriate cell boundaries, configuration and feature settings, capacity, and redundancy to ensure a successful voice and video over WLAN (VVoWLAN) deployment.

APs can be deployed autonomously within the network so that each AP is configured, managed, and operated independently from all other APs, or they can be deployed in a managed mode in which all APs are configured, managed, and controlled by a WLAN controller. In the latter mode, the WLAN controller is responsible for managing the APs as well as handling AP configuration and inter-AP roaming. In either case, to ensure successful VVoWLAN deployment, APs should be deployed using the following general guidelines:

- As shown in Figure 25-2, non-adjacent WLAN AP channel cells should overlap by a minimum of 20%. This overlap ensures that a wireless device can successfully roam from one AP to the next as the device moves around within the campus location while still maintaining voice and data network connectivity. A device that successfully roams between two APs is able to maintain an active voice call without any noticeable change in the voice quality or path.

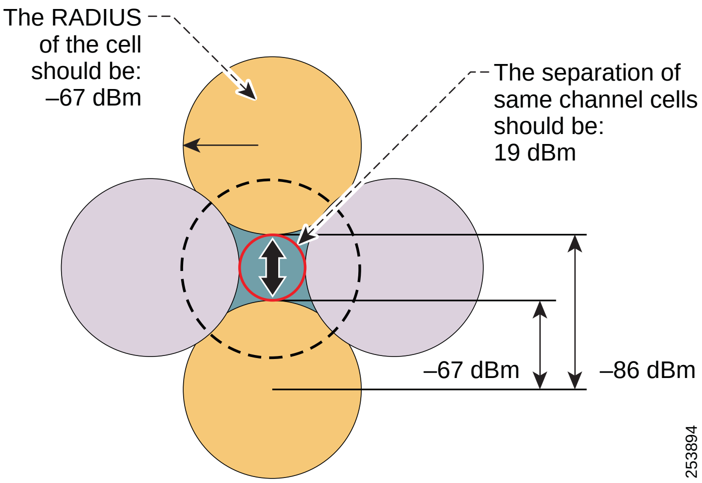
Figure 25-2 WLAN Channel Cell Overlap



- As shown in Figure 25-3, WLAN AP channel cells should be deployed with cell power-level boundaries (or channel cell radius) of -67 decibels per milliwatt (dBm). Additionally, the same-channel cell boundary separation should be approximately 19 dBm.

  A cell radius of approximately -67 dBm (or less) minimizes packet loss, which can be problematic for real-time voice and video traffic. A same-channel cell separation of 19 dBm is critical to ensure that APs or clients do not cause co-channel interference to other devices associated to the same channel, which would likely result in poor voice quality. The cell radius guideline of -67 dBm applies for both 2.4 GHz (802.11b/g) and 5 GHz (802.11a) deployments.

Figure 25-3 WLAN Cell Radius and Same Channel Cell Separation


The RADIUS of the cell should be: –67 dBm
The separation of same channel cells should be: 19 dBm
–67 dBm
–86 dBm
253894

Note The 19 dBm same-channel cell separation is simplified and is considered ideal. It is very unlikely that this 19 dBm of separation can be achieved in most deployments. The most important RF design criteria are the -67 dBm cell radius and the minimum 20% recommended overlap between cells. Designing to these constraints optimizes channel separation.

Wireless roaming is not limited to wireless phones but also applies to software-based phones running on wireless personal computers. For example, a user can roam wirelessly throughout the campus with a laptop computer running Cisco IP Communicator or Cisco Jabber.

Most wireless APs, wireless phones, and wireless PC clients provide a variety of security options for providing secure access to the enterprise WLAN. In all cases, select a security method supported by both the WLAN infrastructure and the wireless devices that matches the security policies and requirements of the enterprise.

For more information on the Cisco Unified Wireless Network Infrastructure, see Wireless LAN Infrastructure, page 3-54. For more details on Voice over WLAN design, refer to the *Voice over Wireless LAN Design Guide*, available at

http://www.cisco.com/en/US/solutions/ns340/ns414/ns742/ns820/landing_voice_wireless.html

Extension Mobility (EM)

As shown in Figure 25-1, in addition to physical movement of wired and wireless phones, the users themselves can also move around within the campus infrastructure without phone or PC hardware. In these cases, a user can move their enterprise extension or number from one device to another by applying a device profile containing the user's enterprise number and other settings.

The EM feature allows users to log on to IP phones located throughout the campus using a set of security credentials (user ID and PIN number). Once logged on, the user's personal device profile, including their enterprise phone number, calling privileges, and even their configured speed dials, is applied to the phone temporarily until the user logs out of the device or the login times out. The EM feature is available as part of Unified CM.

This feature is particularly useful for mobile enterprise users who spend considerable amounts of time outside the enterprise and are physically in the office only occasionally. By providing temporary office space for these types of mobile users, sometimes referred to as hot seating or free seating, a system administrator can accommodate large numbers of mobile users who only occasionally and temporarily need to use IP phone hardware.

To leverage EM within the campus the Unified CM administrator must configure user device profile(s) and user credentials, and subscribe IP phone(s) to the EM phone service.

For more information about EM, see Extension Mobility, page 19-8.

Campus Enterprise Mobility High Availability

Campus enterprise mobility features and solutions should be configured and deployed in a redundant fashion to ensure high availability of mobility functions and features.

For example, to effectively support hard-wired IP phones and computers running software-based IP phones, redundant and prevalent network connections or ports should be made available. Furthermore, these redundant network connections should be deployed with appropriate characteristics, including appropriate security, quality of service, and other network-based features to ensure optimal operation and voice quality for wired devices as they are moved from location to location. Ultimately a successful campus mobility deployment is possible only if the underlying network connectivity, PSTN connectivity, and other applications and services are deployed in a highly available fashion.

Likewise, when deploying or tuning a WLAN network for wireless device connectivity and roaming, it is also important to consider high availability for wireless services. To ensure resilient and sufficient coverage for the number of devices being deployed, a WLAN network should be deployed in a manner that ensures that adequate and redundant cells of coverage are provided without overlapping same-channel cells. Network connectivity for wireless devices and clients can be made highly available by providing ample cell coverage without same-channel cell overlap and sufficient overlap of different channel cells in order to facilitate roaming between APs.

Finally, when leveraging EM for user mobility within the campus, you should deploy this feature in a redundant fashion so that the failure of a single node within the Unified CM cluster does not prevent the operation of the Extension Mobility feature. For information on deploying Cisco Extension Mobility in a highly available manner, see High Availability for Extension Mobility, page 19-16.

Capacity Planning for Campus Enterprise Mobility

Deploying campus enterprise mobility successfully requires providing ample capacity to accommodate all mobile users exercising these mobility features and solutions.

Capacity considerations for physical movement of wired devices and computers depend completely on the number of network ports that are made available within the campus network infrastructure. In order for users to move devices around the campus, there must be some number of available network ports in each location that can be used to connect these mobile users' devices. A shortage of network ports to accommodate this wired device movement can result in an inability to move a device physically from one location to another.

When deploying wireless devices and leveraging wireless device roaming within the enterprise WLAN, it is also important to consider the device connectivity and call capacity of the WLAN infrastructure. Oversubscription of the campus WLAN infrastructure in terms of number of devices or number of active calls will result in dropped wireless connections, poor voice and video quality, and delayed or failed call setup. The chances of oversubscribing a deployment of voice and video over WLAN (VVoWLAN) are greatly minimized by deploying sufficient numbers of APs to handle required call capacities. AP call capacities are based on the number of simultaneous voice and/or video bidirectional streams that can be supported in a single channel cell area. The general rule for VVoWLAN call capacities is as follows:

- Maximum of 27 simultaneous voice over WLAN (VoWLAN) bidirectional streams per 802.11g/n (2.4 GHz) channel cell with Bluetooth disabled and 24 Mbps or higher data rates.
- Maximum of 27 simultaneous VoWLAN bidirectional streams per 802.11a/n (5 GHz) channel cell with 24 Mbps or higher data rates.
- Assuming a video resolution of 720p (high-definition) and a video bit rate of up to 1 Mbps, a maximum of 8 simultaneous VVoWLAN bidirectional streams per 802.11 g/n (2.4 GHz) with Bluetooth disabled or 802.11 a/n (5 GHz) channel cell.

These voice and video call capacity values are highly dependent upon the RF environment, the configured or supported video resolution and bit rates, the wireless endpoint and its specific capabilities, and the underlying WLAN system features. Actual capacities for a particular deployment could be less.

Note A single call between two wireless endpoints associated to the same AP is considered to be two simultaneous bidirectional streams.

Scalability of EM is dependent almost completely on the login/logout rate of the feature within Unified CM. It is important to know the number of extension mobility users enabled within the Unified CM cluster as well as how many users are moving around the campus and exercising this feature at any given time to ensure that sufficient EM login/logout capacity can be provided to these mobile users. For more information on EM capacity planning, see the chapter on Unified Communications Design and Deployment Sizing Considerations, page 29-1.

In all cases, the Unified CM cluster within the campus must have sufficient device registration capacity to handle device registration for moved devices, regardless of whether they are wired or wireless devices. Of course, assuming all devices being moved throughout the campus are already deployed within the campus network, then sufficient capacity within Unified CM should already be in place prior to the movement of devices. If new devices are added to the deployment for mobility purposes, however, device registration capacity should be considered and, if necessary, additional capacity should be added.

Finally, given the many features and functions provided by Unified CM, configuration and deployment of these mobility solutions does have sizing implications for the overall system. Determining actual system capacity is based on considerations such as number of endpoint devices, EM users, and busy hour call attempt (BHCA) rates to number of CTI applications deployed. For more information on general system sizing, capacity planning, and deployment considerations, see the chapter on Unified Communications Design and Deployment Sizing Considerations, page 29-1.

Design Considerations for Campus Enterprise Mobility

Observe the following design recommendations when deploying campus enterprise mobility features and solutions:

- To accommodate physical device mobility within the campus ensure that the network connection used at a new location has the same type of IP connectivity (VLANs, inter-VLAN routing, and so forth), connection speed, quality of service, security, and network services (in-line power, dynamic host control protocol (DHCP), and so forth) as provided by the previous network connection. Failure to replicate these connection parameters, services, and features will lead to diminished functionality and in some case complete loss of functionality.

- When deploying wireless IP devices and software-based clients, it is imperative to conduct a WLAN radio frequency (RF) site survey before, during, and after the deployment to determine appropriate cell boundaries, configuration and feature settings, capacity, and redundancy to ensure a successful voice over WLAN (VoWLAN) deployment.
- APs should be deployed with a minimum cell overlap of 20%. This overlap ensures that a dual-mode device can successfully roam from one AP to the next as the device moves around within a location, while still maintaining voice and data network connectivity.
- APs should be deployed with cell power level boundaries (or channel cell radius) of -67 dBm in order to minimize packet loss. Furthermore, the same-channel cell boundary separation should be approximately 19 dBm. A same-channel cell separation of 19 dBm is critical for ensuring that APs or clients do not cause co-channel interference to other devices associated to the same channel, which would likely result in poor voice and video quality.
- Deploy EM services in a highly redundant manner so that the loss of a single Unified CM node does not have adverse effects on the feature operation. If EM services are critical, consider deploying a server load balancing solution to route around Unified CM node failures and provide highly available functionality. For more information on EM high availability, see High Availability for Extension Mobility, page 19-16.
- Provide sufficient wireless voice and video call capacity on the campus network by deploying the appropriate number of wireless APs to handle the desired call capacity based on wireless user BHCA rates. Each 802.11g/n (2.4 GHz) or 802.11a/n (5 GHz) channel cell can support a maximum of 27 simultaneous voice-only calls with 24 Mbps or higher data rates. Each 802.11g/n (2.4 GHz) or 802.11a/n (5 GHz) channel cell can support a maximum of 8 simultaneous video calls assuming 720p video resolution at up to 1 Mbps bit rate. For 2.4 GHz WLAN deployments, Bluetooth must be disabled to achieve this capacity. Actual call capacity could be lower depending on RF environment, wireless endpoint type, and WLAN infrastructure.

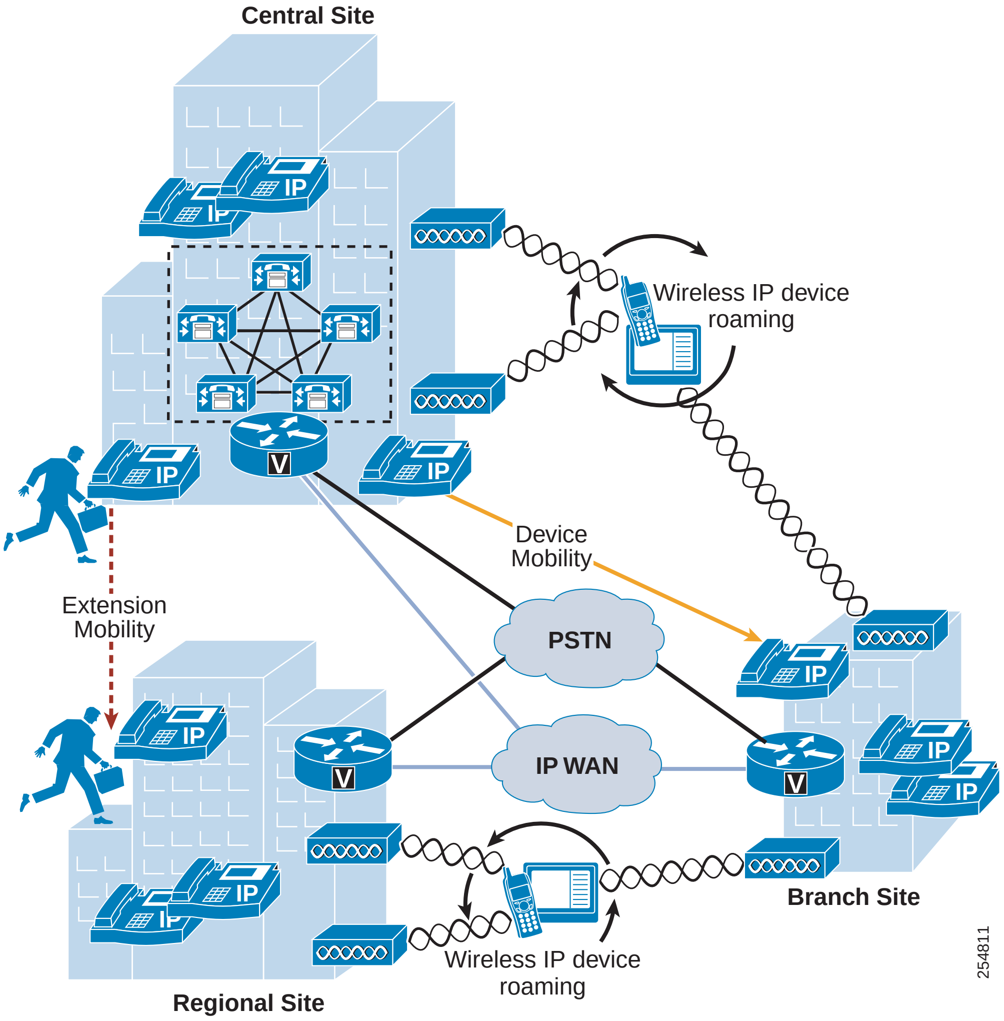
Multisite Enterprise Mobility

Multisite enterprise mobility refers to mobility within an enterprise with multiple physical locations, each with a unique IP address space and PSTN egress/ingress boundary. Mobility in this case includes not only the movement of users and endpoint devices within each physical location but also movement of users and endpoint devices between sites and locations.

Multisite Enterprise Mobility Architecture

As shown in Figure 25-4, the multisite enterprise mobility architecture is based on two or more locations or sites geographically separated. Sites may vary in size from large numbers of users and devices in a central or campus site to smaller numbers of users and devices in medium-sized regional sites or smaller branch sites. Typically multisite enterprise deployments consist of IP WAN links interconnecting sites as well as local PSTN egress/ingress at each location. In addition, critical services are often replicated at each physical site in order to maintain features and functions during network outages between sites. From a mobility perspective, users and their devices may be mobile within a site or between sites.

Figure 25-4 Multisite Enterprise Mobility Architecture



Note While Figure 25-4 depicts a multisite deployment with centralized call processing (as evidenced by a single Unified CM cluster within the central site), the same design and deployment considerations for multisite enterprise mobility deployments apply to distributed call processing environments. Differences in mobility feature operation when deployed in distributed call processing environments are described in the following discussions.

Types of Multisite Enterprise Mobility

Mobility within a multisite enterprise deployment involves not only the movement of devices, users, or both within a single site, but also movement of users and devices between sites.

The same types of mobility features and solutions supported with campus or single site enterprise deployments apply to intra-site movement of users and devices within any single site of a multisite deployment. These include physical wired phone movement, wireless phone roaming, and extension mobility. For information on these types of mobility solutions and functions, see Campus Enterprise Mobility, page 25-4.

For inter-site mobility in a multisite deployment, these same mobility features are also supported in much the same way. However, the key difference with these features when applied between two or more sites is that they are augmented with the Device Mobility feature. The Device Mobility feature provides a mechanism for dynamic location awareness of devices based on the IP address the device uses when connecting to the enterprise network.

Physical Wired Device Moves

Movement of physical wired phones is easily accommodated within each site of a multisite deployment as well as between sites. Just as with a campus or single-site deployment, wired device movement limited to a single site of a multisite deployment simply involves unplugging a Cisco Unified IP Phone from the network, moving it to another location within the site, and plugging it into another wired network port. Once connected to the new network location, the phone simply re-registers to Unified CM and is able to make and receive calls just like in the previous location.

Movement of wired devices between sites or locations in a multisite deployment involve the same basic behavior. However, the Device Mobility feature, when combined with this type of mobility, ensures that call admission control operations and gateway and codec selection are appropriate once the device re-registers in the new location to which it has been moved. See Device Mobility, page 25-14, for information about this feature.

Wireless Device Roaming

Just as with a single-site campus deployment, wireless devices can move or roam throughout a multisite enterprise deployment, as shown in Figure 25-4, provided wireless LAN network infrastructure is available at each site to provide wireless network connectivity. However, as with the movement of wired phones between sites, the Device Mobility feature should also be deployed for wireless devices to ensure that the correct gateway and codec are used when making and receiving calls and that call admission control manages bandwidth appropriately. See Device Mobility, page 25-14, for information about this feature.

For distributed call processing environments, just as with wired phones, wireless devices should be configured to register with only a single Unified CM cluster to avoid potential issues with call routing.

Extension Mobility (EM)

In addition to supporting EM within a single site, as illustrated in Figure 25-4, this feature is also supported between sites to enable users to move between sites within the enterprise and log on to phones in each locations.

EM is also supported in distributed call processing deployments when users move between sites and phones on different Unified CM clusters. To support extension mobility in distributed call processing environments, you might need to configure the Cisco Extension Mobility Cross Cluster (EMCC) feature. For information about this feature, see Extension Mobility Cross Cluster (EMCC), page 19-10.

Device Mobility

In Cisco Unified Communications Manager (Unified CM), a site or a physical location is identified using various settings such as locations, regions, calling search spaces, and media resources. Cisco Unified IP Phones residing in a particular site are statically configured with these settings. Unified CM uses these settings for proper call establishment, call routing, media resource selection, and so forth. However, when dual-mode phones and other mobile client devices such as Cisco Cius or Cisco Unified Wireless IP Phones are moved from their home site to a remote site, they retain the home settings that are statically configured on the phones. Unified CM then uses these home settings on the phones in the remote site. This situation is undesirable because it can cause problems with call routing, codec selection, media resource selection, and other call processing functions.

Cisco Unified CM uses a feature called Device Mobility, which enables Unified CM to determine if the IP phone is at its home location or at a roaming location. Unified CM uses the device's IP subnets to determine the exact location of the IP phone. By enabling device mobility within a cluster, mobile users can roam from one site to another, thus acquiring the site-specific settings. Unified CM then uses these dynamically allocated settings for call routing, codec selection, media resource selection, and so forth.

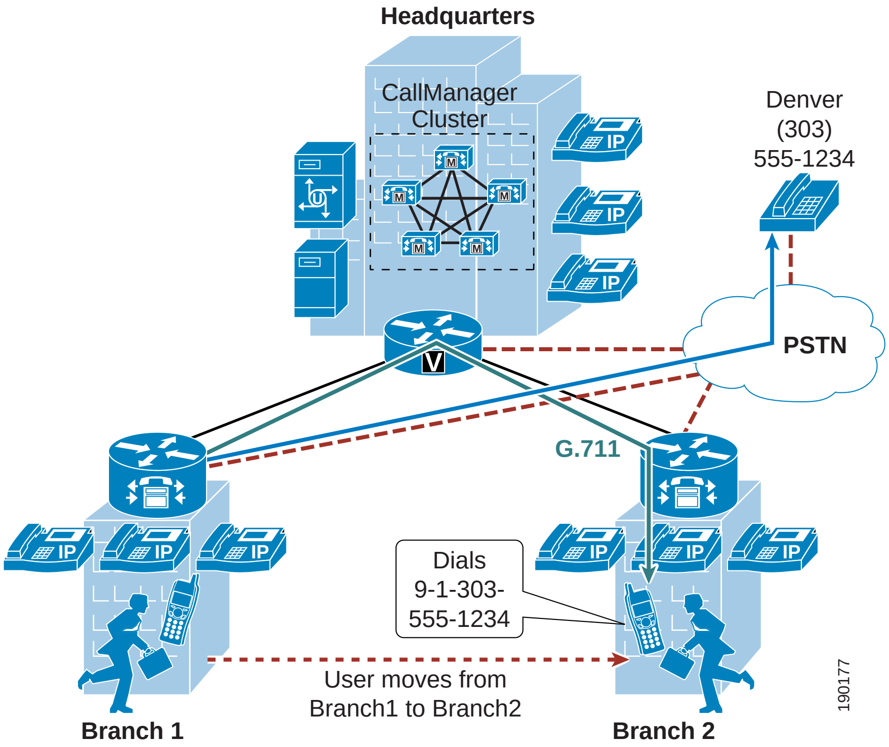
This section begins with a discussion surrounding the main purpose for the Device Mobility feature, followed by an in-depth discussion of the Device Mobility feature itself. This discussion covers the various components and configuration constructs of the Device Mobility feature. This section also presents an in-depth discussion of the impact of the Device Mobility feature on the enterprise dial plan, including the implication for various dial plan models.

Need for Device Mobility

This section explains the need for device mobility when there are many mobile users in a Unified CM cluster.

Figure 25-5 illustrates a hypothetical network containing a Unified CM cluster without the Device Mobility feature, located at the headquarter site (HQ). The cluster has two remote sites, Branch1 and Branch2. All intra-site calls use G.711 voice codecs, while all inter-site calls (calls across the IP WAN) use G.729 voice codecs. Each site has a PSTN gateway for external calls.

Figure 25-5 Example Network with Two Remote Sites



When a user in Branch1 moves to Branch2 and calls a PSTN user in Denver, the following behavior occurs:

- Unified CM is not aware that the user has moved from Branch1 to Branch2. An external call to the PSTN is sent over the WAN to the Branch1 gateway and then out to the PSTN. Thus, the mobile user continues to use its home gateway for all PSTN calls.
- The mobile user and Branch1 gateway are in the same Unified CM region and location. Location-based call admission control is applicable only for devices in different locations, and an intra-region call uses the G.711 voice codec. Thus, the call over the IP WAN to the Branch1 gateway uses the G.711 codec and is not tracked by Unified CM for purposes of call admission control. This behavior can result in over-subscription of the IP WAN bandwidth if all the remote links are low-speed links.
- The mobile user creates a conference by adding multiple Branch2 users to the existing call with the PSTN user in Denver. The mobile user uses the conferencing resource that is on the Branch1 gateway, therefore all conference streams flow over the IP WAN.

Note Device Mobility is an intra-cluster feature and does not span multiple Unified CM clusters. In distributed call processing environments, Device Mobility must be enabled and configured on each Unified CM cluster within the deployment.

Device Mobility Architecture

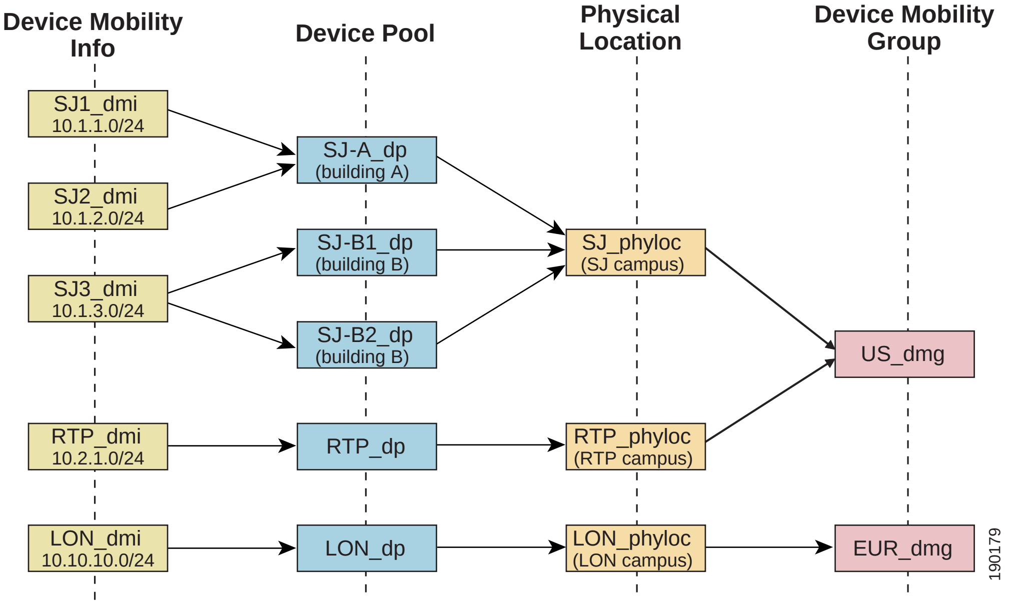
The Unified CM Device Mobility feature helps solve the problems mentioned above. This section briefly explains how the feature works. However, for a detailed explanation of this feature, refer to the product documentation available on http://www.cisco.com.

Some of the device mobility elements include:

- Device Mobility Info — Configures IP subnets and associates device pools to the IP subnets.
- Device Mobility Group — Defines a logical group of sites with similar dialing patterns (for example, US_dmg and EUR_dmg in Figure 25-6).
- Physical Location — Defines the physical location of a device pool. In other words, this element defines the geographic location of IP phones and other devices associated with the device pool. (For example, all San Jose IP phones in Figure 25-6 are defined by physical location SJ_phyloc.)

Figure 25-6 illustrates the relationship between all these terms.

Figure 25-6 Relationship of Device Mobility Components



Unified CM assigns a device pool to an IP phone based on the device's IP subnet. The following steps, illustrated in Figure 25-7, describe the behavior:

1. The IP phone tries to register to Unified CM by sending its IP address in the Skinny Client Control Protocol (SCCP) or Session Initiation Protocol (SIP) registration message.
2. Unified CM derives the device's IP subnet and matches it with the subnet configured in the Device Mobility Info.

3. If the subnet matches, Unified CM provides the device with a new configuration based on the device pool configuration.

Figure 25-7 Phone Registration Process

Unified CM uses a set of parameters under the device pool configuration to accommodate Device Mobility. These parameters are of the following two main types:

- Roaming Sensitive Settings, page 25-17
- Device Mobility Related Settings, page 25-18

Roaming Sensitive Settings

The parameters under these settings will override the device-level settings when the device is roaming within or outside a Device Mobility Group. The parameters included in these settings are:

- Date/time Group
- Region
- Media Resource Group List
- Location
- Network Locale
- SRST Reference
- Physical Location
- Device Mobility Group

The roaming sensitive settings primarily help in achieving proper call admission control and voice codec selection because the location and region configurations are used based on the device's roaming device pool.

For more details on various call admission control techniques, see the chapter on Call Admission Control, page 11-1.

The roaming sensitive settings also update the media resource group list (MRGL) so that appropriate remote media resources are used for music on hold, conferencing, transcoding, and so forth, thus utilizing the network efficiently.

The roaming sensitive settings also update the Survivable Remote Site Telephony (SRST) gateway. Mobile users register to a different SRST gateway while roaming. This registration can affect the dialing behavior when the roaming phones are in SRST mode.

For example, if a user moves with their phone to a new location that loses connectivity to Unified CM, then based on the roaming sensitive Device Mobility settings, a new SRST reference is configured for the moved phone and the moved phone will now be under control of the local roaming location SRST router. When this occurs, not only would the user's phone be unreachable from the PSTN or other sites because the device's DID will not have changed and will still be anchored at their home location, but in addition reachabililty from devices within the local failed site might be difficult without the use of abbreviated dialing as implemented within SRST.

As an example, assume that a user moves a phone from their home location in San Jose, which has a directory number of 51234 and an associated DID of 408 555 1234 to a remote location in New York, and that the link between the New York site and San Jose fails shortly after the user roams to the New York location. In this scenario the phones in the New York site will all fail-over to the SRST router in that site. The roaming/moved phone will also register to the New York SRST router because its SRST reference was updated based on the device mobility roaming sensitive settings. In this scenario, the local New York devices will register to the SRST router with five-digit extensions just as they do to Unified CM, and as a result the roaming phone still has a directory number of 51234. To reach the roaming phone from all other sites and from the PSTN, the number 408 555 1234 will be routed to the San Jose PSTN gateway to which this particular DID is anchored. Because the New York site is disconnected from the San Jose site, any such calls will be routed to the users' voicemail boxes since they will be unreachable at their desk phones. Likewise, calls internally within the local failed site will have to be dialed using five-digit abbreviated dialing or based on the configured digit prefixing as defined by the **dialplan-pattern** and **extension-length** commands within the SRST router. In either case, local callers will have to be understand the required dialing behavior for reaching the local roaming device by abbreviated dialing. In some cases this may be simply five-digit dialing or it may be that users have to dial a special digit prefix to reach the local roaming phone. The same logic applies to outbound dialing from the moved or roaming phone in New York because its dialing behavior might have to be altered in order to reach local extensions using abbreviated dialing. Outbound dialing to the PSTN from the local roaming device should remain the same, however.

Device Mobility Related Settings

The parameters under these settings will override the device-level settings only when the device is roaming within a Device Mobility Group. The parameters included in these settings are:

- Device Mobility Calling Search Space
- AAR Calling Search Space
- AAR Group
- Calling Party Transformation CSS

The device mobility related settings affect the dial plan because the calling search space dictates the patterns that can be dialed or the devices that can be reached.

Device Mobility Group

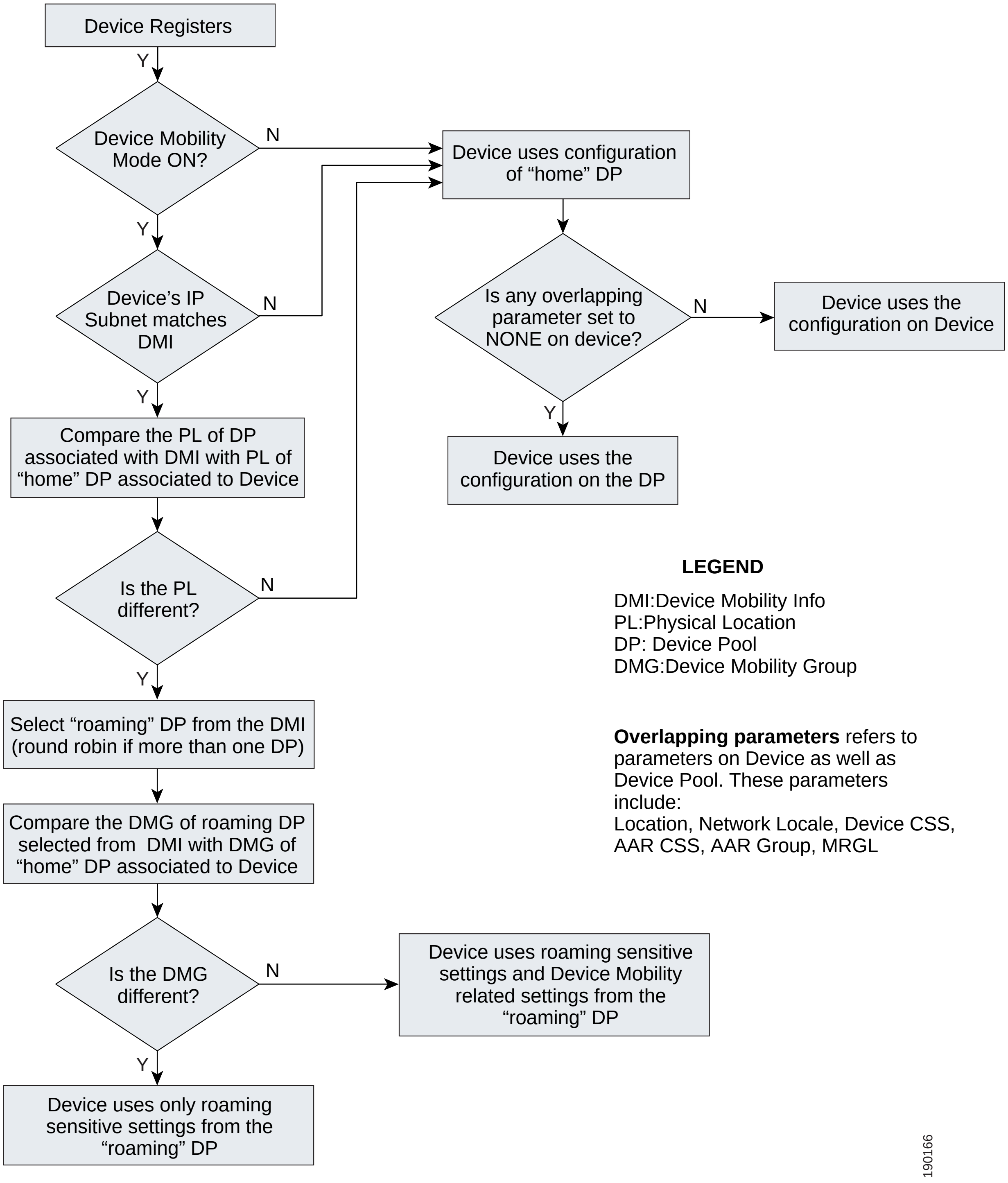
Device Mobility Group, as explained earlier, defines a logical group of sites with similar dialing patterns (for example, sites having the same PSTN access codes and so forth). With this guideline, all sites have similar dialing patterns in the site-specific calling search spaces. Sites having different dialing behavior are in a different Device Mobility Group. As illustrated in Figure 25-6, the San Jose and RTP sites' Device Mobility Info, Device Pools, and Physical Locations are different; however, all of these have been assigned to the same Device Mobility Group US_dmg because the required dialing patterns and PSTN access codes are the same between the two locations. On the other hand, the London site is assigned to a separate Device Mobility Group EUR_dmg due to the fact that the required dialing patterns and PSTN access codes there are different than those of the US sites. A user roaming within a Device Mobility Group may preserve his dialing behavior at the remote location even after receiving a new calling search space. A user roaming outside the Device Mobility Group may still preserve his dialing behavior at the remote location because he uses his home calling search space.

However, if a Device Mobility Group is defined with sites having different dialing patterns (for example, one site requires users to dial 9 to get an outside line while another site requires users to dial 8 to get an outside line), then a user roaming within that Device Mobility Group might not preserve his same dialing behavior at all locations. A user might have to dial digits differently at different locations after receiving a new calling search space at each location. This behavior can be confusing for users, therefore Cisco recommends against assigning sites with different dialing patterns to the same Device Mobility Group.

Device Mobility Operation

The flowchart in Figure 25-8 represents the operation of the Device Mobility feature.

Figure 25-8 Operation of the Device Mobility Feature


Device Registers
Y
Device Mobility Mode ON?
N
Device uses configuration of “home” DP
Y
Device’s IP Subnet matches DMI
N
Is any overlapping parameter set to NONE on device?
N
Device uses the configuration on Device
Y
Compare the PL of DP associated with DMI with PL of “home” DP associated to Device
Y
Device uses the configuration on the DP
Is the PL different?
N
Y
Select “roaming” DP from the DMI (round robin if more than one DP)
Compare the DMG of roaming DP selected from DMI with DMG of “home” DP associated to Device
Is the DMG different?
N
Device uses roaming sensitive settings and Device Mobility related settings from the “roaming” DP
Y
Device uses only roaming sensitive settings from the “roaming” DP
LEGEND
DMI:Device Mobility Info
PL:Physical Location
DP: Device Pool
DMG:Device Mobility Group
Overlapping parameters refers to parameters on Device as well as Device Pool. These parameters include:
Location, Network Locale, Device CSS, AAR CSS, AAR Group, MRGL
190166

The following guidelines apply to the Device Mobility feature:

- If the overlapping parameters listed in Figure 25-8 have the same configurations on the device as well as the device pool, then these parameters may be set to NONE on the device. These parameters must then be configured on the device pool. This practice can greatly reduce the amount of configuration because the devices do not have to be configured individually with all the parameters.
- Define one physical location per site. A site may have more than one device pool.
- Define sites with similar dialing patterns for PSTN or external/off-net access with the same Device Mobility Group.
- A "catch-all" Device Mobility Info with IP subnet 0.0.0.0 may be defined for all non-defined subnets, depending on the company policy. This Device Mobility Info may be used to assign a device pool that can restrict access or usage of the network resources. (For example, the device pool may be configured with a calling search space NONE that will block any calls from the device associated with this device pool while roaming.) However, by doing so, administrators must be aware of the fact that this will block all calls, even 911 or other emergency calls. The calling search space may be configured with partitions that will give access only to 911 or other emergency calls.

Dial Plan Design Considerations

The Device Mobility feature uses several device and device pool settings that are based on the settings in the roaming device pool selected and on the IP address with which the endpoint registers. For details of which settings are updated with the settings of the device pool for the subnet, refer to the *Cisco Unified Communications Manager Features and Services Guide*, available at

http://www.cisco.com/en/US/products/sw/voicesw/ps556/prod_maintenance_guides_list.html

From the dial plan perspective, mainly the AAR group, AAR CSS, device CSS, Local Route Group, and outgoing call's calling party transformation CSS settings are relevant.

Egress Gateway Selection for Roaming Devices

Typically the desired egress gateway selection behavior of roaming devices is to use gateways local to the visited site. The recommended way to implement egress gateway selection that is specific to the calling device is to use PSTN route patterns pointing to route lists that use Standard Local Route Group. Using Standard Local Route Group in a route list effectively means that Standard Local Route Group, when routing an actual call, will be replaced with the Local Route Group configured in the device pool of the calling endpoint. This schema ensures that site-unspecific route patterns and route lists are used; site-specific egress gateway selection completely relies on device pool-level Local Route Group configuration.

For roaming devices (whether roaming inside or between device mobility groups), the device mobility feature always ensures that the Local Route Group of the roaming device pool is used as Standard Local Route Group. This guarantees that, with Local Route Group egress gateway selection, a visited site-specific route group (and thus gateways local to the visited site) will typically be used. This behavior ensures that, for example, emergency calls routed via route patterns that use a Standard Local Route Group route list will always use egress gateways local to the visited site.

Local Route Group egress gateway selection can be used with all dial plan approaches explained in the chapter on Dial Plan, page 9-1.

If certain calls from roaming endpoints need to be routed through gateways local to the home site of the roaming phone, then routing for these calls has to be implemented through route patterns pointing to route lists that use fixed site-specific route groups instead of Standard Local Group.

In a line/device dial plan approach, these route patterns would be addressed by the device CSS configured on the endpoint. When roaming but not leaving the device mobility group, the calling endpoint's device CSS is replaced by the Device Mobility CSS configured on the roaming device pool.

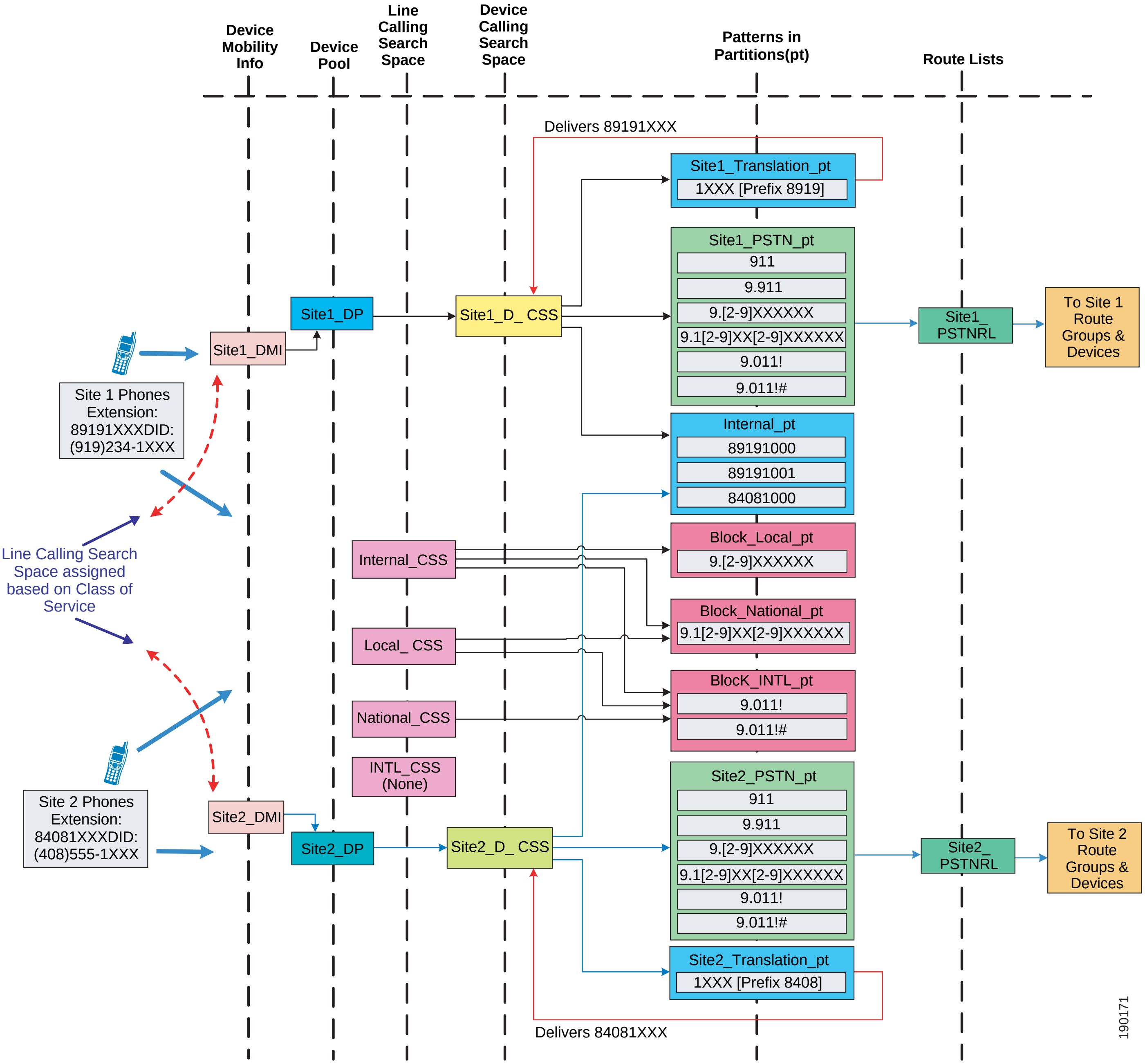
If fixed egress gateway selection is required for some calls and the route patterns for those calls are addressed by the device CSS, you have to make sure that roaming devices always roam across device mobility groups. This will guarantee that roaming endpoints always use the device CSS configure on the endpoint.

When using the +E.164 dial plan approach explained in the chapter on Dial Plan, page 9-1, all PSTN route patterns are accessible by the line CSS, which is not changed or updated for roaming devices. In this dial plan, site-specific route patterns tying specific PSTN destinations to fixed gateways (for example, in the home location of the roaming device) are not affected by device mobility operation.

Variable Length On-Net Dialing with Flat Addressing Using the Line/Device Approach without Local Route Group

Figure 25-9 shows a variable-length on-net dial plan with flat addressing for Device Mobility.

Figure 25-9 Variable-Length On-Net Dial Plan with Flat Addressing for Device Mobility



The following design considerations apply to the dial plan model in Figure 25-9:

- In this dial plan the translation patterns implementing 4-digit intra-site dialing are addressed by the device CSS. This is done to avoid the requirement to have site-specific line CSSs. Mobile users inherit the intra-site dialing of the visited site because the device CSS is updated with the roaming device pool's device mobility CSS (assuming the user is roaming inside the device mobility group). If this behavior is not desired, consider defining each site as a Device Mobility Group. However, users must be aware that, for any external PSTN calls, the mobile phone continues to use the home gateway and therefore consumes WAN bandwidth. This can be avoided by using Standard Local Route Group (see Egress Gateway Selection for Roaming Devices, page 25-21).
- Additional device calling search spaces may be configured for roaming users with access only to the PSTN and internal phones partitions. This configuration will need at least one additional device pool and calling search space per site. Thus, *N* sites will need *N* device pools and *N* calling search spaces. However, this configuration will not require defining each site as a Device Mobility Group. With this configuration mobile users, when roaming, will not have access to dialing habits through translation patterns in their device CSS.
- Mobile users registered with a remote SRST gateway have unique extensions. However, mobile users must be aware that no PSTN user can call them when they are registered to a remote SRST gateway.

+E.164 Dial Plan with Traditional Approach and Local Route Group

As described in the chapter on Dial Plan, page 9-1, the line/device approach has some specific issues, and creating a +E-164 dial plan based on the line/device approach is not recommended. The recommended approach for +E.164 dial plans is to combine class of service selection and dialing normalization on the line CSS and use the Local Route Group feature to address the requirement for site-specific egress gateway selection. In this approach the device CSS on the phone is not used at all. If you combine this approach with device mobility, the only roaming sensitive component of the design is the device pools' local route group. For a roaming phone (whether roaming inside or between device mobility groups), the local route group defined on the phone's home device pool will always be updated with the local route group defined on the roaming device pool. This guarantees that all calls always egress through a gateway local to the visited site.

Multisite Enterprise Mobility High Availability

Multisite enterprise mobility features and solutions should be configured and deployed in a redundant fashion in order to ensure high availability of mobility functionality. High availability considerations for wired phone moves, wireless roaming, and EM in multisite mobility deployments are similar to those for campus mobility deployments. Just as with campus environments, redundant network ports, wireless cell coverage, and Unified CM nodes handling extension mobility logins and logouts should be provided to ensure highly available services.

Similarly, it is important to consider high availability of the Device Mobility feature. Because Device Mobility is natively integrated within Unified CM, the failure of a cluster node should have no impact on the functionality of Device Mobility. Device pool, Device Mobility Info, Device Mobility Group, and all other configurations surrounding Device Mobility are preserved if there is a failure of the publisher node or a call processing (subscriber) node. Additionally, if there is a call processing node failure, affected phones will fail-over to their secondary call processing node or Survivable Remote Site Telephony (SRST) reference router as usual based on the Unified CM Group construct.

Capacity Planning for Multisite Enterprise Mobility

As for Device Mobility scalability considerations, there are no specific or enforced capacity limits surrounding this feature and the various configuration constructs (device pools, device mobility groups, and so forth). For more information on general system sizing, capacity planning, and deployment considerations, see the chapter on Unified Communications Design and Deployment Sizing Considerations, page 29-1

Design Considerations for Multisite Enterprise Mobility

All campus enterprise mobility design considerations apply to multisite enterprise mobility deployments as well (seeDesign Considerations for Campus Enterprise Mobility, page 25-10). The following additional design recommendations apply specifically to multisite mobility environments:

- Ensure that all critical services (device registration, PSTN connectivity, DNS, DHCP, and so forth) are deployed at each site in a multisite deployment so that failure of the connection between the site and other sites does not disrupt critical operations. In addition, ensure that a sufficient number of physical network ports and wireless LAN APs are available at each site to support movement of devices and required call capacity.
- In situations in which sites with different dialing patterns (for example, sites having different PSTN access codes) are configured in the same Device Mobility Group, roaming users might have to dial numbers differently based on their location, which can be confusing. For this reason, Cisco recommends assigning sites with similar dialing patterns (for example, sites having the same PSTN access codes) to the same Device Mobility Group. Doing so ensures that roaming users can dial numbers the same way at all sites within the Device Mobility Group.
- The Device Mobility settings from the "roaming" device pool are applied only when users roam within the same Device Mobility Group; therefore, avoid roaming between different Device Mobility Groups because the resulting call routing behavior will cause originated calls from the moved phone to be routed using the "home" or device-configured calling search space. This can lead to unnecessary consumption of WAN bandwidth because the call might be routed through a different site's gateway rather than the local "roaming" gateway.
- Define only one physical location per site. This ensures that device mobility is engaged only in scenarios in which a user is roaming between sites. For roaming within the same site, the concerns that mandate Device Mobility (for example, WAN bandwidth consumption, codec selection, and call admission control) are not present because low-speed links typically are not deployed within a single site.
- In failover scenarios, "roaming" phones will utilize the SRST reference/gateway as dictated by the "roaming" device pool's roaming sensitive settings. Therefore, in these situations the "roaming" phone is unreachable from the PSTN due to the fact that the DID for this phone is anchored in another location's PSTN gateway. Furthermore, for outbound calls from the "roaming" phone, dialing behavior might have to be altered for things such as PSTN access codes, and speed dials configured on the phone might not be usable.
- If your system requires the ability to use abbreviated dialing or to use speed dials that rely on abbreviated dialing, Cisco recommends using a Uniform On-net dial plan model because it will ensure that abbreviated dialing (direct or through speed dials) continues to work even when the mobile user's phone is in a roaming location. Abbreviated dialing is still possible with this dial plan model because all extensions or directory numbers are unique across all sites, and therefore abbreviated dialing can be used universally due to the fact that there are no overlapping extensions.

- If your system uses a Variable Length On-net dial plan model (using either the line/device or the line-CSS-only +E.164 dial plan approach), Cisco recommends configuring speed dials in a universal way so that a single unique extension can be reached when called. By configuring speed dials using full +E.164 numbers or using site or access codes, you can enable roaming users to use the same speed dials at any location.
- If Device Mobility is enabled for users who on occasion access the enterprise network through a VPN connection, Device Mobility Info (DMI) for VPN attached phones should contain IP subnets distributed or owned by the VPN concentrators to ensure that "roaming" to a VPN location results in appropriate dynamic Device Mobility configuration changes. Be sure to associate the DMI with the same device pool that is used for any devices co-located with the VPN concentrators.

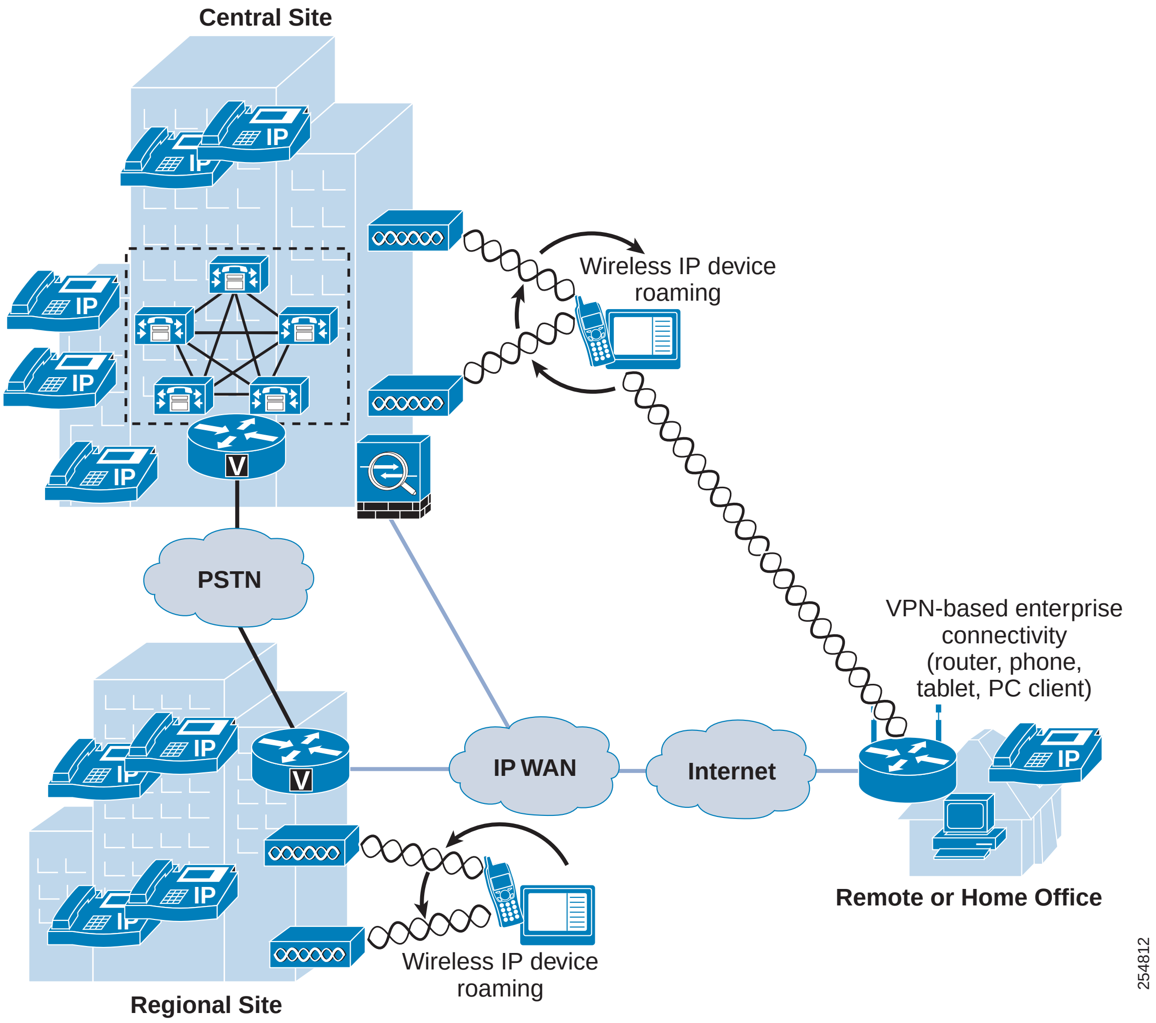
Remote Enterprise Mobility

Remote enterprise mobility refers to mobile users in locations remote from the enterprise but still attached to the enterprise network infrastructure through secure connections over the public Internet. Mobility here deals with the placement of endpoint devices in these remote locations and the movement of users, and in some cases their mobile devices, between the enterprise and these locations either frequently or on occasion.

Remote Enterprise Mobility Architecture

As illustrated in Figure 25-10, the remote enterprise mobility architecture is based on a remote physical location, typically an employee home office but also any remote location capable of secure connection back to the enterprise over the Internet. These remote sites typically consist of an IP network with connections for a user's computer, telephone, and other equipment or endpoints. In some cases this IP network may be behind an enterprise controlled and configured VPN router that provides a secure tunnel between the remote location and the enterprise network. In other cases, the remote site IP network is connected to the Internet through a user-provided router, and user computer or endpoint devices must use software-based VPN client capabilities to create secure connections back to the enterprise network. Wireless connectivity may also be provided in the remote location to allow wireless attachment of the user's computer or endpoint. When wireless connectivity is provided at the remote location, wireless phones may be moved from the enterprise network to the home office, allowing users to leverage wireless enterprise devices or mobile phones within the remote location to make and receive calls.

Figure 25-10 Remote Enterprise Mobility Architecture



Types of Remote Enterprise Mobility

Remote enterprise mobility deployments focus predominately on supporting remote users as opposed to supporting regular user or device movement. Certainly users may regularly move with or without an endpoint device between the enterprise location or locations and the remote site; however, the predominate purpose of these deployments is to support remote connectivity for enterprise users. Remote site mobility typically involves two main types of remote connectivity: router-based secure connectivity and client-based secure connectivity. Both types support remote site secure connectivity and both can accommodate various endpoint devices that can be moved between the remote site and the enterprise, including dual-mode mobile phones, wireless IP phones and tablets, and even wired IP phones.

Client-Based Secure Remote Connectivity

Wireless and wired IP phones and software-based PC telephony clients can be connected to remote site locations, as shown in Figure 25-10. These devices and endpoints are responsible for creating secure VPN connections back to the enterprise VPN head-end termination concentrator.

Examples of these types of devices include wirelessly attached mobile client devices using VPN client or application capabilities such as the Cisco Jabber for iPhone and Android clients (see Cisco Mobile Clients and Devices, page 25-60), wired Cisco Unified IP Phones such as the Cisco Unified IP Phone 7965 that uses a built-in VPN client, and personal computers running software-based telephony clients such as Cisco Jabber that uses a software-based VPN client for connectivity to the enterprise network.

Router-Based Secure Remote Connectivity

On the other hand, remote site connectivity can be handled through router-based secure VPN tunnels. In these types of scenarios the deployed remote site router, which may be able to provide wireless network connectivity as well, is responsible for setting up and securing a VPN tunnel back to the enterprise network. This in effect extends the enterprise network boundary to the remote site location. The advantage of this type of connectivity is that a wider range of devices and endpoints may be deployed in the remote site because these devices are not responsible for providing secure connectivity and therefore do not require special software or configuration. Instead, these devices simply connect to the remote site network and leverage the secure VPN IP path from the remote site router to the enterprise VPN head-end.

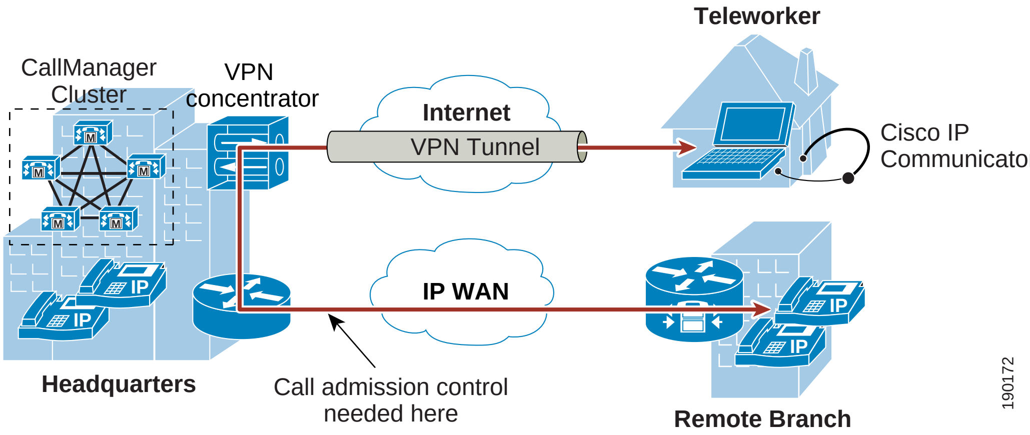
An example of this type of route-based remote site connectivity is the Cisco Virtual Office solution.

Device Mobility and VPN-Based Remote Enterprise Connectivity

Whether you are deploying client-based or router-based secure remote connectivity, the Device Mobility feature may be used to ensure that call admission control and codec are correctly negotiated for endpoint devices and that the appropriate enterprise site PSTN gateway and media resources are utilized. Based on the IP address of the endpoint device as received over the VPN connection, Unified CM will dynamically determine the location of the device.

Figure 25-11 shows an example of client-based secure remote connectivity where a Cisco IP Communicator software phone is running on a remote site computer. This software-based IP phone is connected through a client-based VPN back to the enterprise and registered to Unified CM.

Figure 25-11 Client-Based VPN Connection for Remote Site Cisco IP Communicator



The following design guidelines pertain to enabling the Device Mobility feature for user devices at a remote site connected to the enterprise through a VPN connection:

- Configure Device Mobility Info (DMI) with the IP subnets distributed or owned by the VPN concentrators.
- Associate the DMI with the same device pool that is used for devices co-located with the VPN concentrators. However, parameters such as calling privileges, network locale, and so forth, must be taken into consideration.
- Educate the remote site users to point to the geographically nearest enterprise VPN concentrator when making client-based or router-based VPN connections.

These guidelines ensure that call admission control is correctly applied on the enterprise WAN and over the connection to the remote site.

For information on deploying a VPN, refer to the various VPN design guides available under the *Security in WAN* subsection of the Design Zone for Security, available at:

http://www.cisco.com/en/US/solutions/ns340/ns414/ns742/ns744/landing_wan_security.html

Remote Enterprise Mobility High Availability

For remote site mobility environments, it is imperative that enterprise VPN services are configured and deployed in a redundant manner within the enterprise. This ensures that both client-based and router-based secure connections are highly available. If a VPN concentrator within the enterprise fails, a new secure connection can be set up with another VPN concentrator. Device registration and voice services are highly available in this type of deployment simply by virtue of the built-in Unified CM cluster node redundancy.

Capacity Planning for Remote Enterprise Mobility

The most critical scalability consideration for remote enterprise mobility environments is VPN concentrator capacity. Administrators must deploy sufficient VPN session capacity to accommodate all remote site connectivity, whether they are client-based or router-based secure tunnel connections. Failure to provide appropriate capacity will prevent some remote sites from connecting to the enterprise, thus eliminating access to even basic telephony services. Furthermore, just as with campus and multisite enterprise mobility deployments, it is important to provide sufficient device registration capacity within the enterprise to handle all remote user devices.

Design Considerations for Remote Enterprise Mobility

Consider the following design recommendations when enabling remote site connectivity for mobile users:

- When using Device Mobility, remember to configure Device Mobility Info (DMI) with the IP subnets distributed or owned by the VPN concentrators, and assign the DMI to the same device pool that is configured for devices deployed in the same location as the VPN concentrators.
- Educate remote site users to select the nearest VPN concentrator for VPN connection.
- Ensure appropriate VPN session capacity is available in order to provide connectivity to all remote site users.

Mobility Beyond the Enterprise

With Cisco's mobile Unified Communications, mobility users can handle calls to their enterprise directory number, not only on their desk phone, but also on one or more remote phones. Mobility users can also make calls from a remote phone as if they are dialing inside the enterprise. In addition, mobility users can take advantage of enterprise features such as hold, transfer, and conference as well as enterprise applications such as voicemail, conferencing, and presence on their mobile phones. This ensures continued productivity for users even when they are traveling outside the organization.

Further, with dual-mode phones that provide connectivity to the mobile voice and data provider network as well as the 802.11 WLAN, users not only have the ability to leverage enterprise applications while away from the enterprise, but they can also leverage the enterprise telephony infrastructure when inside the enterprise or remotely attached to the enterprise network to make and receive calls without incurring mobile voice network per-minute charges.

The fixed mobile convergence (FMC) mobility functionality delivered within the Cisco Unified Mobility solution is provided through Cisco Unified Communications Manager (Unified CM) and can be used in conjunction with Cisco mobile clients and devices such as Cisco Jabber and Cisco Cius.

Cisco Unified Mobility provides the following mobility application functionality:

- Mobile Connect

  Mobile Connect, also known as Single Number Reach, provides Cisco Unified Communications users with the ability to be reached at a single enterprise phone number that rings on both their IP desk phone and their mobile phone simultaneously. Mobile Connect users can pick up an incoming call on either their desk or mobile phones and at any point can move the in-progress call from one of these phones to the other without interruption.

- Mid-Call Features

  Mid-call features allow a user to invoke hold, resume, transfer, conferencing, and directed call park features from their mobile phone during in-progress mobility calls. These features are invoked from the mobile phone keypad and take advantage of enterprise media resources such as music on hold and conference bridges.

- Single Enterprise Voicemail Box

  Single Enterprise Voicemail box provides mobile voicemail avoidance capabilities and ensures that any unanswered calls made to the user's enterprise number and extended to the user's mobile phone will end up in the enterprise voicemail system rather than in a mobile voicemail system. This provides a single consolidated voicemail box for all business calls and eliminates the need for users to check multiple voicemail systems for messages.

- Mobile Voice Access and Enterprise Feature Access two-stage dialing

  Mobile Voice Access and Enterprise Feature Access two-stage dialing provide mobile users with the ability to make calls from their mobile phone as if they were calling from their enterprise IP desk phone. These features provide a cost savings in terms of toll charges for long distance or international calls as well as calls to internal non-DID extensions on the system that would not normally be reachable from outside the enterprise. These two-stage dialing features also provide the enterprise with an easy way to track phone calls made by users via a uniform and centrally located set of call detail records. Furthermore, these features provide the ability to mask a user's mobile phone number when sending outbound caller ID. Instead, the user's enterprise number is sent as caller ID. This ensures that returned calls to the user are made to the enterprise number, thus resulting in enterprise call anchoring.

Cisco mobile clients and devices provide the ability to attach to both the mobile provider network and 802.11 wireless networks for voice and data connectivity. This enables users to leverage both enterprise call control and in some cases mobile network call control from a single device. By leveraging the enterprise telephony infrastructure for making and receiving calls whenever possible and, in the case of dual-mode phones, falling back to the mobile voice network only when enterprise connectivity is unavailable, mobile clients and devices can help reduce telephony costs. Dual-mode phones and the clients that run on them also provide a handoff mechanism so that in-progress voice calls can be moved easily between the WLAN and mobile voice interfaces as a user moves in or out of the enterprise.

In addition to enabling mobile devices to make voice-over-IP calls via 802.11 WLAN or mobile data networks, Cisco mobile clients and devices also provide other unified communications services such as corporate directories access, presence and instant messaging (IM). These devices and clients enable mobile users to remain productive whether inside or outside the enterprise by providing access to collaboration applications while at the same time enabling users to make and receive enterprise calls from their mobile devices, whether outside the enterprise over public or private WiFi hot spots or the mobile data network, or inside the enterprise and over the WLAN network.

This section begins with a discussion of Unified Mobility features, functionality, and design and deployment considerations. Given the various benefits of Unified Mobility and the fact that mobile clients and devices can be integrated to take advantage of the features provided, this discussion paves the way for examination of mobile client applications such as Cisco Jabber. Following the mobile client and device discussion, this section includes a discussion of architecture, functionality, and design and deployment implications for the following mobility applications and features:

- Cisco Unified Mobility, page 25-32
- Cisco Mobile Clients and Devices, page 25-60

Cisco Unified Mobility

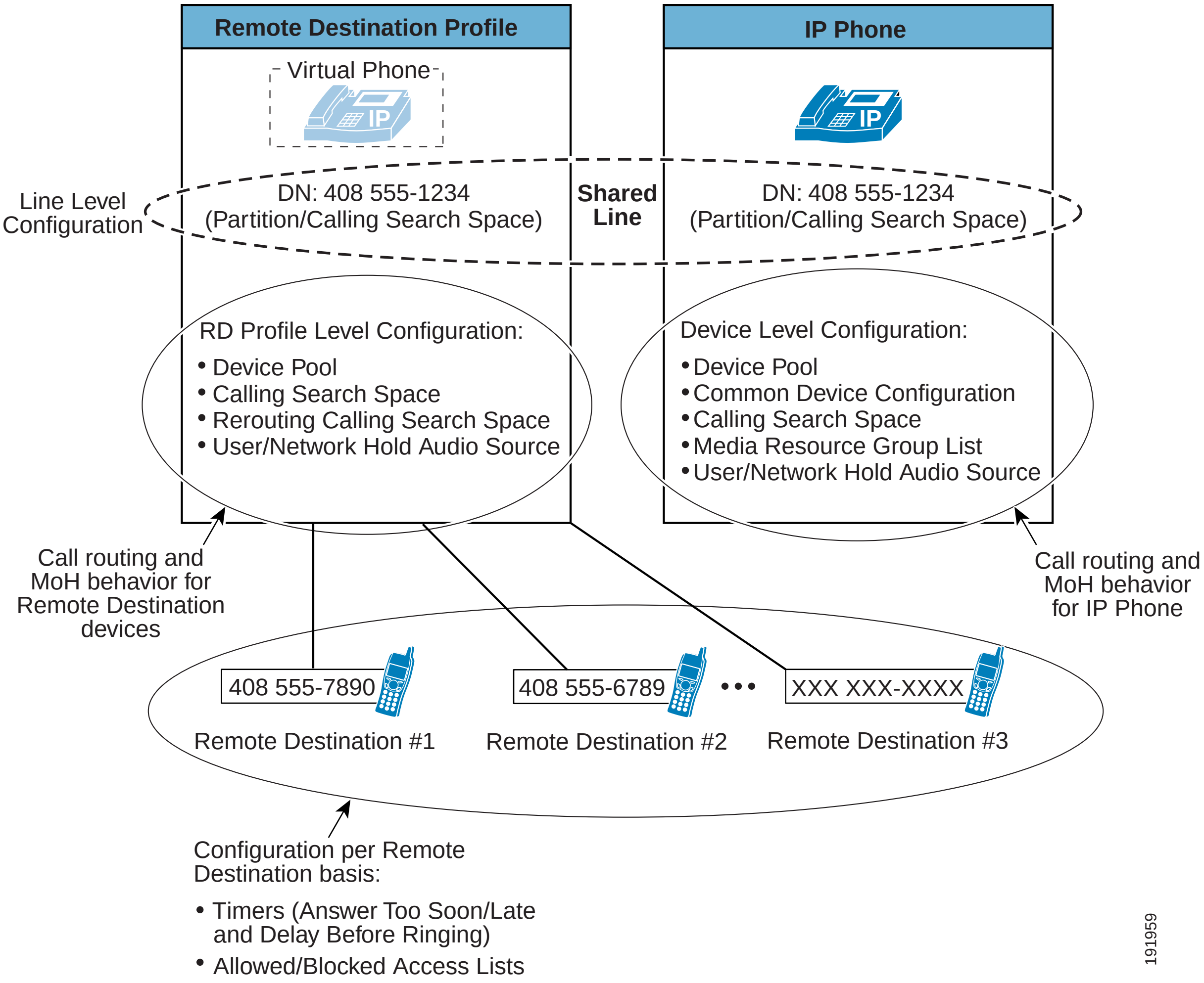
Cisco Unified Mobility refers to the native mobility functionality within the Cisco Unified Communications Manager (Unified CM) and includes the Mobile Connect, Mobile Voice Access, and Enterprise Feature Access features.

Unified Mobility functionality depends on the appropriate configuration of Unified CM. For this reason, it is important to understand the nature of this configuration as well as the logical components.

Figure 25-12 illustrates the configuration requirements for Unified Mobility. First, as for all users, a mobility user's enterprise phone is configured with appropriate line-level settings such as directory number, partition, and calling search space. In addition, the device-level settings of the enterprise phone include parameters such as device pool, common device configuration, calling search space, media resource group list, and user and network hold audio sources. All of these line and device settings on the user's enterprise phone affect the call routing and music on hold (MoH) behavior for incoming and outgoing calls.

Next, a remote destination profile must be configured for each mobility user in order for them to take advantage of Unified Mobility features. The remote destination profile is configured at the line level with the same directory number, partition, and calling search space as the user's enterprise phone line. This results in a shared line between the remote destination profile and the enterprise phone. The remote destination profile configuration includes device pool, calling search space, rerouting calling search space, and user and network hold audio source parameters. The remote destination profile should be thought of as a virtual phone whose configuration mirrors the user's line-level enterprise phone settings, but whose profile-level configuration combined with the line-level settings determines the call routing and MoH behavior that the user's remote destination phone will inherit. The user's enterprise directory number, which is shared between the remote destination profile and the enterprise phone, allows calls to that number to be extended to the user's remote destination.

Figure 25-12 Cisco Unified Mobility Configuration Architecture


Remote Destination Profile
Virtual Phone
IP
IP Phone
IP
Line Level Configuration
DN: 408 555-1234
(Partition/Calling Search Space)
Shared Line
DN: 408 555-1234
(Partition/Calling Search Space)
RD Profile Level Configuration:
• Device Pool
• Calling Search Space
• Rerouting Calling Search Space
• User/Network Hold Audio Source
Device Level Configuration:
• Device Pool
• Common Device Configuration
• Calling Search Space
• Media Resource Group List
• User/Network Hold Audio Source
Call routing and MoH behavior for Remote Destination devices
Call routing and MoH behavior for IP Phone
408 555-7890
408 555-6789
XXX XXX-XXXX
Remote Destination #1
Remote Destination #2
Remote Destination #3
Configuration per Remote Destination basis:
• Timers (Answer Too Soon/Late and Delay Before Ringing)
• Allowed/Blocked Access Lists
191959

As further shown in Figure 25-12, a mobility user can have one or more remote destinations configured and associated with their remote destination profile. A remote destination represents a single PSTN phone number where a user can be reached. A user can have up to 10 remote destinations defined. Call routing timers can be configured for each remote destination to adjust the amount of time a call will be extended to a particular remote phone, as well as the amount of time to wait before extending the call and the amount of time that must pass before a call can be answered at the remote phone. Mobility users can also configure filters for each remote destination to allow or deny calls from certain phone numbers to be extended to that remote phone.

Note Cisco Business Edition supports a maximum of four remote destinations per mobility user.

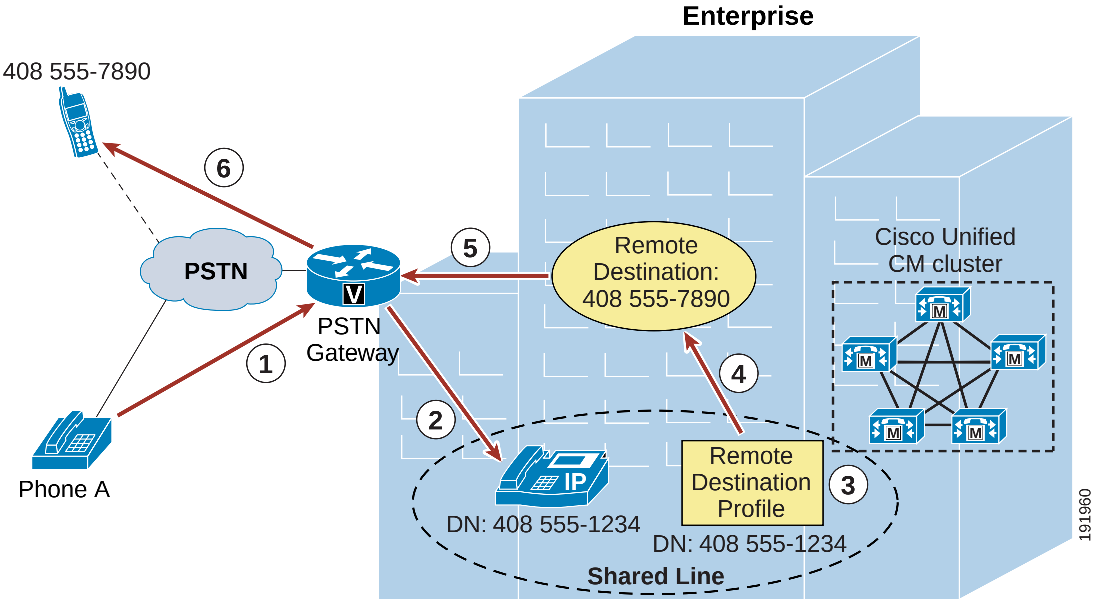
Mobile Connect

The Mobile Connect feature allows an incoming call to an enterprise user to be offered to the user's IP desk phone as well as up to 10 configurable remote destinations. Typically a user's remote destination is their mobile or cellular telephone. Once the call is offered to both the desktop and remote destination phone(s), the user can answer at any of those phones. Upon answering the call on one of the remote destination phones or on the IP desk phone, the user has the option to hand off or pick up the call on the other phone.

Mobile Connect Functionality

Figure 25-13 illustrates a basic Mobile Connect call flow. In this example, Phone A on the PSTN calls a Mobile Connect user's enterprise directory number (DN) 408-555-1234 (step 1). The call comes into the enterprise PSTN gateway and is extended through Unified CM to the IP phone with DN 408-555-1234 (step 2), and this phone begins to ring. The call is also extended to the user’s Remote Destination Profile, which shares the same DN (step 3). In turn, a call is placed to the remote destination associated with the user's remote destination profile (in this case 408-555-7890) (step 4). The outgoing call to the remote destination is routed through the PSTN gateway (step 5). Finally the call rings at the remote destination PSTN phone with number 408 555-7890 (step 6). The call can then be answered at either phone.

Figure 25-13 Mobile Connect



Typically a Mobile Connect user's configured remote destination is their mobile phone on a mobile voice or cellular provider network; however, any destination reachable by means of the PSTN can be configured as a user's remote destination. Furthermore, a Mobile Connect user can have up to 10 remote destinations configured, so an incoming call could potentially ring as many as 10 PSTN phones as well as the user's desk phone. Once the call is answered at the desk phone or at a remote destination phone, any other call legs that have been extended to ring additional remote destinations or the desk phone (if not answered at the desk phone) will be cleared. If the incoming call is answered at the remote destination, the voice media path will be hairpinned within the enterprise PSTN gateway utilizing two gateway ports. This utilization must be considered when deploying the Mobile Connect feature.

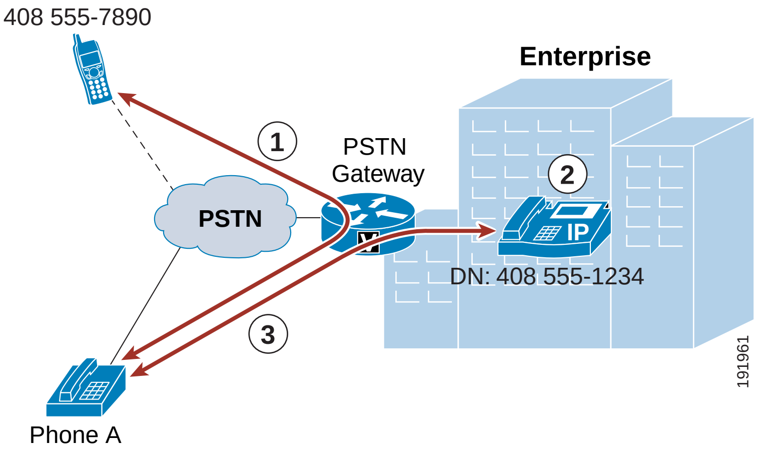
Note Mobility users on a Cisco Business Edition system can have a maximum of four remote destinations.

Note In order for Mobile Connect to work as in Figure 25-13, ensure that the user-level Enable Mobility check box under the End User configuration page has been checked and that at least one of the user's configured remote destinations has the Enable Mobile Connect check box checked.

Desk Phone Pickup

As illustrated in Figure 25-14, once a user answers a Mobile Connect call at the remote destination device (step 1: in this case, 408 555-7890), at any point the user can hang up the call at the remote destination and pick it up again at their desk phone by simply pressing the Resume softkey on the desk phone (step 2: at DN 408 555-1234 in this case). The call resumes between the original caller at Phone A and the desk phone (step 3).

Figure 25-14 Desk Phone Pickup



Desk phone pickup can be performed whenever an enterprise-anchored call is in progress at a configured remote destination phone and that phone hangs up the call.

Note An enterprise-anchored call refers to any call that has at least one call leg connected through an enterprise PSTN gateway and that originated either from a remote destination to an enterprise DID or from Mobile Connect, Mobile Voice Access, Enterprise Feature Access, or Intelligent Session Control.

The option to pick up or resume the call at the desk phone is available for a certain amount of time. For this reason, it is good practice for the Mobile Connect user to ensure that the calling phone hangs up before the remote destination phone is hung up. This ensures that the call cannot be resumed at the desk phone by someone else. By default, the call remains available for pickup at the desk phone for 10 seconds after the remote destination phone hangs up; however, this time is configurable and can be set from 0 to 30000 milliseconds on a per-user basis by changing the Maximum Wait Time for Desk Pickup parameter under the End User configuration page. Desk phone pickup can also be performed after invoking the mid-call hold feature at the remote destination phone. However, in these cases, the Maximum Wait Time for Desk Pickup parameter setting has no effect on the amount of time the call will be available for pickup. A call placed on mid-call hold will remain on hold and be available for desk phone pickup until manually resumed at either the remote or desktop phone.

Another method for performing desk phone pickup is to use the mid-call session handoff feature. This mid-call feature is invoked by manually keying *74, the default enterprise feature access code for session handoff, which in turn generates a DTMF sequence back to Unified CM. When this feature is invoked, Unified CM sends a new call to the user's enterprise desk phone. Once this new call is flashing or ringing at the desk phone, the user then must answer the call to complete the session handoff.

The benefit of this desk phone pickup method over other methods (such as hanging up the call at the mobile phone or using the mid-call hold feature) is that the conversation between the user and the far-end phone is maintained throughout the handoff process. Once the *74 sequence has been keyed, the user can continue the conversation because the handoff call is sent to the user's desk phone. When the user answers the call at the desk phone, the call legs are shuffled so that the call leg to the far-end is connected to the new call leg created at the desk phone, thus resulting in an uninterrupted or near-instantaneous cut-through of the audio path. The original call leg at the mobile device is subsequently cleared.

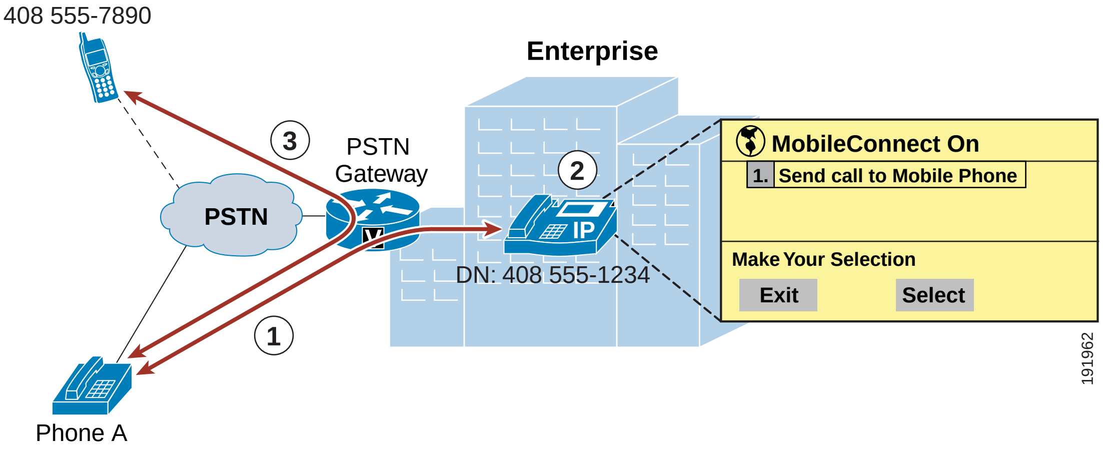
Unlike the hang-up method for invoking desk phone pickup, where the end-user’s Maximum Wait Time for Desk Pickup setting determines how long the call will be available for pickup at the desk phone, with session handoff the Session Handoff Alerting Timer service parameter determines the amount of time the call will ring or flash at the desk phone before the handoff call is cleared. The default handoff alerting time is 10 seconds. Further, with session handoff, any call forward settings configured on the desk phone do not get invoked. As a result, the handoff feature does not forward to voicemail or any other call-forward destination. If a call is not answered by the end of Session Handoff Alerting Timer period, then the call is cleared and the Remote In Use state is removed from the user's desk phone line. However, in this scenario the original call is maintained at the mobile phone.

For additional information about session handoff and other mid-call features, see Mid-Call Features, page 25-37.

Remote Destination Phone Pickup

Figure 25-15 illustrates Mobile Connect remote destination phone pickup functionality. Assuming Phone A calls the Mobile Connect user's enterprise DN 408 555-1234 and the call is answered at the user's desk phone and is in progress (step 1), the user must push the Mobility softkey. Assuming the Mobile Connect feature is enabled for this phone and remote destination pickup is available, the user presses the Select softkey (step 2). A call is generated to the user's remote destination phone (in this case, 408 555-7890), and the remote phone begins to ring. Once the call is answered at the remote phone, the call resumes between Phone A and the Mobile Connect user's remote phone with number 408 555-7890 (step 3).

Figure 25-15 Remote Destination Phone Pickup



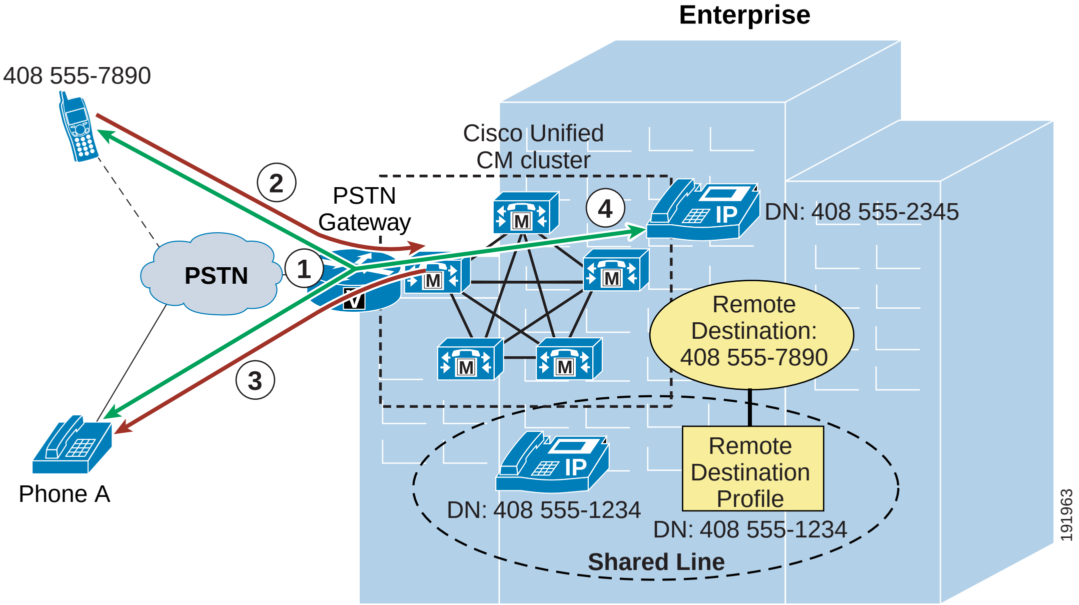
When a Mobile Connect user has multiple remote destinations configured, each remote destination will ring when the Select softkey is pressed, and the user can answer the desired phone.

Note In order for remote destination phone pickup to work as in Figure 25-15, ensure that at least one of the user's configured remote destinations has the Mobile Phone check box checked. In addition, the Mobility softkey must be configured for all mobility users by adding the softkey to each user's associated desk phone softkey template. Failure to check the Mobile Phone check box and to make the Mobility softkey available to mobility users will prevent the use of remote destination phone pickup functionality.

Mid-Call Features

As illustrated in Figure 25-16, once a user answers a Mobile Connect call at the remote destination device (step 1: in this case, 408 555-7890), the user can invoke mid-call features such as hold, resume, transfer, conference, directed call park, and session handoff by sending DTMF digits from the remote destination phone to Unified CM via the enterprise PSTN gateway (step 2). When the mid-call feature hold, transfer, conference, or directed call park is invoked, MoH is forwarded from Unified CM to the held party (step 3: in this case, Phone A). In-progress calls can be transferred to another phone or directed call park number, or additional phones can be conferenced using enterprise conference resources (step 4).

Figure 25-16 Mobility Mid-Call Feature



Mid-call features are invoked at the remote destination phone by a series of DTMF digits forwarded to Unified CM. Once received by Unified CM, these digit sequences are matched to the configured Enterprise Feature Access Codes for Hold, Exclusive Hold, Resume, Transfer, Conference, and Session Handoff, and the appropriate function is performed.

Note To enable the Directed Call Park mid-call feature, you must configure Cisco Unified CM with directed call park numbers and call park retrieval prefixes.

Note In order to perform the transfer, conference, and directed call park mid-call features, a second call leg is generated by the remote destination phone to a system-configured Enterprise Feature Access DID that answers the call, takes user input (including PIN number, mid-call feature access code, and target number), and then creates the required call leg to complete the transfer, conference, or directed call park operation.

With the mid-call session handoff feature, MoH is not forwarded to the far-end because the far-end is never placed on hold. Instead, the original audio path is maintained until the mobile user answers the handoff call at the desk phone. Once the call is answered, the call legs are shuffled at the enterprise gateway and the audio path is maintained.

Mid-call features are invoked by manually keying the feature access codes and entering the appropriate key sequences. Table 25-2 indicates the required key sequences for invoking mid-call features.

Table 25-2 Manual Mid-Call Feature Key Sequences

Mid-Call Feature	Enterprise Feature Access Code (default)	Manual Key Sequence
Hold	*81	Enter: *81
Exclusive Hold	*82	Enter: *82
Resume	*83	Enter: *83
Transfer	*84	1. Enter: *82 (Exclusive Hold) 2. Make new call to Enterprise Feature Access DID. 3. On connect, enter: *<PIN_number>* # *84 # *<Transfer_Target/DN>* # 4. Upon answer by transfer target (for consultive transfer) or upon ringback (for early attended transfer), enter: *84
Directed Call Park	N/A	1. Enter: *82 (Exclusive Hold) 2. Make new call to Enterprise Feature Access DID. 3. On connect, enter: *<PIN_number>* # *84 # *<Directed_Call_Park_Number>* # *84 # **Note** To retrieve a parked call, the user must use Mobile Voice Access or Enterprise Feature Access Two-Stage Dialing to place a call to the directed call park number. When entering the directed call park number to be dialed, it must be prefixed with the appropriate call park retrieval prefix.
Conference	*85	1. Enter: *82 (Exclusive Hold) 2. Make new call to Enterprise Feature Access DID. 3. On connect enter: *<PIN_number>* # *85 # *<Conference_Target/DN>* # 4. Upon answer by conference target, enter: *85
Session Handoff	*74	1. Enter: *74 2. Answer at the desk phone upon ring and/or flash.

Note Media resource allocation for mid-call features such as hold and conference is determined by the Remote Destination Profile configuration or, in the case of dual-mode phones and Unified Mobile Communicator, the device configuration. The media resource group list (MRGL) of the device pool configured for the Remote Destination Profile or the mobile client device is used to allocate a conference bridge for the conferencing mid-call feature. The User Hold Audio Source and Network Hold MoH Audio Source settings of the Remote Destination Profile or the mobile client device, in combination with the media resource group list (MRGL) of the device pool, is used to determine the appropriate MoH stream to be sent to a held device.

Mobile Voicemail Avoidance with Single Enterprise Voicemail Box

An additional consideration with Cisco Unified Mobility Mobile Connect is mobile voicemail avoidance. The single enterprise voicemail box feature ensures that all unanswered enterprise business calls end up at the enterprise voicemail system. This prevents a user from having to check multiple mailboxes (enterprise, mobile, home, and so forth) for calls to their enterprise phone number that are unanswered. This feature provides two methods for avoiding mobile or non-enterprise voicemail:

- Timer Control method — With this method the system relies on a set of timers (one per remote destination) in conjunction with system call-forward timers to ensure that, when and if a call is forwarded to a voicemail system on ring-no-answer, the enterprise voicemail system receives the call.
- User Control method — With this method the system relies on a DTMF confirmation tone from the remote destination when the call is answered to determine if the call was received by the user or a non-enterprise voicemail system.

System settings determine whether the timer control or user control method is used. The method used can be set globally via the Voicemail Selection Policy service parameter or for individual remote destinations via the Single Number Reach Voicemail Policy. By default the system and all remote destinations use the timer control method

Timer Control Mobile Voicemail Avoidance

For this method, the system relies on a set of timers on the Remote Destination configuration page. The purpose of these timers is to ensure that, when and if a call is forwarded to a voicemail system on ring-no-answer, the call is forwarded to the enterprise voicemail system rather than any remote destination voicemail system. These timers in conjunction with other system forward-no-answer timers should be configured to avoid non-enterprise voicemail systems as follows:

- Ensure the system forward-no-answer time is shorter at the desk phone than at the remote destination phones.

  To do so, ensure that the global Forward No Answer Timer field in Unified CM or the No Answer Ring Duration field under the individual phone line is configured with a value that is less than the amount of time a remote destination phone will ring before forwarding to the mobile voicemail system. In addition, the Delay Before Ringing Timer parameter under the Remote Destination configuration page can be used to delay the ringing of the remote destination phone in order to further lengthen the amount of time that must pass before a remote destination phone will forward to its own mobile voicemail box. However, when adjusting the Delay Before Ringing Timer parameter, take care to ensure that the global Unified CM Forward No Answer Timer (or the line-level No Answer Ringer Duration field) is set sufficiently high enough so that the mobility user has time to answer the call on the remote destination phone. The Delay Before Ringing Timer parameter can be set for each remote destination and is set to 4,000 milliseconds by default.

- Ensure that the remote destination device stops ringing before the incoming call is forwarded to the mobile voicemail system.

  You can accomplish this with the Answer Too Soon and Answer Too Late timers for each remote destination. First the Answer Too Soon Timer parameter under the Remote Destination configuration page should be configured with a value that is more than the amount of time it takes a call extended to a powered-off or out-of-range mobile phone to be forwarded to the mobile voicemail system. By default this timer is set 1,500 milliseconds (or 1.5 seconds). If the call is answered before the Answer Too Soon Timer expires, the system will disconnect the call leg to the remote destination. This ensures that calls forwarded immediately to the mobile voicemail system will not be connected, but those answered by the user after ring-in are connected.

Next configure the Answer Too Late Timer parameter under the Remote Destination configuration page with a value that is less than the amount of time that a remote destination phone will ring before forwarding to its voicemail box. By default this timer is set to 19,00 milliseconds (or 19 seconds). If the call is not answered before this timer expires, the system will disconnect the call leg to the remote destination. This ensures that the remote destination phone stops ringing before the call is forwarded to the mobile voicemail system.

Note Incoming calls to a remote destination that are manually diverted by the mobility user can end up in the mobile voicemail box if the manual diversion occurs after the Answer Too Soon timer has expired. To prevent this from happening, mobility users should be configured for the user control method or advised to ignore or silence the ringing of incoming calls they wish to divert to voicemail. This will ensure that unanswered calls always end up in the enterprise voicemail system.

Note In most deployment scenarios, the default Delay Before Ringing Timer, Answer Too Late Timer, and Answer Too Soon Timer values are sufficient and do not need to be changed.

User Control Mobile Voicemail Avoidance

For this method, the system relies on DTMF confirmation tone from the remote destination when the call is answered. If a DTMF tone is received by the system, then the system knows that the user answered the call and pressed a key to generate the DTMF tone. On the other hand, if the DTMF tone is not received by the system, the system assumes the call leg was answered by a non-enterprise voicemail system and it disconnects the call leg.

When the user control method is enabled, on answer the end user will hear an audio prompt requesting that they press a key pad button to generate a DTMF tone. By default the audio prompt is played to the user one second after the call is answered. The user may not hear the audio prompt if they press the keypad to generate a DTMF tone immediately upon answering. The audio prompt is played only on the remote destination call leg and therefore the far-end party will not hear this prompt. Once the audio prompt is played to the user, by default the system will wait 5 seconds to receive the DTMF tone. If the tone is not received, the system disconnects the call leg but continues to ring the user's other configured devices until the call is answered by the user or forwarded to the enterprise voicemail system.

Note The user control mobile voicemail avoidance method is completely dependent on successful relay of the DTMF tone from the remote destination on the mobile voice network or PSTN all the way to Unified CM. The DTMF tone must be sent out-of-band to Unified CM. If DTMF relay is not properly configured on the network and system, DTMF will not be received and all call legs to remote destinations relying on the user control method will be disconnected. The system administrator should ensure proper DTMF interoperation and relay across the enterprise telephony network prior to enabling the user control method. If DTMF cannot be effectively relayed from the PSTN to Unified CM, then the timer control mobile voicemail avoidance method should be used instead.

Enabling and Disabling Mobile Connect

The Mobile Connect feature can be enabled or disabled by using one of the following methods:

- Cisco Unified CM Administration or Cisco Unified CM User Options pages

  An administrator or user unchecks the Mobile Connect box to disable, or checks the Mobile Connect box to enable, the feature. This is done per remote destination.

- Mobile Voice Access or Enterprise Feature Access

  A Mobility-enabled user dials into the Mobile Voice Access or Enterprise Feature Access DID and, after entering appropriate credentials, enters the digit 2 to enable or 3 to disable. With Mobile Voice Access, the user is prompted to enable or disable Mobile Connect for a single remote destination or all of their remote destinations. With Enterprise Feature Access, the user can enable or disable Mobile Connect only for the remote destination device from which they are calling.

- Desk phone Mobility softkey

  The user presses the Mobility softkey when the phone is in the on-hook state and selects either Enable Mobile Connect or Disable Mobile Connect. With this method, Mobile Connect is enabled or disabled for all of the user's remote destinations.

Access Lists for Allowing or Blocking Mobile Connect Calls

Access lists can be configured within Cisco Unified CM and associated to a remote destination. Access lists are used to allow or block inbound calls (based on incoming caller ID) from being extended to a mobility-enabled user's remote destinations. Furthermore, these access lists are invoked based on the time of day.

Access lists are configured for mobility-enabled users as either blocked or allowed. Access lists contain one or more members or filters consisting of a specific number or number mask, and the filters are compared against the incoming caller ID of the calling party. In addition to containing specific number strings or number masks for matching caller ID, access lists can also contain a filter for incoming calls where the caller ID is not available or is set to private. A blocked access list contains an implicit "allow all" at the end of the list so that calls from any numbers entered in the access list will be blocked but calls from all other numbers will be allowed. An allowed access list contains an implicit "deny all" at the end of the list so that calls from any numbers entered in the access list will be allowed but calls from all other numbers will be blocked.

Once configured access lists are associated with a configured Ring Schedule under the Remote Destination configuration screen, the configured Ring Schedule in combination with the selected access list provides time-of-day call filtering for Mobile Connect calls on a per-remote-destination basis. Access lists and Ring Schedules can be configured and associated to a remote destination by an administrator using the Cisco Unified CM Administration interface or by an end user using the Cisco Unified CM User Options interface.

Mobile Connect Architecture

The architecture of the Mobile Connect feature is as important to understand as its functionality. Figure 25-17 depicts the message flows and architecture required for Mobile Connect. The following sequence of interactions and events can occur between Unified CM, the Mobile Connect user, and the Mobile Connect user's desk phone:

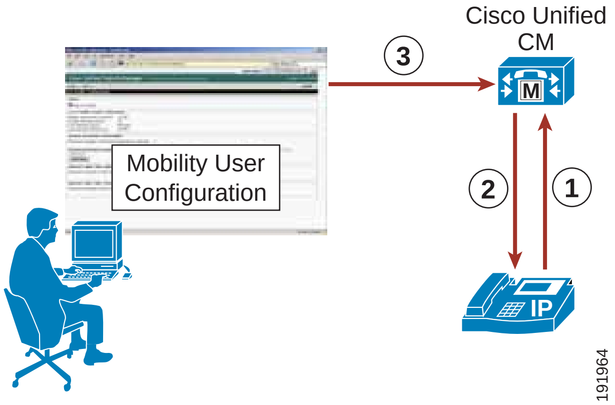
1. The Mobile Connect phone user who wishes to either enable or disable the Mobile Connect feature or to pick up an in-progress call on their remote destination phone pushes the Mobility softkey on their desk phone (see step 1 in Figure 25-17).

2. Unified CM returns the Mobile Connect status (On or Off) and offers the user the ability to select the Send Call to Mobile Phone option when the phone is in the Connected state, or it offers the user the ability to enable or disable the Mobile Connect status when the phone is in the On Hook state (see step 2 in Figure 25-17).
3. Mobile Connect users can use the Unified CM User Options interface to configure their own mobility settings via the web-based configuration pages at

   http://<*Unified-CM_Server_IP_Address*>/ccmuser/

   where <*Unified-CM_Server_IP_Address*> is the IP address of the Unified CM publisher server (see step 3 in Figure 25-17).

Figure 25-17 Mobile Connect Architecture



High Availability for Mobile Connect

The Mobile Connect feature relies on the following components:

- Unified CM servers
- PSTN gateway

Each component must be redundant or resilient in order for Mobile Connect to continue functioning fully during various failure scenarios.

Unified CM Server Redundancy

The Unified CM server is required for the Mobile Connect feature. Unified CM server failures are non-disruptive to Mobile Connect functionality, assuming phone and gateway registrations are made redundant using Unified CM Groups.

In order for Mobile Connect users to use the Unified CM User Options web interface to configure their mobility settings (remote destinations and access lists), the Unified CM publisher server must be available. If the publisher is down, users will not be able to change mobility settings. Likewise, administrators will be unable to make mobility configuration changes to Unified CM; however, existing mobility configurations and functionality will continue. Finally, changes to Mobile Connect status must be written by the system on the Unified CM publisher server; if the Unified CM publisher is unavailable, then enabling or disabling Mobile Connect will not be possible.

PSTN Gateway Redundancy

Because the Mobile Connect feature relies on the ability to extend additional call legs to the PSTN to reach the Mobile Connect users' remote destination phones, PSTN gateway redundancy is important. Should a PSTN gateway fail or be out of capacity, the Mobile Connect call cannot complete.   Typically, enterprise IP telephony dial plans provide redundancy for PSTN access by providing physical gateway redundancy and call re-routing capabilities as well as enough capacity to handle expected call activity. Assuming that Unified CM has been configured with sufficient capacity, multiple gateways, and route group and route list constructs for call routing resiliency, the Mobile Connect feature can rely on this redundancy for uninterrupted functionality.

Mobile Voice Access and Enterprise Feature Access

Mobile Voice Access (also referred to as System Remote Access) and Enterprise Feature Access two-stage dialing are features built on top of the Mobile Connect application. Both features allow a mobility-enabled user who is outside the enterprise to make a call as though they are directly connected to Unified CM. This functionality is commonly referred to as Direct Inward System Access (DISA) in traditional telephony environments. These features benefit the enterprise by limiting toll charges and consolidating phone billing directly to the enterprise rather than billing to each mobile user. In addition, these features allow the users to mask their mobile phone or remote destination numbers when sending outbound caller ID. Instead, the user's enterprise directory number is sent as caller ID. This ensures that returned calls to the user are made to the enterprise number, thus resulting in enterprise call anchoring. These features also enable mobile users to dial internal extensions or non-DID enterprise numbers that would not normally be reachable from outside the enterprise.

Mobile Voice Access is accessed by calling a system-configured DID number that is answered and handled by an H.323 or SIP VoiceXML (VXML) gateway. The VoiceXML gateway plays interactive voice response (IVR) prompts to the Mobile Voice Access user, requesting user authentication and input of a number to be dialed via the user phone keypad.

Enterprise Feature Access functionality includes the previously discussed mid-call transfer and conference features as well as two-stage dialing functionality. Two-stage dialing works the same way as Mobile Voice Access, but without the IVR prompts. The system-configured Enterprise Feature Access DID is answered by Unified CM. The user then uses the phone keypad or Smart Phone softkeys to input authentication and the number to be dialed. These inputs are received without prompts.

With both the Mobile Voice Access and Enterprise Feature Access two-stage dialing features, once the call to the input number is connected, users can invoke mid-call features or pick up the call on their desk phones just as with a Mobile Connect calls. This is possible because the call is anchored at the enterprise gateway.

Mobile Voice Access IVR VoiceXML Gateway URL

The Mobile Voice Access feature requires the Unified CM VoiceXML application to reside on the H.323 or SIP gateway. The URL used to load this application is:

http://<*Unified-CM-Publisher_IP-Address*>:8080/ccmivr/pages/IVRMainpage.vxml

where <*Unified-CM-Publisher_IP-Address*> is the IP address of the Unified CM publisher node.

Mobile Voice Access Functionality

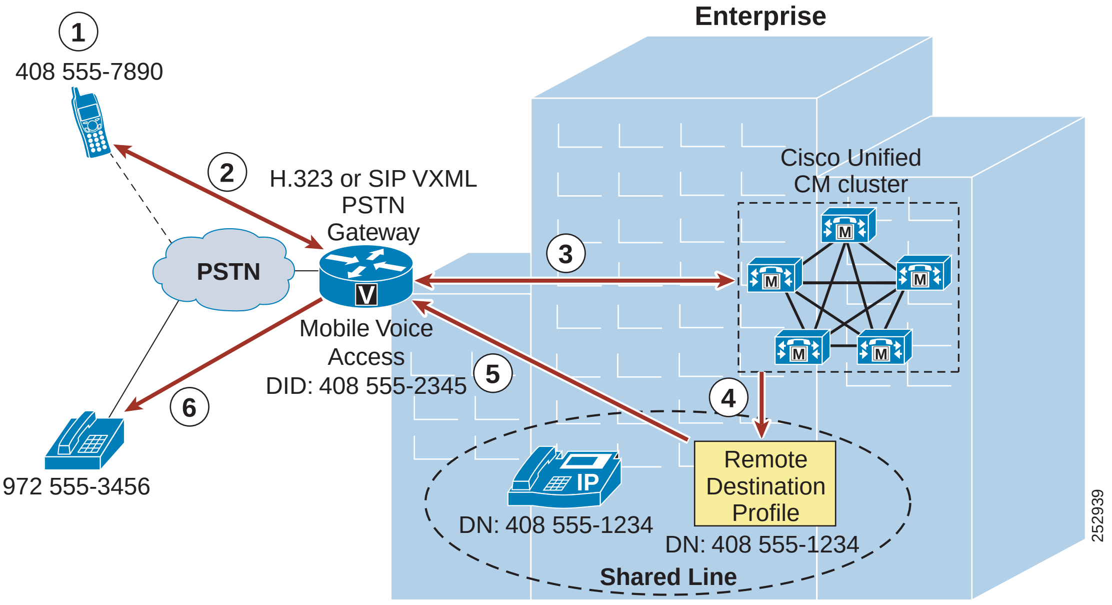
Figure 25-18 illustrates a Mobile Voice Access call flow. In this example, the Mobile Voice Access user on PSTN phone 408 555-7890 dials the Mobile Voice Access enterprise DID DN 408-555-2345 (step 1).

The call comes into the enterprise PSTN H.323 or SIP gateway, which also serves as the VoiceXML gateway. The user is prompted via IVR to enter their numeric user ID (followed by the # sign), PIN number (followed by the # sign), and then a 1 to make a Mobile Voice Access call, followed by the phone number they wish to reach. In this case, the user enters 9 1 972 555 3456 as the number they wish to reach (followed by the # sign) (step 2).

Note If the PSTN phone from which the Mobile Voice Access user is calling is configured as a Mobile Connect remote destination for that user and the incoming caller ID can be matched against this remote destination by Unified CM, the user does not have to enter their numeric user ID. Instead they will be prompted to enter just the PIN number.

In the meantime, Unified CM has forwarded IVR prompts to the gateway, the gateway has played these prompts to the user, and the gateway has collected user input including the numeric ID and PIN number of the user. This information is forwarded to Unified CM for authentication and to generate the call to 9 1 972 555 3456 (step 3). After authenticating the user and receiving the number to be dialed, Unified CM generates a call via the user's Remote Destination Profile (step 4). The outbound call to 972 555-3456 is routed via the PSTN gateway (step 5). Finally, the call rings at the PSTN destination phone with number 972 555-3456 (step 6).

Figure 25-18 Mobile Voice Access



Note In order for Mobile Voice Access to work as in Figure 25-18, ensure that the system-wide Enable Mobile Voice Access service parameter is set to True and that the per-user Enable Mobile Voice Access check box on the End User configuration page is also checked.

Note The Mobile Voice Access feature relies on the Cisco Unified Mobile Voice Access Service, which must be activated manually from the Unified CM Serviceability configuration page. This service can be activated on the publisher node only.

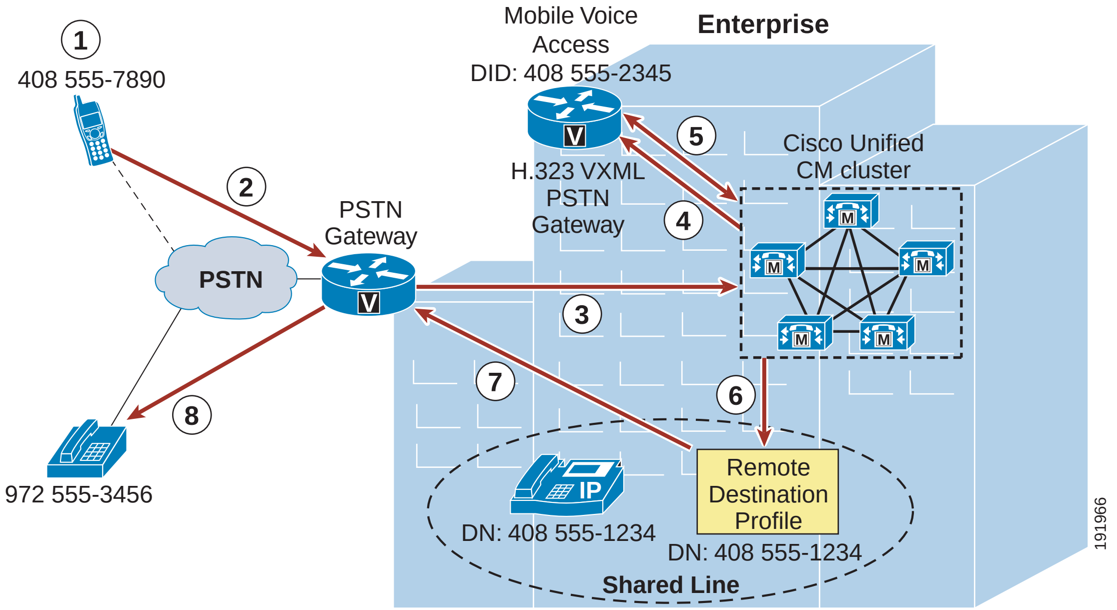
Mobile Voice Access Using Hairpinning

In deployments where the enterprise PSTN gateways are not using H.323 or SIP, Mobile Voice Access functionality can still be provided using hairpinning on a separate gateway running H.323. Mobile Voice Access using hairpinning relies on off-loading the VoiceXML functionality to a separate H.323 gateway. Figure 25-19 illustrates a Mobile Voice Access call flow using hairpinning. In this example, just as in the previous example, the Mobile Voice Access user on PSTN phone 408 555-7890 dials the Mobile Voice Access enterprise DID DN 408-555-2345 (step 1). The call comes into the enterprise PSTN gateway (step 2) and is forwarded to Unified CM for call handling (step 3). Unified CM next routes the inbound call to the H.323 VoiceXML gateway (step 4). The user is then prompted by IVR to enter their numeric user ID, PIN, and then a 1 to make a Mobile Voice Access call, followed by the phone number they wish to reach. Again the user enters 9 1 972 555 3456 as the number they wish to reach (followed by the # sign).

Note When using Mobile Voice Access with hairpinning, users calling into the system will not be identified automatically by their caller ID. Instead, users will have to key in their remote destination number manually prior to entering their PIN. The reason the user is not automatically identified is that, for hairpinning deployments, the PSTN gateway must first route the call to Unified CM to reach the hairpinned Mobile Voice Access gateway. Because the call is routed to Unified CM first, the conversion of the calling number from a mobile number to an enterprise directory number occurs prior to the call being handled by the Mobile Voice Access gateway. This results in the Mobile Voice Access gateway being unable to match the calling number with a configured remote destination, and therefore the system prompts the user to enter their remote destination number. This is unique to hairpinning deployments; with normal Mobile Voice Access flows, the PSTN gateway does not have to route the call to Unified CM first in order to access Mobile Voice Access because the functionality is available on the local gateway.

In the meantime, the H.323 VoiceXML gateway collects and forwards the user input to Unified CM and then plays the forwarded IVR prompts to the PSTN gateway and the Mobile Voice Access user. Unified CM in turn receives user input, authenticates the user, and forwards appropriate IVR prompts to the H.323 VoiceXML gateway based on user input (step 5). After receiving the number to be dialed, Unified CM generates a call using the user's Remote Destination Profile (step 6). The outbound call to 972 555-3456 is routed through the PSTN gateway (step 7). Finally, the call rings at the PSTN destination phone with number 972 555-3456 (step 8).

Figure 25-19 Mobile Voice Access Using Hairpinning



Note When deploying Mobile Voice Access in hairpinning mode, Cisco recommends configuring the Mobile Voice Access DID at the PSTN gateway and the Mobile Voice Access Directory Number within Cisco Unified CM (under **Media Resources** > **Mobile Voice Access**) as different numbers. A translation pattern within Unified CM can then be used to translate the called number of the Mobile Voice Access DID to the configured Mobile Voice Access directory number. Because the Mobile Voice Access directory number configured within Unified CM is visible to the administrator only, translation between the DID and directory number will be invisible to the end user and there will be no change in end-user dialing behavior. This is recommended in order to prevent mobility call routing issues in multi-cluster environments. This recommendation does not apply to Mobile Voice Access in non-hairpinning mode.

Note Mobile Voice Access in hairpinning mode is supported only with H.323 VXML gateways.

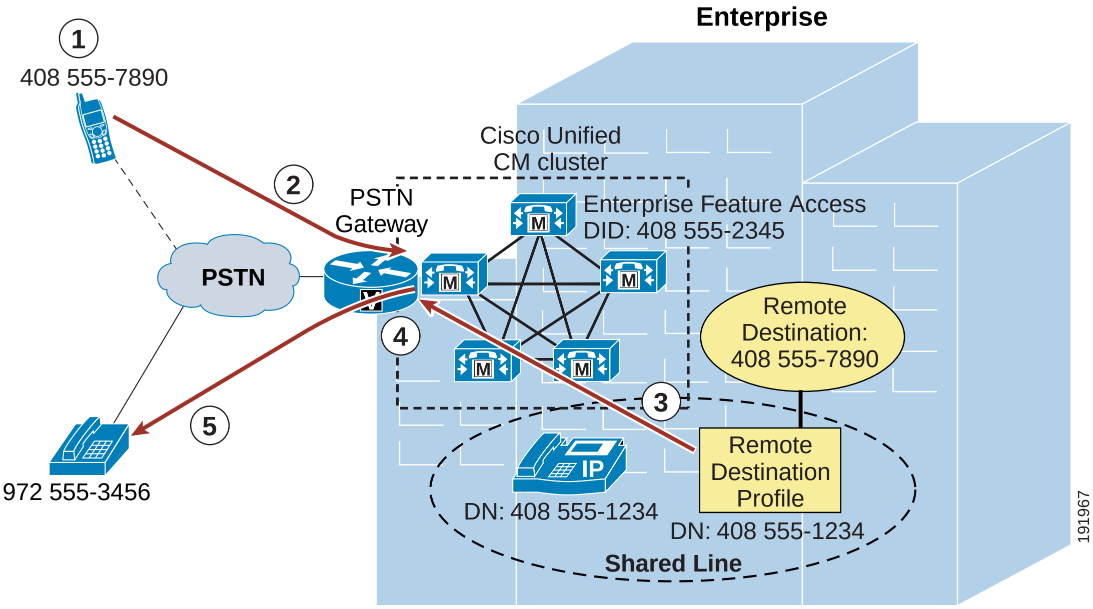
Enterprise Feature Access with Two-Stage Dialing Functionality

Figure 25-20 illustrates the call flow for Enterprise Feature Access two-stage dialing. In this example, the mobility user at remote destination phone 408 555-7890 dials the Enterprise Feature Access DID 408 555-2345 (step 1). Once the call is connected, the remote destination phone is used to send DTMF digits to Unified CM via the PSTN gateway, beginning with the user's PIN (followed by the # sign) which is authenticated with Unified CM. Next a 1 (followed by the # sign) is sent to indicate a two-stage dialed call is being attempted, followed by the phone number the user wishes to reach. In this case the user enters 9 1 972 555 3456 as the destination number (step 2).

Note Unlike with Mobile Voice Access, Enterprise Feature Access requires that all two-stage dialed calls must originate from a phone that has been configured as a remote destination in order to match the caller ID and PIN against the end-user account. There is no provision within Enterprise Feature Access in which the mobility user can enter their remote destination number or ID to identify themselves to the system. Identity can be established only via the combination of incoming caller ID and entered PIN.

Next the outgoing call is originated via the user's remote destination profile (step 3), and the call to PSTN number 972 555-3456 is routed via the enterprise PSTN gateway (step 4). Finally, the call rings the PSTN phone (step 5: in this case, 972 555-3456). As with Mobile Voice Access, the voice media path of each Enterprise Feature Access two-stage dialed call is hairpinned within the enterprise PSTN gateway utilizing two gateway ports.

Figure 25-20 Enterprise Feature Access Two-Stage Dialing Feature



Note In order for Enterprise Feature Access two-stage dialing to work as in Figure 25-20, ensure that the system-wide Enable Enterprise Feature Access service parameter is set to True.

Desk and Remote Destination Phone Pickup

Because Mobile Voice Access and Enterprise Feature Access functionality is tightly integrated with the Mobile Connect feature, once a Mobile Voice Access or Enterprise Feature Access two-stage dialed call has been established, the user does have the option of using Mobile Connect functionality to pick up the in-progress call on their desk phone by simply hanging up the call on the originating phone and pushing the Resume softkey on their desk phone or by using the mid-call hold feature. In turn, the call can then be picked up on the user's configured remote destination phone by pressing the Mobility softkey and selecting Send Call to Mobile Phone.

Enabling and Disabling Mobile Connect

In addition to providing users of Mobile Voice Access and Enterprise Feature Access with the ability to make calls from the PSTN as though they are within the enterprise, the functionality provided by Mobile Voice Access on the H.323 or SIP VoiceXML gateway and provided by Enterprise Feature Access also gives users the ability to remotely enable and disable their Mobile Connect functionality for each remote destination via their phone keypad. Rather than entering a 1 to make a call, users enter a 2 to turn the Mobile Connect feature on and a 3 to turn the Mobile Connect feature off.

If a user has more than one remote destination configured when using Mobile Voice Access, they are prompted to key in the remote destination phone number for which they wish to enable or disable the Mobile Connect feature. When using Enterprise Feature Access, a user can enable or disable Mobile Connect only for the remote destination phone from which they are calling.

Note When the Enable Mobile Voice Access service parameter is set to False, resulting in an inability to make two-stage dialed calls, Mobile Voice Access still provides users with the ability to enable and disable mobile connect remotely. As long as the Mobile Voice Access Directory Number has been configured on the system, the user's account has been enabled for Mobile Voice Access, and the Cisco Unified Mobile Voice Access service is running on the publisher, a calling user can still enable or disable Mobile Connect.

Mobile Voice Access and Enterprise Feature Access Number Blocking

Administrators might want to prevent users of Mobile Voice Access and Enterprise Feature Access two-stage dialing from dialing certain numbers when using these features. In order to restrict or block calls to certain numbers when using these features for off-net calls, a comma-separated list of those numbers can be configured in the System Remote Access Blocked Numbers service parameter field. Once this parameter is configured with blocked numbers, those numbers will not be reachable from a user's remote destination phone when using Mobile Voice Access or Enterprise Feature Access features. Numbers that administrators might want to block can include emergency numbers such as 911. When configuring blocked numbers, ensure they are configured as they would be dialed by an enterprise user, with appropriate prefixes or steering digits. For example, if an emergency number is to be blocked and the emergency number is dialed by system users as 9911, then the number configured in the System Remote Access Blocked Numbers field should be 9911.

Access Numbers for Mobile Voice Access and Enterprise Feature Access

While the Unified CM system allows the configuration of only a single Mobile Voice Access Directory Number, this does not preclude the use of multiple externally facing numbers that can access these internally configured numbers. For example, consider a system deployed in the US in New York with a remote site in San Jose as well as an overseas site in London. Even though the system may have the Mobile Voice Access directory number configured as 555-1234, the gateways at each location can be configured to map a local or toll-free DID number to this Mobile Voice Access directory number. For example, the gateway in New York may have DIDs of +1 212 555 1234 and +1 800 555 1234, which both map to the Mobile Voice Access number, while the gateway in San Jose has a DID of +1 408 666 5678 and the gateway in London has a DID of +44 208 777 0987, which also map to the Mobile Voice Access number of the system. The Unified CM system does permit the configuration of multiple Enterprise Feature Access Numbers so that location-specific system access numbers can be configured for each geographic location of the deployment. This enables local or toll-free Enterprise Feature Access two-stage dialing functionality for all users regardless of geographic location.

By acquiring multiple local or toll-free DID numbers, system administrators can ensure that two-stage dialed calls will always originate as a call into the system that is either local or toll-free, thus providing further reductions in telephony costs.

Remote Destination Configuration and Caller ID Matching

When authenticating users for Mobile Voice Access and Enterprise Feature Access two-stage dialing functionality as well as the DTMF-based mid-call features Transfer and Conference, the caller ID of the calling remote destination phone is matched against all remote destinations configured within the system. Matching of this caller ID depends on a number of factors, including how the remote destination numbers are configured, whether digit prefixing is required to include PSTN steering digits on the system, and whether the Matching Caller ID with Remote Destination parameter is set to Partial or Complete Match. In all cases, the requirement is to be able to uniquely identify each mobility user based on the their remote destination number or numbers. For this reason, it is critical not only that remote destination numbers be configured uniquely within the system, but also that inbound caller ID matching (whether using complete or partial matching) must always uniquely correspond to a single remote destination. If a single or unique match is not found, caller ID matching will fail.

To control the nature of this matching, consider the following two approaches.

Using Complete Caller ID Matching

With this approach, remote destination numbers are configured exactly as the caller ID would be presented from the PSTN. For example, if the caller ID from the PSTN for a remote destination phone is presented to the system as 4085557890, then this number should be configured on the Remote Destination configuration page.

In order to route Mobile Connect calls appropriately to this remote destination, it is necessary to configure the dial plan to use either +E.164 dialing methods or a digit prefix mechanism to prefix necessary PSTN access codes and other required digits. For example, if you are not using a global +E.164 dial plan and assuming a 9 or other PSTN steering digits or country codes are required to reach the PSTN when dialing calls from the enterprise, then digit prefixing must be configured to add the appropriate PSTN steering digit and country code to the beginning of the configured remote destination number. Digit prefixing should be facilitated by using translation patterns, route patterns, or route list constructs within the Unified CM system. When using this complete match approach and a digit prefixing method, the Matching Caller ID with Remote Destination parameter should be left at the default setting of **Complete Match**.

Application Dial Rules may also be used to provide digit prefixing in these scenarios. However, it is worth noting that Application Dial Rules are applied based on called digit-string length and cannot be partitioned, meaning that they are applied globally across the system. This severely limits the use of Application Dial Rules, especially in scenarios where multiple dialing domains (for example, different countries) need to be supported on a single Unified CM cluster.

Note Not only are Application Dial Rules applied to Mobile Connect, Mobile Voice Access, and Enterprise Feature Access calls, but they are also applied to calls made with Cisco WebDialer, Cisco Unified CM Assistant, and Cisco Jabber applications. For this reason, exercise care when configuring these rules to ensure that dialing behavior across all applications is as expected.

The recommended dial plan approach is always to globalize the caller ID to +E.164 on ingress from the PSTN and always to configure remote destinations as +E.164. This will guarantee that the caller ID from the PSTN (after normalization) will always provide a unique match when compared against all configured remote destinations. Combined with a dial plan supporting +E.164 dialing, this eliminates the need for digit prefixing and ensures unique identification of remote destination users and numbers even when supporting multiple international numbering plans. Because the recommended dial plan approach is to globalize the caller ID on ingress and localize on egress according to trunk requirements and/or user expectations, using the unmodified caller ID as presented from the PSTN is not compatible with this approach.

Using Partial Caller ID Matching

With this approach, remote destinations are configured as they would be dialed from the system to the PSTN. For example, if the number for the remote destination is 14085557890 and PSTN access from the system requires a 9, then this number should be configured on the Remote Destination configuration page as 914085557890. This approach precludes the need for configuration of a digit prefixing mechanism on the system, but it requires setting the Matching Caller ID with Remote Destination service parameter to Partial Match and setting the Number of Digits for Caller ID Partial Match to the appropriate number of consecutive digits that should be matched against the remote destination caller ID. For example, if the caller ID for a remote destination is 14085557890 and the remote destination is configured as 914085557890, then the Number of Digits for Caller ID Partial Match would ideally be set to 10 or 11. In this example, this parameter could be set to a lower number of digits; however, always ensure that enough consecutive digits are matched so that all configured remote destinations in the system are matched uniquely. If there is no exact match or if more than one configured remote destination number is matched when using partial caller ID matching, the system treats this as if there is no matching remote destination number, thus requiring the user to enter their remote destination number/ID manually in the case of Mobile Voice Access before providing their PIN. With Enterprise Feature Access, there is no mechanism for the user to enter their remote destination number; therefore, when using this functionality, ensure that only unique matches occur.

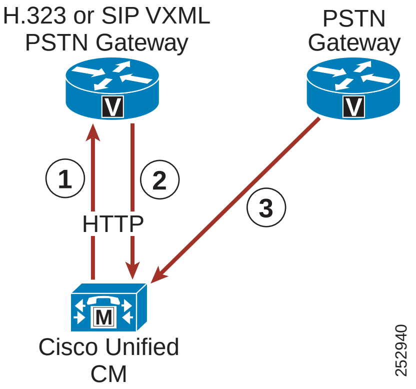
Note If the PSTN service provider sends variable-length caller IDs, using partial caller ID matching is not recommended because ensuring a unique caller ID match for each inbound call might not be possible. In these scenarios, using complete caller ID matching and/or a +E.164 dial plan is the preferred method.

Mobile Voice Access and Enterprise Feature Access Architecture

The architecture of the Mobile Voice Access and Enterprise Feature Access feature is as important to understand as their functionality. Figure 25-21 depicts the message flows and architecture required for Mobile Voice Access and Enterprise Feature Access. The following sequence of interactions and events can occur between Unified CM, the PSTN gateway, and the H.323 or SIP VXML gateway:

1. Unified CM forwards IVR prompts and instructions to the H.323 or SIP VXML gateway via HTTP (see step 1 in Figure 25-21). This provides the VXML gateway with the ability to play these prompts for the inbound Mobile Voice Access callers.
2. The H.323 or SIP VXML gateway uses HTTP to forward Mobile Voice Access user input back to Unified CM (see step 2 in Figure 25-21).
3. The PSTN gateway forwards DTMF digits in response to user or Smart Phone key sequences from the remote destination phone for Enterprise Feature Access two-stage dialing and mid-call features (see step 3 in Figure 25-21).

Figure 25-21 Mobile Voice Access and Enterprise Feature Access Architecture



Note While Figure 25-21 depicts the H.323 or SIP VoiceXML gateway as a separate box from the PSTN gateway, this is not an architectural requirement. Both VoiceXML functionality and PSTN gateway functionality can be handled by the same box, provided there are no requirements for the PSTN gateway to run a protocol other than H.323 or SIP. An H.323 or SIP gateway is required for Mobile Voice Access VoiceXML functionality.

High Availability for Mobile Voice Access and Enterprise Feature Access

The Mobile Voice Access and Enterprise Feature Access features rely on the same components and redundancy mechanisms as the Mobile Connect feature (see High Availability for Mobile Connect, page 25-43). Unified CM Groups are necessary for PSTN gateway registration redundancy. Likewise, PSTN physical gateway and gateway connectivity redundancy should be provided. Redundant access between the PSTN and the enterprise is required for remote destination phones to access Mobile Voice Access and Enterprise Feature Access features in the event of a gateway failure. However, while physical redundancy can and should be provided for the H.323 or SIP VoiceXML gateway, there is no redundancy mechanism for the Cisco Unified Mobile Voice Access service on Unified CM. This service can be enabled and run on the publisher node only. Therefore, if the publisher node fails, Mobile Voice Access functionality will be unavailable. Enterprise Feature Access and two-stage dialing functionality have no such dependency on the publisher and can therefore provide equivalent functionality to mobility users (without the IVR prompts).

Designing Cisco Unified Mobility Deployments

The Cisco Unified Mobility solution delivers mobility functionality via Cisco Unified CM. Functionality includes Mobile Connect, Mobile Voice Access, and Enterprise Feature Access. When deploying this functionality it is important to understand dial plan implications, guidelines and restrictions, and performance and capacity considerations.

Dial Plan Considerations for Cisco Unified Mobility

In order to configure and provision Unified Mobility appropriately, it is important to understand the call routing behavior and dial plan implications of the remote destination profile configuration.

Remote Destination Profile Configuration

When configuring Unified Mobility, you must consider the following two settings on the Remote Destination Profile configuration page:

- Calling Search Space

  This setting combines with the directory number or line-level calling search space (CSS) to determine which partitions can be accessed for mobility dialed calls. This affects calls made by the mobility user from the remote destination phone, including Mobile Voice Access and Enterprise Feature Access two-stage dialing as well as calls made in conjunction with mid-call transfer and conferencing features. Ensure that this CSS, in combination with the line-level CSS, contains all partitions that need to be accessed for enterprise calls originating from a user's remote destination phone. In a +E.164 dial plan using the line-only traditional approach with local route groups, this CSS is not required and can be set to **<None>**.

- Rerouting Calling Search Space

  This setting determines which partitions are accessed when calls are sent to a user's remote destination phone. This applies to all Mobile Connect calls. When a call to a user's enterprise directory number is also sent via Mobile Connect to a user's remote destination, this CSS determines how the system reaches the remote destination phone. For this reason, the CSS should provide access to partitions with appropriate route patterns and gateways for reaching the PSTN or mobile voice network.

When configuring the Remote Destination Profile Rerouting CSS, Cisco recommends that the route patterns within this CSS point to a gateway that is in the same call admission control location as the gateway used to route the inbound call to the user's desk phone. This ensures that a call admission control denial due to insufficient bandwidth between two locations will not occur when routing calls out to the remote destination. Further, because subsequent call admission control checks after the initial Mobile Connect call is routed will not result in a denial if there is insufficient WAN bandwidth, routing the inbound and outbound call legs out a gateway or gateways in the same call admission control location ensures that subsequent desk phone or remote destination pickup operations during this call will not require call admission control, which could result in WAN bandwidth oversubscription.

When using route patterns pointing to route lists that use Standard Local Route Group, the local route group configured on the caller’s device pool will be used. In this case the egress gateway for the call leg to the remote destination will be local to the original calling device. For calls coming in from the PSTN, this will help to fulfill the above requirement to use egress gateways in the same call admission control location as the original caller (in this case the incoming gateway).

Likewise, it is equally important to ensure that call admission control denials are minimized when placing two-stage dialed calls. Call admission control denials for two-stage dialed calls can be minimized or avoided by using local route group constructs so that the egress gateway used to route the outbound call leg is chosen by the ingress gateway of the inbound call leg. With this method, the ingress and egress gateways used will be in the same call admission control location. Alternatively, the route patterns within the Remote Destination Profile device-level CCS should point to an egress gateway that is in the same call admission control location as the ingress gateway that handled the inbound call leg to the Mobile Voice Access or Enterprise Feature Access system access number. However, be aware that a subsequent desk phone pickup can result in WAN bandwidth oversubscription if the desk phone is in a different call admission control location than the gateway through which the Mobile Voice Access or Enterprise Feature Access system access numbers are reached.

Automatic Caller ID Matching and Enterprise Call Anchoring

Another aspect of the Unified Mobility dial plan that is important to understand is the system behavior with regard to automatic caller ID identification for inbound calls from configured remote destination phones. Whenever an inbound call comes into the system, the presented caller ID for that call is compared against all configured remote destination phones. If a match is found, the call will automatically be anchored in the enterprise, thus allowing the user to invoke mid-call features and to pick up in-progress calls at their desk phone. This behavior occurs for all inbound calls from any mobility user's remote destination phone, even if the inbound call is not originated as a mobility call using Mobile Voice Access or Enterprise Feature Access.

Note Automatic inbound caller ID matching for configured remote destination numbers is affected by whether the Matching Caller ID with Remote Destination service parameter is set to Partial or Complete Match. See Remote Destination Configuration and Caller ID Matching, page 25-50, for more information about this setting.

In addition to automatic enterprise call anchoring, inbound and outbound call routing must also be considered when a configured remote destination phone is calling into the enterprise. Inbound call routing for calls from configured remote destinations occurs in one of two ways, depending on the setting of the service parameter Inbound Calling Search Space for Remote Destination. By default, this service parameter is set to **Trunk or Gateway Inbound Calling Search Space**. With the service parameter set to the default value, inbound calls from configured remote destinations will be routed using the Inbound Calling Search Space (CSS) of the PSTN gateway or trunk on which the call is coming in. If, on the other hand, the parameter Inbound Calling Search Space for Remote Destination is set to the value **Remote Destination Profile + Line Calling Search Space**, inbound calls coming from remote destinations will bypass the Inbound CSS of the PSTN gateway or trunk and will instead be routed using the associated Remote Destination Profile CSS (in combination with the line-level CSS).

Given the nature of inbound call routing from remote destination phones, it is important to make sure that calling search spaces are configured appropriately in order to provide access for these inbound calls to any partitions required for reaching internal enterprise phones, thus ensuring proper call routing from remote destination phones.

Note Incoming calls that do not come from a configured remote destination phone are not affected by the Inbound Calling Search Space for Remote Destination service parameter because they will always use the trunk or gateway inbound CSS.

Outbound call routing for Mobile Voice Access or Enterprise Feature Access calls always uses a concatenation of the Remote Destination Profile line CSS and device-level CSS, therefore it is important to make sure that these calling search spaces are configured appropriately in order to provide access to any route patterns necessary for off-net or PSTN access, thus ensuring proper outbound call routing from remote destination phones.

Intelligent Session Control and Ring All Shared Lines

The Intelligent Session Control feature enables automatic call anchoring for enterprise-originated calls made directly to configured remote destination numbers. Normally, mobility call anchoring is dependent exclusively on calls made to or on behalf of a user's enterprise number. The system already anchors externally originated calls made by enterprise two-stage dialing because these call are routed as internal calls. With the Intelligent Session Control feature enabled, the system will also anchor internally originated calls made directly to configured remote destinations.

This feature is enabled by setting the Reroute Remote Destination Calls to Enterprise Number service parameter to True. By default, this service parameter is set to False and the feature is disabled. When the feature is enabled, not only will the system route the call to the dialed remote destination by way of the PSTN, but it will also automatically anchor the call inside the enterprise gateway. By anchoring these types of calls, the system enables the called mobile user to invoke mid-call features and desk phone pickup or session handoff.

As an example, assume that the Intelligent Session Control feature has been enabled and that a mobility-enabled user has a remote destination number configured as 408 555 1234, which corresponds to their mobile number. If another system user dials the mobility-enabled user's remote destination number (408 555 1234) from their desk phone, the system will route the call through the PSTN to the remote destination and will simultaneously anchor the call in the enterprise gateway. Once the call is set up and anchored, the called mobility-enabled user now has the ability to invoke mid-call features such as hold, transfer, and conference, as well as the ability to perform a desk phone pickup or session handoff.

Taking this same example and assuming instead that the Intelligent Session Control feature is disabled, then when a system user dials the mobility-enabled user’s remote destination directly from a desk phone inside the enterprise, the call will still be routed to the called remote destination through the PSTN; however, the call will not be anchored. As a result, the mobile user would not be able to invoke mid-call hold or transfer and would have no ability to perform a desk phone pickup or session handoff.

When enabling this feature, it is important to understand the implications to dial plan configuration and call routing. To invoke the feature, the number dialed by an internal user to reach a remote destination number on the PSTN (including any required PSTN steering digits) must match the remote destination (or mobility identity) number as it is configured on the system. For example, if the remote destination number is configured on the system as 408 555 1234 but internal users must normally dial PSTN steering digits 91 in addition to the number they are calling, then rerouting and resulting enterprise call anchoring will not occur. This is because the user dialed 91 408 555 1234 to reach the remote destination on the PSTN but the remote destination was configured as 408 555 1234, so there is no match.

For this feature to function properly, matching must occur between the configured remote destination and the number that must be dialed to reach this remote destination on the PSTN. To ensure that this matching happens, set the service parameter Matching Caller ID with Remote Destination to **Partial Match**. By setting this parameter to Partial Match and then specifying the number of digits to partially match using the Number of Digits for Caller ID Partial Match service parameter, it is still possible to match the configured remote destination number with the dialed number even if it contains PSTN steering digits.

Using the previous example and assuming that system has been set to use partial match on ten digits, the dialed number 9 1 408 555 1234 can be matched to the configured remote destination 408 555 1234. This is because, with partial matching, the system attempts to match the same number of digits as specified by the Number of Digits for Caller ID Partial Match, which in this case is ten digits. The system attempts to match the two numbers by matching digits from right to left. The last ten digits of the dialed number 9 1 408 555 1234 are 408 555 1234, and these ten digits match the ten digits of the configured remote destination (408 555 1234). In this example, the resulting call is anchored in the enterprise and the called mobile user is able to invoke mid-call features and perform desk phone pickup or session handoff.

At first glance it might appear that an easier way to handle this feature would be to configure remote destination or mobility identity numbers that include any required PSTN steering digits. However, when configuring these numbers with required PSTN steering digits, if you do not also configure partial caller ID matching, the system will not be able to perform automatic caller ID matching and enterprise anchoring for inbound calls from configured remote destinations or mobility identities. In the previous example, if the remote destination number had been configured as 9 1 408 555 1234 and complete caller ID matching had been used, an inbound call from the remote destination would present caller ID of 408 555 1234 and a match would not occur, meaning the inbound call from the remote destination would not be anchored as expected.

Based on this potential for mismatch between dialed numbers for outbound calls and configured remote destination numbers for inbound calls, Cisco recommends enabling partial (rather than complete) caller ID matching when using the Intelligent Session Control feature for all deployments that require one or more steering digits to reach the PSTN. This ensures that calls made directly to the remote destination number using PSTN steering digits are still matched and anchored. On the other hand, if steering digits are not required to reach the PSTN and users are able to dial the full E.164 number to route calls to the PSTN, then Cisco recommends the complete caller ID matching setting because the remote destination is configured to match the caller ID and is the same number as dialed by internal users to reach the remote destination or mobility identity on the PSTN.

When enabling the Intelligent Session Control feature, it is also important to understand the behavior of the enterprise and remote destination lines during the reroute feature operation. On call reroute, remote destination line settings Do Not Disturb (DND), Access Lists and Time of Day call filtering, and the Delay Before Ringing Timer are ignored. All reroute calls are routed unfiltered and immediately. Enterprise desk phone line settings are also ignored or bypassed by default. However, Call Forward All settings on the enterprise desk phone line can be honored during reroute feature operation by setting the Ignore Call Forward All on Enterprise DN service parameter to False. If this parameter is set to False, on reroute operation, calls will not be routed to the remote destination if the enterprise desk phone line has a call-forward-all destination set. Instead, the call will be routed to the call-forward-all destination. By default, this service parameter is set to True, and call-forward-all settings on enterprise desk phone lines are ignored.

Intelligent Session Control functionality may be further enhanced by using the Ring All Shared Lines feature. This feature is enabled by setting the Ring All Shared Lines service parameter to True. By default, this service parameter is set to True and the feature is enabled. However, the Ring All Shared Lines feature is dependent on the Intelligent Session Control feature, which must also be enabled in order use the Ring All Shared Line functionality. When the Ring All Shared Lines and Intelligent Session Control features are both enabled, not only will the system route internally originated calls to the dialed remote destination by way of the PSTN, but all of the user's other shared-line devices will also receive the call. This includes the user's enterprise desk phone as well as other configured remote destinations. The called user will then be able to answer the incoming call on any of their devices and the call will be anchored in the enterprise.

Note If Ring All Shared Lines is enabled, mobile client devices will not receive calls at the cellular voice interface of the device when the device is registered to Unified CM.

Caller ID Transformations

Any calls made into the cluster by configured remote destination numbers will automatically have their caller ID or calling number changed from the calling remote destination phone number to the enterprise directory number of the associated desk phone. For example, if a remote destination phone with number 408 555-7890 has been configured and associated to a user's enterprise desk phone with number 555-1234, then any call from the user's remote destination phone destined for any directory number in the cluster will automatically have the caller ID changed from the remote destination number of 408 555-7890 to the enterprise directory number of 555-1234. This ensures that the active call caller ID display and call history log caller ID reflect a mobility user's enterprise desk phone number rather than their mobile phone number, and it ensures that any return calls are made to the user's enterprise number, thus anchoring those calls within the enterprise.

Likewise, calls from a remote destination phone to external PSTN destinations and anchored in the enterprise via Mobile Voice Access or Enterprise Feature Access two-stage dialing, or those calls forked to the PSTN as a result of Mobile Connect, will also have caller ID changed from the calling remote destination phone number to the associated enterprise directory number.

Finally, in order to deliver the calling party number as an enterprise DID number rather than an enterprise directory number to external PSTN phones, calling party transformation patterns can be used. By using calling party transformation patterns to transform caller IDs from enterprise directory numbers to enterprise DIDs, return calls from external destinations will be anchored within the enterprise because they will be dialed using the full enterprise DID number. For more information about these transformations and dial plan implications, see Special Considerations for Cisco Unified Mobility, page 9-125.

Guidelines and Restrictions for Unified Mobility

The following guidelines and restrictions apply with regard to deployment and operation of Mobile Connect within the Unified CM telephony environment:

- Mobile Connect is supported only with PRI TDM PSTN connections. T1 or E1-CAS, FXO, FXS, and BRI PSTN connections are not supported. This PRI requirement is based on the fact that Cisco Unified CM must receive expeditious answer and disconnect indication from the PSTN in order to ensure full feature support. Answer indication is needed in order for Cisco Unified CM to stop ringing the desk phone and other remote destinations when a Mobile Connect call is answered at a particular remote destination. In addition, answer indication is required in order to support the single enterprise voicemail box feature. Finally, disconnect indication is required for desk phone pickup. A PRI PSTN connection will always provide answer or disconnect indication.
- Mobile Connect is also supported over SIP trunk VoIP PSTN connections as long as the Cisco IOS Unified Border Element provides the demarcation point between the Unified CM SIP trunk and the service provider trunk and as long as mid-call features (or other DTMF-dependent features) are not in use. Mid-call features are not supported over VoIP PSTN connections. A VoIP-based PSTN connection is still able to provide expeditious answer and disconnect indication to Unified CM due to the end-to-end signaling path provided by VoIP-based PSTN connections.
- Mobile Connect can support up to two simultaneous calls per user. Any additional calls that come in are automatically transferred to the user’s voicemail.
- Mobile Connect does not work with Multilevel Precedence and Preemption (MLPP). If a call is preempted with MLPP, Mobile Connect features are disabled for that call.
- Mobile Connect services do not extend to video calls. A video call received at the desktop phone cannot be picked up on the cellular phone.

- The Unified CM Forced Authorization Code and Client Matter Code (FAC/CMC) feature does not work with Mobile Voice Access or Enterprise Feature Access.
- Remote destinations must be Time Division Multiplex (TDM) devices or off-system IP phones on other clusters or systems. You cannot configure IP phones within the same Unified CM cluster as remote destinations.

For additional guidelines and restrictions, refer to the information on Cisco Unified Mobility in the latest version of the *Cisco Unified Communications Manager Features and Services Guide*, available at

http://www.cisco.com/en/US/products/sw/voicesw/ps556/prod_maintenance_guides_list.html

Capacity Planning for Cisco Unified Mobility

Cisco Unified Mobility supports a maximum of 15,000 remote destinations or mobility identities per Unified CM cluster. The maximum number of mobility-enabled users would thus be 15,000 users, assuming a single remote destination or mobility identity per user. As the number of remote destinations or mobility identities per user increases, the number of supported mobility-enabled users decreases.

Note A mobility-enabled user is defined as a user that has a remote destination profile and at least one remote destination or a mobility identity configured.

Note A mobility identity is configured just like a remote destination within the system, and it has the same capacity implications as a remote destination. Unlike a remote destination, however, the mobility identity is associated directly to a phone device rather than a remote destination profile. The mobility identity applies only to mobile client devices such as Cisco Jabber.

Scalability and performance of Cisco Unified Mobility ultimately depends on the number of mobility users, the number of remote destinations or mobility identities each user has, and the busy hour call attempt (BHCA) rates of those users. Multiple remote destinations per user and/or high BHCA per user can result in lower capacity for Cisco Unified Mobility. For more information on Cisco Unified Mobility sizing, including Unified CM server node capacities and hardware specific per-node and per-cluster capacities, see the chapter on Unified Communications Design and Deployment Sizing Considerations, page 29-1.

Design Considerations for Cisco Unified Mobility

Observe the following design recommendations when deploying Unified Mobility:

- Ensure that the PSTN gateway protocol is capable of out-of-band DTMF relay or allocate media termination points (MTPs) in order to covert in-band DTMF to out-of-band DTMF. When using Cisco IOS gateways for PSTN connectivity, out-of-band DTMF relay will be supported. However, third-party gateways might not support a common out-of-band DTMF method, and as a result an MTP might be required. In order to use Enterprise Feature Access Two-Stage Dialing and mid-call features, DTMF digits must be received out-of-band by Cisco Unified CM.

  Note When relying on MTP for converting in-band DTMF to out-of-band DTMF, be sure to provide sufficient MTP capacity. If heavy or frequent use of Enterprise Feature Access Two-Stage Dialing or mid-call features is anticipated, Cisco recommends a hardware-based MTP or Cisco IOS software-based MTP.

- Prior to deploying Unified Mobility, it is important to work with the PSTN provider to ensure the following:
  - Caller ID is provided by the service provider for all inbound calls to the enterprise. This is a requirement if Enterprise Feature Access Two-Stage Dialing or mid-call transfer, conference, and directed call park features are needed.
  - Outbound caller ID is not restricted by the service provider. This is a requirement if there is an expectation that mobility-enabled users will receive the caller ID of the original caller at their remote destination rather than a general enterprise system number or other non-meaningful caller ID.

Note Some providers restrict outbound caller ID on a trunk to only those DIDs handled by that trunk. For this reason, a second PRI trunk that does not restrict caller ID might have to be acquired from the provider. To obtain an unrestricted PRI trunk, some providers might require a signed agreement from the customer indicating they will not send or make calls to emergency numbers over this trunk.

Note Some providers allow unrestricted outbound caller ID on a trunk as long as the Redirected Dialed Number Identification Service (RDNIS) field or SIP Diversion Header contains a DID handled by the trunk. The RDNIS or SIP Diversion Header for forked calls to remote destinations can be populated with the enterprise number of the user by checking the Redirecting Number IE Delivery - Outbound check box on the gateway or trunk configuration page. Contact your service provider to determine if they honor the RDNIS or SIP Diversion Header and allow unrestricted outbound caller ID.

- Because mobility call flows typically involve multiple PSTN call legs, planning and allocation of PSTN gateway resources is extremely important for Unified Mobility. In cases where there are large numbers of mobility-enabled users, PSTN gateway resources will have to be increased. The following methods are recommend to minimize or reduce PSTN utilization:
  - Limit the number of remote destinations per mobility-enabled user to one (1). This will reduce the number of DS0s that are needed to extend the inbound call to the user's remote destination. One DS0 is consumed for each configured remote destination when a call comes into the user's enterprise directory number, even if the call is not answered at one of the remote destinations. Note that a DS0 per remote destination may be used for as long as 10 seconds, even if the call is not answered at the remote destination.
  - Use access lists to block or restrict the extension of calls to a particular remote destination based on incoming caller ID. Because access lists can be invoked based on the time of day, this eliminates the need for repeated updates of access lists by the end-user or the administrator.
  - Educate end-users to disable Mobile Connect when not needed, to further eliminate DS0 utilization when a call comes in for that user's enterprise number. If Mobile Connect is disabled, incoming calls will still ring the desk phone and will still forward to enterprise voicemail if the call goes unanswered.
- Due to the potential for call admission control denials resulting from insufficient WAN bandwidth between locations and the possibility that a desk phone pickup or remote destination pickup may result in a WAN bandwidth oversubscription, Cisco recommends configuring Remote Destination Profile CSS and Rerouting CSS so that route patterns within these CSSs point to gateways that are located within the same call admission control location as the gateway on which the inbound call leg comes in. For more information, see Remote Destination Profile Configuration, page 25-53.

- If you enable the Intelligent Session Control feature in deployments where PSTN steering digits must be dialed to access the PSTN, Cisco recommends setting the Matching Caller ID with Remote Destination service parameter to **Partial Match** and configuring the appropriate number of digits (Number of Digits for Caller ID Partial Match service parameter) to achieve a partial match of configured remote destinations or mobility identities. This will ensure proper functioning of the Intelligent Session Control feature and the mobility automatic caller ID matching and anchoring features.

Cisco Mobile Clients and Devices

As the prevalence of mobile users, mobile phones, and mobile carrier services continues to increase, the ability to use a single device for voice, video, and data services both inside and outside the enterprise becomes increasingly attractive. Mobile devices, including dual-mode smartphones and the clients that run on them, afford an enterprise the ability to provide customized voice, video, and data services to users while inside the enterprise and to leverage the mobile carrier network as an alternate connection method for general voice and data services. By enabling voice, video, and data services inside the enterprise and providing network connectivity for mobile client devices, enterprises are able to provide these services locally at reduced connectivity costs. For example, voice over IP (VoIP) calls made on the enterprise network will typically incur less cost than those same calls made over the mobile voice network.

In addition to providing voice and video over IP (VVoIP) capabilities, these mobile clients and devices enable mobile users to access and leverage other back-end collaboration applications and services. Other services and applications that can be leveraged through Cisco mobile clients and services include enterprise directory, enterprise voicemail, and XMPP-based enterprise IM (instant messaging) and presence. Further, these clients and devices can be deployed in conjunction with Cisco Unified Mobility so that users can leverage additional features and functions with their mobile device, such as Mobile Connect, enterprise two-stage dialing through Mobile Voice Access or Enterprise Feature Access, and single enterprise voicemail box.

This section examines mobile client architecture and common functions and features provided by Cisco mobile clients and devices, including remote secure attachment and handoff considerations related to moving an active voice call between the enterprise WLAN network and the mobile voice network. After covering the general mobile client solution architecture and features and functions, this section provides coverage of various capabilities and integration considerations for the following specific mobile clients and devices:

- Cisco Jabber — A set of mobile clients for Android and iOS mobile devices including iPhone and iPad, providing the ability to make voice and/or video calls over IP on the enterprise WLAN network or over the mobile data network as well as the ability to access the corporate directory and enterprise voicemail services and XMPP-based enterprise IM and presence.
- Cisco Cius — A purpose-built Android-based enterprise tablet device providing the ability to make voice and video calls over WLAN or mobile data networks and to access additional enterprise services such as directory, voicemail, IM and presence.

In addition, this section discusses high availability and capacity planning considerations for Cisco mobile clients and devices.

Cisco Mobile Clients and Devices Architecture

Cisco mobile clients are deployed on a wide range of mobile devices including tablets and handheld devices with only IP-based network connectivity capabilities (IEEE 802.11 wireless local area network or mobile provider data network) and dual-mode phones, which contain two physical interfaces or radios that enable the device to connect to both mobile voice and data carrier networks by means of traditional cellular or mobile network technologies and to connect to wireless local area networks (WLANs) using 802.11. Cisco mobile clients and devices enable on-premises data and real-time traffic (voice and video) connectivity through an 802.11 WLAN. In addition, these clients and devices provide remote data and real-time traffic (voice and video) connectivity to the enterprise through public or private WLANs or over the mobile data network. For devices with provider cellular voice radios, voice connectivity may also be enabled through the mobile voice network and PSTN.

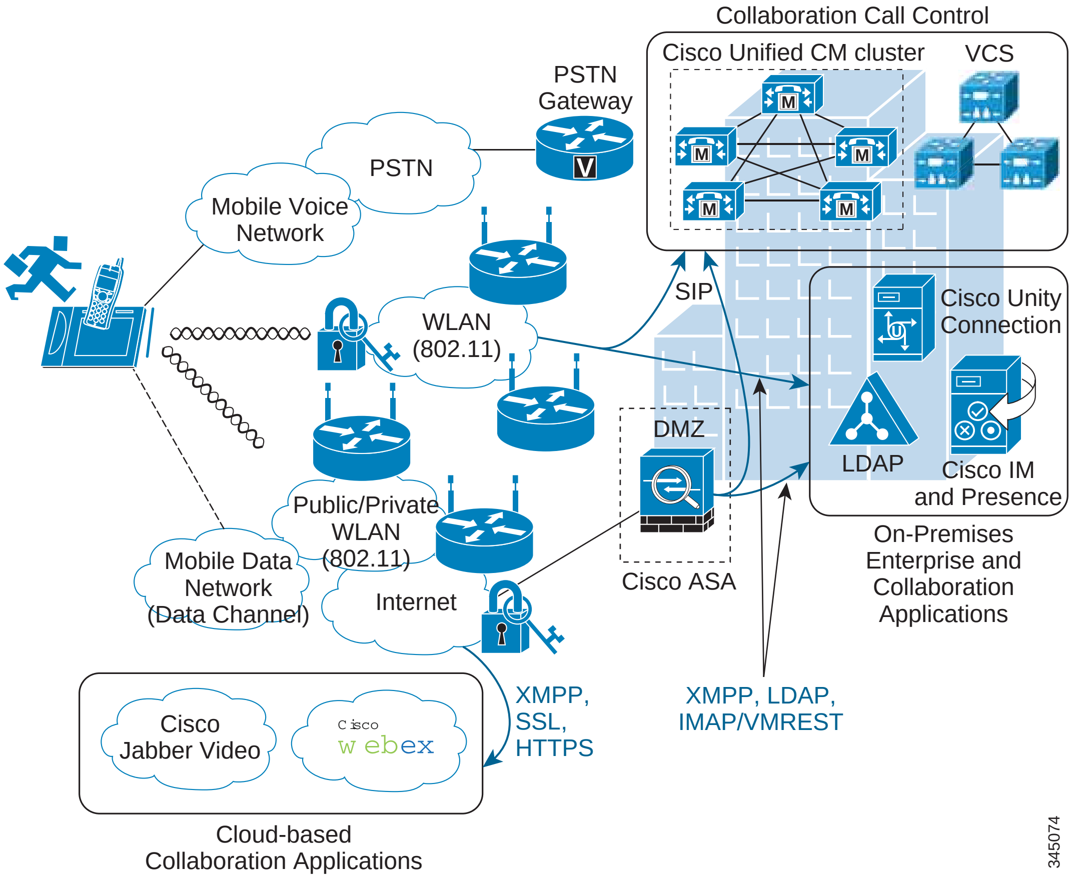
Note The use of the term *dual-mode phone* in this section refers specifically to devices with 802.11 radios in addition to the cellular radio for carrier voice and data network connectivity. Dual-mode devices that provide Digital Enhanced Cordless Telecommunications (DECT) or other wireless radios and/or multiple cellular radios are outside the scope of this section.

Figure 25-22 depicts the basic Cisco mobile clients and devices solution architecture for connecting mobile client devices into a Cisco Unified Communications System. For voice and video services, mobile client devices associate to the enterprise WLAN or connect over the Internet (from a public or private WLAN hot spot or the mobile data network), and the Cisco mobile client registers to Cisco Unified CM as an enterprise phone using the Session Initiation Protocol (SIP). Alternatively, some mobile clients may instead register to Cisco Video Communications System (VCS) as a video endpoint using SIP. Once registered, the client device relies on the underlying enterprise Cisco IP telephony network for making and receiving calls. When the mobile device is connected to the enterprise network and the client is registered to Unified CM, the device is reachable through the user's enterprise number. Any inbound calls to the user's enterprise number will ring the mobile client device. If the user has a Cisco IP desk phone, then the mobile client registration enables a shared line instance for the user's enterprise number so that an incoming call rings both the user's desk phone and the mobile device. When unregistered, the mobile client device will not receive incoming enterprise calls unless the mobile device has an active cellular voice radio, the user has been enabled for Cisco Unified Mobility, and Mobile Connect (or single number reach) has been turned on for the user's mobile phone number. In these scenarios the mobile voice network and PSTN are used for making and receiving voice-only calls.

Unified Mobility features such as Mobile Connect are not compatible with tablets and other mobile client devices that do not have cellular voice radios because these non-dual-mode devices do not have a native PSTN reachable number. Non-dual-mode devices are able to make and receive enterprise calls only when connected to the enterprise and registered to the enterprise call control system.

As shown in Figure 25-22, when attached to the enterprise, Cisco mobile clients and devices can also communicate directly with other back-end application servers such as the LDAP corporate directory, Cisco Unity Connection enterprise voicemail system, and the Cisco IM and Presence Service for access to additional enterprise unified communications services such as messaging and presence. Cisco mobile clients and devices also integrate with cloud-based collaboration services such as Cisco WebEx, which delivers IM and presence, point-to-point IP calling, and web conferencing services.

Figure 25-22 Cisco Mobile Clients and Devices Architecture



Note The voice and video quality of calls will vary depending on the Wi-Fi or mobile data network connection. Cisco Technical Assistance Center (TAC) is not able to troubleshoot connectivity or voice and video quality issues over 3G/4G mobile data networks or non-corporate Wi-Fi networks.

Dual-mode mobile client devices must be capable of dual transfer mode (DTM) in order to be connected simultaneously to both the mobile voice and data network and the WLAN network. This allows the device to be reachable and able to make and receive calls on both the cellular radio and WLAN interface of the device. In some cases proper mobile client operation might not be possible if mobile voice and data networks do not support dual-connected devices.

Voice and Video over Wireless LAN Network Infrastructure

Before considering the various mobile client device features and functions and the impact these features and functions have on the enterprise telephony infrastructure, it is critical to plan and deploy a finely tuned, QoS-enabled, and highly available WLAN network. Because dual-mode phones and other mobile devices rely on the underlying WLAN infrastructure for carrying both critical signaling and other traffic for setting up calls and accessing various applications as well as the real-time voice and video media traffic, deploying a WLAN network optimized for both data and real-time media traffic is necessary. A poorly deployed WLAN network will be subjected to large amounts of interference and diminished capacity, leading not only to poor voice and video quality but in some cases dropped or missed calls.

This will in turn render the WLAN deployment unusable for making and receiving calls. Therefore, when deploying dual-mode phones and other mobile devices, it is imperative to conduct a WLAN radio frequency (RF) site survey before, during, and after the deployment to determine appropriate cell boundaries, configuration and feature settings, capacity, and redundancy to ensure a successful deployment of voice and video over WLAN. Each mobile device type and/or client should be tested on the WLAN deployment to ensure proper integration and operation prior to a production deployment. Using a WLAN that has been deployed and configured to provide optimized real-time traffic over WLAN services (such as the Cisco Unified Wireless Network), including quality of service, will ensure a successful mobile client device deployment.

Cisco recommends relying on the 5 GHz WLAN band (802.11a/n) whenever possible for connecting mobile clients and devices capable of generating voice and video traffic. 5 GHz WLANs provide better throughput and less interference for voice and video calls.

For more information on voice and video over WLAN deployments and wireless device roaming, see Wireless Device Roaming, page 25-6.

Note While dual-mode phones and other mobile client devices are capable of connecting back to the enterprise through the Internet for call control and other Unified Communications services, Cisco cannot guarantee voice and video quality or troubleshoot connectivity or voice and video quality issues for these types of connections. These types of connections include remote connections to the enterprise through public or private WLAN access points (APs) or hot spots or through the mobile data network. Cisco recommends an enterprise class voice and video-optimized WLAN network for connecting dual-mode phones and other mobile client devices. Most public and private WLAN APs and hot spots are tuned for data applications and devices. In these cases, the AP radios are turned to maximum power, and dynamic-power control results in devices enabling maximum power on network attachment, which allows for larger client capacities. While this may be ideal for data applications that are capable of retransmitting dropped or lost packets, for real-time traffic applications this can result in poor voice and video quality due to the potential for large numbers of dropped packets. Likewise, mobile provider data networks are susceptible to congestion and/or dropped connections, which can also result in poor call quality and dropped calls.

Mobile Client and Device Quality of Service

Cisco mobile client applications and devices generally mark Layer 3 QoS packet values in accordance with Cisco collaboration QoS marking recommendations. Table 25-3 summarizes these markings.

Table 25-3 Cisco Mobile Client Layer 3 QoS Markings

	Layer 3 Marking	
Traffic Type	DSCP[1]	PHB[2]
Voice media (audio only)	DSCP 46	PHB EF
Video media (audio and video)	DSCP 34	PHB AF41
Call Signaling	DSCP 24	PHB CS3

[1] Differentiated Services Code Point

[2] Per-Hop Behavior

Cisco mobile client Layer 2 802.11 WLAN packet marking (User Priority, or UP) presents challenges given the various mobile platform and firmware restrictions. Because Cisco mobile clients run on a variety of mobile devices, Layer 2 wireless QoS marking is inconsistent and therefore Layer 2 wireless QoS marking cannot be relied on to provide appropriate treatment to traffic on the WLAN. In deployments with Cisco Unified Wireless LAN Controllers, enabling wireless SIP call admission control (CAC) might provide some relief for incorrect or nonexistent Layer 2 WLAN marking. SIP CAC utilizes media session snooping and ensures that downstream voice and video frames are prioritized and/or treated correctly.

Despite appropriate mobile client application Layer 3 or even Layer 2 packet marking, mobile devices present many of the same challenges as desktop PCs in terms of generating many different types of traffic, including both data and real-time traffic. Given this, mobile devices generally fall into the untrusted category of collaboration endpoints. For deployments where mobile client devices are not considered trusted endpoints, packet marking or re-marking based on traffic type and port numbers is required to ensure that network priority queuing and dedicated bandwidth is applied to appropriate traffic. In addition to re-marking the mobile device traffic, Cisco recommends using network-based policing and rate limiting to ensure that the mobile client devices do not consume too much network bandwidth.

Alternatively, given appropriate Cisco mobile client Layer 3 marking and assuming mobile client devices are trusted, Cisco mobile client traffic will be queued appropriately as it traverses the enterprise network by using priority voice queuing and dedicated video media and call signaling bandwidth queues.

Cisco Mobile Clients and Devices Features and Functions

Cisco mobile clients and devices provide a range of features and functions. While features and operations may vary from device to device, the common operations and behaviors described in this section apply to all Cisco mobile client devices.

Enterprise Call Routing

Because Cisco mobile clients and devices are capable of making and receiving calls using the enterprise telephony infrastructure and call control services, it is important to understand the nature and behavior of call routing as it pertains to mobile client devices.

Inbound Call Routing

When mobile clients and devices register to Unified CM as an enterprise device with enterprise number, the mobile device rings when incoming calls to the system are destined for the user's enterprise number. This occurs for incoming calls originated on the PSTN or from other Unified CM clusters or enterprise IP telephony systems as well as for incoming calls originated within the Unified CM cluster by other users. If the mobile client device user has other devices or clients that are also associated to the enterprise number, these devices will also ring as shared lines; and once the call is answered at one of the devices or clients, ringing of all other devices and clients ceases.

In scenarios where a user has been enabled for Cisco Unified Mobility, and when Mobile Connect or single number reach is enabled for the user's dual-mode mobile phone number, the incoming call may also be extended to the mobility identity corresponding to the user's mobile phone number. However, this depends on whether the mobile device is connected to the enterprise WLAN network or attached to the enterprise network through secure VPN and registered to Unified CM. In situations in which the device is connected to the enterprise network directly or through a secure remote connection, an incoming call to the user's enterprise number will not be extended by Mobile Connect to the mobility identity of the mobile device even if Mobile Connect is turned on for this mobile number. The reason an incoming call to the enterprise number is not extended to the mobility identity of a dual-mode mobile device when it is registered to Unified CM is that the system is aware the device is connected to the enterprise network and available. Thus, in order to reduce utilization of enterprise PSTN resources, Unified CM does not extend the call to the dual-mode mobile phone's mobile voice network interface through the PSTN. Instead, only the WLAN or mobile data network interface corresponding to the enterprise number receives the call.

For situations in which the mobile device is not connected to the enterprise network directly or through a secure remote connection or is not registered to Unified CM, incoming calls to the enterprise number will be extended to the dual-mode mobile phone number per the configured mobility identity, assuming that the user has been enabled for Unified Mobility and that Mobile Connect for the mobility identity is turned on. For more information on integration of mobile clients and devices with Unified Mobility, see Interactions Between Cisco Jabber and Cisco Unified Mobility, page 25-75.

The same behavior and logic described above also applies with the Ring All Shared Lines feature. If this feature is enabled, calls are extended to the mobility identity or cellular number only when the dual-mode mobile client device is *not* registered to Unified CM. For more information on the Ring All Share Line feature, see Intelligent Session Control and Ring All Shared Lines, page 25-55.

In all cases, incoming calls made directly to the dual-mode device's mobile network phone number will always be routed directly to the mobile voice interface of the dual-mode device on the mobile network, unless the provider network or device settings are such that calls are not extended to the device by the mobile network. This is considered appropriate behavior because these calls were not made to the user's enterprise number. These would be considered personal calls, and as such should not be routed through the enterprise.

Note Mobile client devices that do not have cellular voice radios, such as the Cisco Cius and other tablet devices, are not dual-mode devices and as such cannot be reached on a mobile voice network interface. These devices can be reached only at the enterprise number by voice-over-IP.

Outbound Call Routing

For outbound calls from the dual-mode mobile device, the interface used depends on the location and connectivity of the device at that particular time. If the dual-mode device is not connected to the enterprise and not registered to Unified CM, then calls are routed by the cellular voice radio interface to the mobile voice network as usual. However, when connected to the enterprise and registered to Unified CM, the mobile device should make all calls through the enterprise telephony infrastructure. If no enterprise connectivity is available or the mobile client is unregistered, then outbound calling is not possible from the enterprise number, and instead calls would have to use the mobile number of the mobile client device for making calls over the mobile voice network. Alternatively, users may use the two-stage dialing features provided with Cisco Unified Mobility (see Mobile Voice Access and Enterprise Feature Access, page 25-44).

Dial Plan

The enterprise dial plan determines the dialing behavior of the mobile client device when it is connected to the enterprise and registered to Unified CM. For example, if the enterprise dial plan is configured to allow abbreviated dialing to reach internal extensions, then a mobile device registered to Unified CM can use this abbreviated dialing. While it is certainly a convenience for dual-mode mobile phone users to be able to dial within the enterprise using enterprise dialing habits and abbreviated dialing as well as site-based and/or PSTN steering digits for outbound calls, it is also a somewhat unnatural dialing scheme because mobile phone users typically dial numbers for outgoing calls on their mobile phone by using full E.164 dial strings since this is what is expected by the mobile voice network for outbound calling.

The enterprise dialing experience for an end-user is ultimately up to the enterprise policies and administrator of the enterprise telephony deployment. However, for dual-mode mobile devices, Cisco recommends normalizing required dialing strings for dual-mode client devices so that user dialing habits are maintained whether the device is connected to the enterprise network and registered to Unified CM or not. Because dialing on the mobile voice network is typically done using full +E.164 (with a preceding '+') and mobile phone contacts are typically stored with full +E.164 numbers, Cisco recommends configuring the enterprise dial plan to accommodate full +E.164 with preceding '+' for dual-mode mobile devices. When the dial plan is configured within Unified CM to handle this type of outbound dialing for dual-mode phones, it is possible for users to store a single set of contacts on the phone in the +E.164 format and, when dialed from these contacts or manually using the full +E.164 number, calls will always be routed to the appropriate destination, whether the device is connected to the enterprise network directly or over secure remote connection and registered to Unified CM or connected only to the mobile voice network. Configuring the enterprise dial plan in this manner provides the best possible end-user dialing experience so that users' mobile device dialing habits are maintained and they do not have to be aware of whether the device has enterprise connectivity and is registered to Unified CM.

To achieve normalized dialing from dual-mode phones, whether connected to the enterprise or just the mobile voice network, configure the dial plan within Unified CM with the following considerations in mind:

- Ensure that the enterprise dial plan is capable of handling dial strings from dual-mode phones typically used on the mobile voice network. For example, the dial plan should be configured to handle the following strings, which might be dialed from a mobile phone to reach a particular phone through the mobile voice network: +1 408 555 1234, 408 555 1234. Supporting the latter 10-digit dialing method (for example, 408 555 1234) might potentially overlap with other dialing habits such as abbreviated intra-site dialing. In that case the administrator has to decide which of the colliding dialing habits (10-digit dialing or abbreviated intra-site) should be available for dual-mode phones registered to the enterprise network. The set of dialing habits supported on dual-mode phones often differs from the set of dialing habits supported on regular endpoints.
- For calls destined for other enterprise numbers, systems configured for abbreviated dialing should be capable of modifying dial strings and rerouting to enterprise extensions as appropriate. For example, assuming the enterprise dial plan is based on five-digit internal dialing, the system should be configured to handle call routing to an enterprise extension so that a call to made to +1 408 555 1234 or 408 555 1234 is modified and rerouted to 51234 if the call is made while the dual-mode device is registered to Unified CM.
- Ensure that all inbound calls to the enterprise destined for dual-mode devices have the calling number and/or caller ID prefixed with appropriate digits so that missed, placed, and received call history lists are in full +E.164 formats. This will allow dual-mode device users to dial from call history lists without the need for editing the dial string. Instead, users will be able to select a number from the call history list to redial, whether connected to the enterprise or not. For example, if an incoming call from inside the enterprise originates from 51234 to a dual-mode user's enterprise number and the call goes unanswered, Unified CM should be configured to manipulate the calling number so that the resulting entry within the history list of the dual-mode device shows either 408 555 1234 or +1 408 555 1234. This number can be dialed whether the dual-mode device is connected to the enterprise or just to the mobile voice network without the need for further manipulation.

The one exception to normalized dialing for dual-mode devices is for scenarios in which some enterprise extensions or phones are reachable only internally (that is, they have no externally reachable corresponding DID number). In these situations, non-externally reachable numbers can be dialed (manually or from contacts) using abbreviated formats. Because these numbers will never be available externally and can be dialed only from inside the enterprise, some sort of enterprise-only indication should be made when storing these numbers in the contact list. Further, incoming calls from these internal-only numbers should not have the calling number modified for call history lists because these numbers may be called only inside the enterprise. Instead, calls from these extensions should be listed in all call history lists without modification so that the abbreviated dial strings can be successfully dialed only while the device is connected to the enterprise and registered to Unified CM.

Mobile client devices that do not have cellular voice radios, such as the Cisco Cius, are dependent exclusively on enterprise connectivity and enterprise voice and video telephony or cloud-based collaboration services.

Emergency Services and Dialing Considerations

Mobile client devices do present a slight challenge when it comes to making calls to emergency service numbers such as 911, 999, and 112. Because the mobile client devices may be located inside or outside the enterprise, providing location indication of a device and its user in the event of an emergency must be considered. Dual-mode mobile devices with cellular voice radios receive location services from their provider networks, and these location services are always available when the device is connected and typically able to pinpoint locations far more precisely than enterprise wireless networks; therefore, Cisco recommends that dual-mode device users rely on the mobile voice network for making emergency calls and determining device and user location. To ensure that Cisco dual-mode client devices rely exclusively on the mobile provider voice network for emergency and location services, these clients force all calls made to numbers configured in the Emergency Numbers field on the mobile client device configuration page to route over the mobile voice network. Further, dual-mode phone users should be advised to make all emergency calls over the mobile voice network rather than the enterprise network.

While making emergency calls over WLANs or mobile data networks is not recommended, mobile devices that do not have cellular voice radios are capable of making calls only through these data interfaces. Mobile devices that do not have cellular voice radios should not be relied upon for making emergency calls.

Enterprise Caller ID

When mobile client devices are connected to the enterprise and registered to Unified CM (either through the mobile data network or a WLAN), all calls made with the enterprise line over the WLAN or mobile data network will be routed with the user's enterprise number as caller ID. This ensures that returned calls made from call history lists at the far-end are always routed through the enterprise because the return call is to the user's enterprise number. If a dual-mode mobile device user has been enabled for Cisco Unified Mobility, and Mobile Connect is turned on for the mobile phone number, return calls to the enterprise number would also be extended to the dual-mode device through the PSTN whenever it is not connected to the enterprise.

Mid-Call Features

When mobile client devices are connected to the enterprise and registered to Unified CM as enterprise endpoints, they are able to invoke call processing supplementary services such as hold, resume, transfer, and conference, using SIP call signaling methods as supported by Unified CM. Just as with any IP phone or client registered to Unified CM, these devices are able to leverage enterprise media resources such as music on hold (MoH), conference bridges, media termination points, and transcoders.

External Call Routing

When dual-mode mobile client devices are not connected to the enterprise and/or not registered to Unified CM, they may make and receive calls only over the mobile voice network. For this reason, Unified CM has no visibility into any calls being made or received at the dual-mode mobile device while it is unregistered. The mobile number is the caller ID being sent to the network when calls are made from dual-mode phones not connected to the enterprise. This will likely result in unanswered calls being made directly back to the dual-mode device's mobile number instead of being routed back through the enterprise.

If the dual-mode mobile client device is integrated with Cisco Unified Mobility, enterprise two-stage dialing services may be leveraged for making calls through the enterprise network even when the dual-mode device is outside the enterprise and not registered to Unified CM. Unified Mobility two-stage dialing is done using either Mobile Voice Access or Enterprise Feature Access and requires the user to dial an enterprise system access DID number and enter credentials prior to dialing the number they are calling. For more information on Unified Mobility two-stage dialing features, see Mobile Voice Access and Enterprise Feature Access, page 25-44.

Likewise, if the dual-mode phone is integrated with Unified Mobility, a user can also receive incoming calls to their enterprise number at the mobile number through Mobile Connect; can invoke mid-call features using DTMF key sequences including hold, resume, transfer, and conference; and can perform desk phone pickup to move an active call from the mobile phone to the enterprise desk phone.

Remote Secure Enterprise Connectivity

Mobile client devices can utilize the IP telephony infrastructure for enterprise voice and video over IP calling even when not connected to the enterprise, provided they have a secure connection back to the enterprise in order to register the client with Unified CM. Remote secure connectivity for these devices requires the use of a VPN solution such as Cisco AnyConnect mobile client in order to secure the client connection over the Internet.

Voice and video quality and user experience for remotely attached mobile client devices will vary depending on the nature of the Internet-based network connection. Cisco cannot guarantee acceptable voice and video quality nor successful connectivity for these types of client connections. Care should be taken when relying on these types of connections for business-critical communications. In the case of dual-mode devices with unreliable or low-bandwidth Internet connections, users with dual-mode devices should be advised to make calls over the mobile voice network if connectivity is available rather than relying on the enterprise telephony infrastructure.

Additional Services and Features

In addition to call processing or call control services, Cisco mobile clients and devices are capable of providing the additional features and services described in this section.

Dual-Mode Call Handoff

One very important aspect of dual-mode device deployments is call preservation as a user moves in and out of the enterprise or as the device connects to and disconnects from the enterprise network and network connectivity changes from the cellular voice radio to the WLAN radio, and vice versa. Because dual-mode phone users are often mobile, it is important to maintain any active call as a dual-mode user moves in or out of the enterprise. For this reason, dual-mode client devices and the underlying enterprise telephony network should be capable of some form of call handoff.

There are two types of call handoff that should be accommodated by both the dual-mode client and the underlying IP telephony infrastructure:

- Hand-out

  Call hand-out refers to the movement of an active call from the WLAN or mobile data network interface of the dual-mode phone to the cellular voice interface of the dual-mode phone. This requires the call to be handed out from the enterprise IP network to the mobile voice network through the enterprise PSTN gateway.

- Hand-in

  Call hand-in refers to the movement of an active call from the cellular voice interface of the dual-mode phone to the WLAN or mobile data network interface of the dual-mode phone. This requires the call to be handed in from the mobile voice network to the enterprise IP network through the enterprise PSTN gateway.

The handoff behavior of a dual-mode phone depends on the nature of the dual-mode client and its particular capabilities. Dual-mode client handoff may be invoked manually by the user or automatically based on network conditions. In manual handoff scenarios, the dual-mode users are responsible for engaging and completing the handoff operation based on their location and needs. With automatic handoff, the mobile client monitors the WLAN signal and makes handoff decision based on strengthening or weakening of the WLAN signal at the client. Hand-out is engaged in the case of a weakening WLAN signal, while hand-in is engaged in the case of a strengthening WLAN signal. Automatic handoff depends on the mobile device to provide capabilities for monitoring WLAN signal strength.

Handoff operations are critical for maximizing utilization of the enterprise IP telephony infrastructure for phone calls.   These operations are also necessary for providing voice continuity and good user experience so that users do not have to hang up the original call and make another call to replace it.

XMPP-Based IM and Presence

Some mobile clients are capable of providing enterprise instant messaging (IM) and presence services based on the Extensible Messaging and Presence Protocol (XMPP), through integration to an on-premises or off-premises application server or service. In either case, the IM and presence capabilities of these mobile clients enable the following:

- Adding users to contact or buddy lists
- Setting and propagating user presence and availability status
- Reception of presence status for a buddy or contact
- Creating and sending of instant messaging (IM) or text messages
- Reception of IM or text messages

While IM and presence are not required functionality for mobile clients, they do enable users to make their availability status visible to contacts and to view the availability status of contacts, thus improving productivity. Further, users can send enterprise-based IM messages rather than incurring costs for mobile Short Message Service (SMS) messages.

Corporate Directory Access

Some mobile clients and devices are capable of accessing LDAP enterprise directory services, including directory lookups and personal contact lists. While this is not a required feature for mobile devices and clients, it does provide productivity improvements for mobile users when they are able to access corporate directory information from their mobile device.

Enterprise Voicemail Services

Many mobile clients and devices are also capable of accessing enterprise voicemail services. Cisco mobile clients are capable of receiving enterprise message waiting indication whenever an unread voicemail is in the user's enterprise voicemail box and the mobile device is attached to the enterprise network. Further, mobile clients can be used to retrieve enterprise voicemail messages. Typically enterprise voicemail messages are retrieved when the user dials the voicemail system number and navigates to their voicemail box after providing required credentials. However, Cisco Jabber mobile clients provide the ability to retrieve voicemail messages from the voicemail box by downloading and displaying a list of all messages in the voicemail box and then by selecting individual messages to be downloaded to the mobile device for listening. This is sometimes referred to as visual voicemail. Both the mobile client and the enterprise voicemail system must be capable of providing and receiving message waiting indication (MWI), voicemail message information, and downloads of the messages over the network. Cisco Unity Connection supports visual voicemail through IMAP or VMREST, and it can provide MWI and voicemail lists and downloads.

Cisco Bring Your Own Device (BYOD) Infrastructure

Cisco mobile client applications such as Cisco Jabber provide core Unified Communications and collaboration capabilities, including voice, video, and instant messaging to users of mobile devices such as Android and Apple iOS smartphones and tablets. When a Cisco mobile client device is attached to the corporate wireless LAN, the client can be deployed within the Cisco Bring Your Own Device (BYOD) infrastructure.

Because Cisco mobile clients and devices rely on enterprise wireless LAN connectivity or remote secure attachment through VPN, they can be deployed within the Cisco Unified Access network and can utilize the identification, security, and policy features and functions delivered by the BYOD infrastructure.

The Cisco BYOD infrastructure provides a range of access use cases or scenarios to address various device ownership and access requirements. The following high-level access use case models should be considered:

- Basic Access — This use case enables basic Internet-only access for guest devices. This use case provides the ability to enable employee-owned personal device network connectivity without providing access to corporate resources.
- Limited Access — This use case enables full access to corporate network resources, but it applies exclusively to corporate-owned devices.
- Enhanced Access — This use case enables granular access to corporate network resources from employee-owned personal devices based on corporate policies.

Cisco collaboration mobile clients, whether running on corporate or personal devices, usually require access to numerous back-end on-premises enterprise application components for full functionality. For this reason the Limited or Enhanced Access use case scenarios generally apply to applications such as Cisco Jabber for Android or iPhone. The chief difference between these two access models is whether the client application is running on a corporate-owned or employee personal device.

In the case of cloud-based collaboration services, Cisco mobile clients and devices connect directly to the cloud through the Internet without the need for VPN or full enterprise network attachment. In these scenarios, user and mobile devices can be deployed using the Basic Access model because these use cases require only Internet access.

For more information about the Cisco BYOD infrastructure and BYOD access use cases, refer to the *Cisco Bring Your Own Device (BYOD) Smart Solution Design Guide*, available at

http://www.cisco.com/go/byoddesign

When deploying Cisco mobile clients and devices within the Cisco BYOD infrastructure, consider the following high-level design and deployment guidelines:

- The network administrator should strongly consider allowing voice and video-capable clients to attach to the enterprise network in the background (after initial provisioning), without user intervention, to ensure maximum use of the enterprises telephony infrastructure. Specifically, use of certificate-based identity and authentication helps facilitate an excellent user experience by minimizing network connection and authentication delay.
- In scenarios where Cisco mobile clients and devices are able to connect remotely to the enterprise network through a secure VPN:
  - The network administrator should weigh the corporate security policy against the need for seamless secure connectivity without user intervention in order to maximize utilization of the enterprise telephony infrastructure. The use of certificate-based authentication and enforcement of a device pin-lock policy provides seamless attachment without user intervention and functionality similar to two-factor authentication because the end user must possess the device and know the pin- lock to access the network. If two-factor authentication is mandated, then user intervention will be required in order for the device to attach remotely to the enterprise.

  - It is important for the infrastructure firewall configuration to allow all required client application network traffic to access the enterprise network. Failure to open access to appropriate ports and protocols at the corporate firewall could result in an inability of the Cisco mobile clients or devices to register to on-premises Cisco call control for voice and video telephony services and/or the loss of other client features such as enterprise directory access or enterprise visual voicemail.

- When enterprise collaboration applications such as Cisco Jabber are installed on employee-owned mobile devices, if the enterprise security policy requires the device to be wiped or reset to factory default settings under certain conditions, device owners should be made aware of the policy and encouraged to backup personal data from their device regularly.
- When deploying Cisco collaboration mobile clients and devices, it is important for the underlying network infrastructure from end-to-end to support the necessary QoS classes of service, including priority queuing for voice media and dedicated video and signaling bandwidth, to ensure the quality of client application voice and video calls and appropriate behavior of all features.

Cisco Mobile Clients and Devices

This section discusses the following Cisco mobile clients and devices:

- Cisco Jabber for Android and iPhone, page 25-71
- Cisco Jabber for iPad, page 25-76
- Cisco Jabber IM, page 25-77
- Cisco Cius, page 25-77

Cisco Jabber for Android and iPhone

This section describes characteristics and deployment considerations for Cisco Jabber.

Cisco Jabber mobile clients are available for Android, iPhone, and other Apple iOS mobile devices. Once the client application is downloaded from the appropriate store or market (Apple Application Store or Google Play, formerly Android Market) and installed on the Apple iOS or Android device, it can connect to the enterprise network and register to Unified CM as a SIP enterprise phone.

To provide registration and call control services to the Cisco Jabber dual-mode Android or iPhone client, the device must be configured within Unified CM as a **Cisco Dual-Mode for Android** or **iPhone** device type. Next, the mobile device must be configured to access the enterprise WLAN for connectivity based on the enterprise WLAN infrastructure and security policies. Alternatively the mobile device can be connected to the enterprise network via the mobile data network or over non-enterprise WLANs. Once the mobile device has been configured to access the enterprise network, when the Cisco Jabber client is launched, it will register the device to Unified CM. To integrate to Unified Mobility and to leverage handoff functionality, the mobile number of the Android or iPhone must be configured as a mobility identity associated to the Cisco Dual-Mode for Android or iPhone device within Unified CM.

The Cisco Jabber client is supported on the following devices:

- Android

  Various models in the Samsung Galaxy family of phones. (Consult the release notes referenced below for specific device and firmware support information.) These devices must be running firmware version 2.3, and in some cases version 4.0 might be required. Although not officially supported, Cisco Jabber for Android runs on many Android devices running version 2.3 or 4.0 with various degrees of limitations depending on the device. The WLAN interfaces of most Android devices support 802.11b, 802.11g, and 802.11n network connectivity. Some devices also support 802.11a.

- Apple iOS

  Various Apple iOS devices including iPhone and iPad. (Consult the release notes referenced below for specific device and firmware support information.) These devices must be running a minimum iOS version of 5.1.1. The WLAN interfaces of these Apple iOS devices support 802.11b, 802.11g, and 802.11n network connectivity.

For details on the latest specific device and firmware version support, refer to the product release notes for:

- Android

  http://www.cisco.com/en/US/products/ps11678/prod_release_notes_list.html
- iPhone

  http://www.cisco.com/en/US/products/ps11596/prod_release_notes_list.html

The Cisco Jabber for iPhone and Android clients not only provide voice-over-IP phone services but also provide directory lookup services when configured to access the enterprise LDAP directory and provide enterprise voicemail message waiting indication (MWI) and visual voicemail when integrated to Cisco Unity Connection.

The Cisco Jabber for iPhone and Android clients are capable of performing only manual hand-out as described in the section on Cisco Jabber Handoff, page 25-72.

For more information about the Cisco Jabber Android and iPhone clients, additional feature details, and supported hardware and software versions, refer to the Cisco Jabber documentation for:

- Android

  http://www.cisco.com/en/US/products/ps11678/tsd_products_support_series_home.html
- iPhone

  http://www.cisco.com/en/US/products/ps11596/tsd_products_support_series_home.html

Cisco Jabber Handoff

To properly deploy Cisco dual-mode clients such as Cisco Jabber, it is important to understand the nature of handoff operations within the client. The handoff method used by the Cisco Jabber dual-mode client depends on the **Transfer to Mobile Network** setting on the Cisco Dual-Mode for iPhone or Cisco Dual-Mode for Android device configuration page.

There are two methods of handoff, depending on the Transfer to Mobile Network setting:

- Mobility Softkey Method of Hand-Out, page 25-72

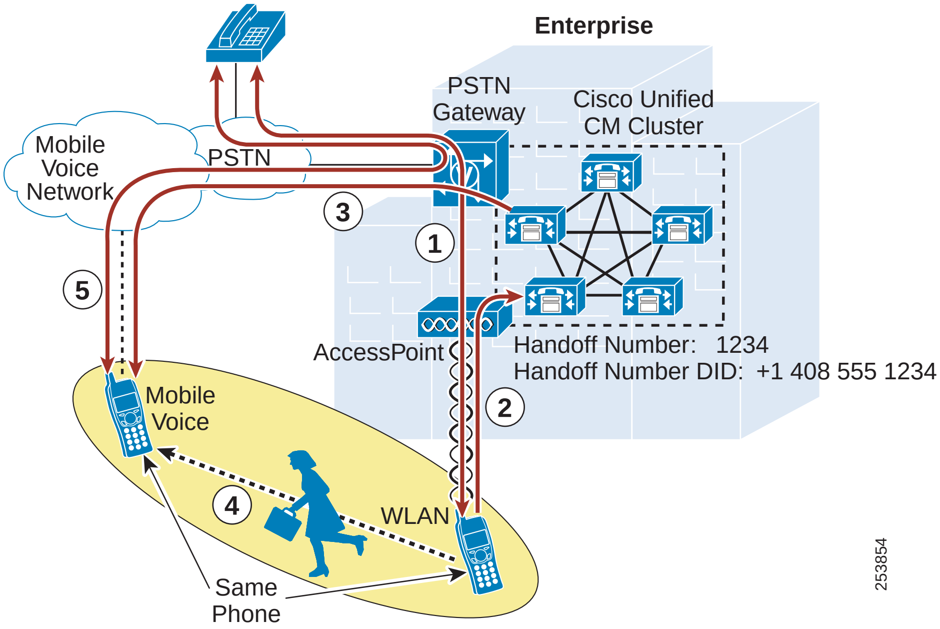
  With this method the Transfer to Mobile Network setting should be set to **Use Mobility Softkey (user receives call)**. In this type of handoff, the Unified CM system generates a call over the PSTN to the user's mobile number. This hand-out method is supported with all Cisco Jabber dual-mode clients.
- Handoff Number Method of Hand-Out, page 25-73

  With this method the Transfer to Mobile Network setting should be set to **Use HandoffDN Feature (user places call)**. In this type of handoff, the mobile client generates a call over the mobile voice network to the handoff number configured within the Unified CM system. This hand-out method is supported only with Cisco Jabber for iPhone clients.

Mobility Softkey Method of Hand-Out

The operation depicted in Figure 25-23 is of an active call on an iPhone or Android dual-mode device within the enterprise being moved manually from the WLAN interface to the mobile voice network or cellular interface of the device through the enterprise PSTN gateway. As shown, there is an existing call between the mobile client device associated to the enterprise WLAN and registered to Unified CM, and a phone on the PSTN network (step 1). Because this is a manual process, the user must select the Use Mobile Network button from the in-call menu within the Cisco Jabber client, which signals to Unified CM the intention to hand-out the call (step 2). Next Unified CM generates a call to the configured mobility identity number corresponding to this mobile device through the enterprise PSTN gateway (step 3). This call to the mobility identity is made to the mobile voice network or cellular interface of the iPhone or Android device. The user can now move out of the enterprise and away from WLAN network coverage (step 4). In the meantime, the inbound call from Unified CM is received at the mobile voice network interface, and the user must answer the call manually to complete the hand-out. Once the inbound call on the cellular interface is answered, the RTP stream that was traversing the WLAN is redirected to the PSTN gateway, and the call continues uninterrupted between the mobile client device and the original PSTN phone, with the call anchored in the enterprise gateway (step 5).

Figure 25-23 Cisco Jabber Dual-Mode Hand-Out (WLAN-to-Mobile Voice Network): Mobility Softkey Method



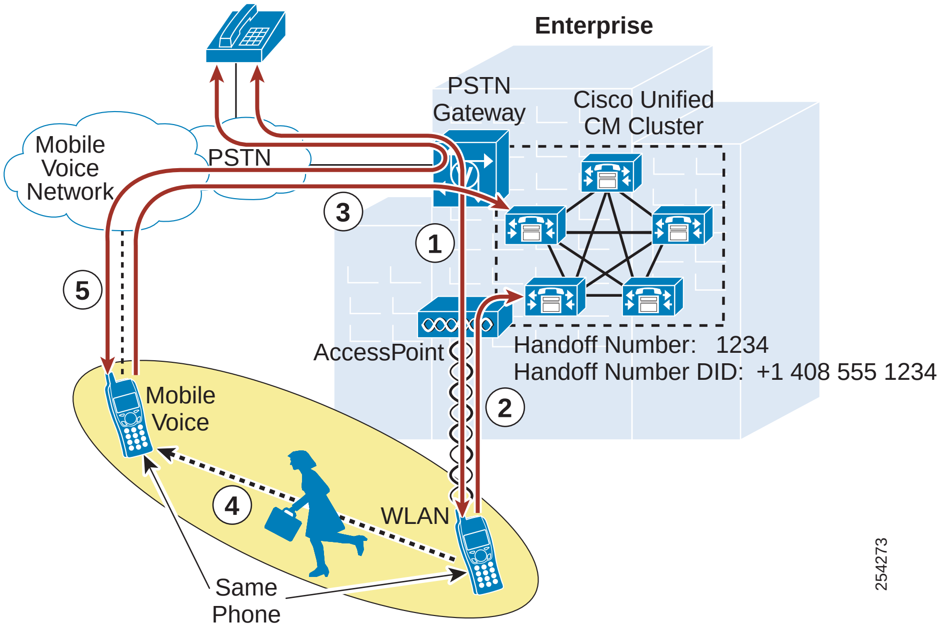
Handoff Number Method of Hand-Out

Figure 25-24 illustrates the same hand-out operation as in Figure 25-23, where an active call on an iPhone dual-mode phone within the enterprise is moved manually from the WLAN interface to the mobile voice network or cellular interface of the device through the enterprise PSTN gateway. However, in this case the Handoff Number method of hand-out is used.

Note The Handoff Number method of hand-out is supported only with Cisco Jabber for iPhone

As shown in Figure 25-24, there is an existing call between the iPhone dual-mode device associated to the enterprise WLAN and registered to Unified CM, and a phone on the PSTN network (step 1). Because this is a manual process, the user must select the Use Mobile Network button from the in-call menu within the Cisco Jabber dual-mode client, which signals to Unified CM the intention to hand-out the call (step 2). Next the Cisco Jabber client automatically generates a call through the cellular interface over the mobile voice network to the configured Handoff Number within the Unified CM system (step 3). The user can now move out of the enterprise and away from WLAN network coverage (step 4). In the meantime, the inbound call from the Cisco Jabber client is received by Unified CM. Assuming the inbound calling number matches the user's configured mobility identity, the RTP stream that was traversing the WLAN is redirected to the PSTN gateway, and the call continues uninterrupted between the Cisco Jabber mobile client and the original PSTN phone, with the call anchored in the enterprise gateway (step 5).

Figure 25-24 Cisco Jabber Dual-Mode Hand-Out: Handoff Number Method



Note The Handoff Number method of hand-out requires Unified CM to receive an inbound calling number from the PSTN network that matches the mobility identity number configured under the Cisco Dual Mode for iPhone device attempting the hand-out. If the caller ID is not sent by the iPhone, if the PSTN provider does not send the inbound caller ID to the enterprise, or if the inbound caller ID does not match the user's configured mobility identity, the hand-out operation will fail.

Note Cisco Jabber dual-mode clients do not support hand-in. In scenarios where an in-progress call is active between the dual-mode mobile voice network or cellular interface and an enterprise phone (or a PSTN phone with the call anchored in the enterprise gateway), the only way to move the call to the WLAN interface of the dual-mode device is to hang up the call and redial once the dual-mode client has connected to the enterprise network and registered to Unified CM.

Cisco Jabber for iPhone Desk Phone Integration

The Cisco Jabber for iPhone mobile client enables the user to move an active or held call from the user's desk phone to the iPhone device. This feature relies on CTI monitoring of the primary line of the user's desk phone as well as the call park feature.

The functionality provided by desk phone integration relies on active CTI monitoring of the primary line of the user's desk phone. Whenever an active or held call is sensed by the Cisco Jabber client, it prompts the user as to whether they want to transfer the call to the dual-mode device. If the user indicates they wish to transfer the call, the desk phone automatically parks the call and the mobile client automatically retrieves the call from the park number.

To enable desk phone integration, ensure that the user's end-user account is assigned to a CTI-enabled user group and that the user's desk phone is enabled to allow CTI control. In addition, the CTI Control Username field on the Cisco Dual-Mode for iPhone device must be configured with the user's end-user account userID.

Cisco Jabber for Android Desk Phone Integration

The Cisco Jabber for Android dual-mode client enables the user to move an active call from the Android device to the IP desk phone sharing a line with the mobile client device. This feature is invoked by placing the active call on hold through the Cisco Jabber client. When the call is placed on hold, the call can be resumed at either the shared-line IP desk phone or on the Cisco Jabber client.

WLAN Design Considerations for Cisco Jabber Mobile Clients

Consider the following WLAN guidelines when deploying Cisco Jabber mobile clients:

- Whenever possible, ensure that Cisco Jabber mobile clients roam on the WLAN only at Layer 2 so that the same IP address can be used on the WLAN interface of the device. In Layer 3 roaming scenarios where subnet boundaries are crossed due to device IP address changes, calls will be dropped.
- Deploy Cisco Jabber mobile clients on WLAN networks where the same SSID is used across all APs. Roaming between APs is much slower if SSIDs are different.
- Ensure all APs in the WLAN broadcast their SSID(s). If the SSID is not broadcast by the AP, the user may be prompted by the device to join other Wi-Fi networks or the device may automatically join other Wi-Fi networks. When this occurs the call is interrupted.
- Whenever possible, deploy Cisco Jabber mobile clients on the 5 GHz WLAN band (802.11a/n). 5 GHz WLANs provide better throughput and less interference for voice and video calls.

Interactions Between Cisco Jabber and Cisco Unified Mobility

The Cisco Jabber mobile clients can be integrated with Cisco Unified Mobility to leverage Cisco Mobile Connect, mid-call DTMF features, two-stage dialing, and single enterprise voicemail box mobile voicemail avoidance.

Integration with Unified Mobility requires the iPhone or Android dual-mode mobile phone number to be configured within Unified CM as a mobility identity associated with the Cisco Dual-Mode for iPhone or Cisco Dual-Mode for Android device. Once the mobile number is configured as a mobility identity within the system, Mobile Connect can be leveraged so that incoming calls to the user's enterprise number will be extended to the iPhone or Android dual-mode device through the mobile voice network as long as the iPhone or Android dual-mode device is not connected to the enterprise and not registered to Unified CM. In situations where the dual-mode device is connected to the enterprise and registered to Unified CM, an inbound call to the enterprise number will not be extended to the mobile voice network interface of the device. When the iPhone or Android dual-mode device is connected to the enterprise, only the WLAN or mobile data interface of the device will receive the inbound call. This prevents unnecessary consumption of enterprise PSTN gateway resources.

When not connected to the enterprise and not registered to Unified CM, the iPhone or Android dual-mode device can invoke mid-call features by means of DTMF and perform desk phone pickup for any enterprise anchored call. The dual-mode device can also leverage Mobile Voice Access and Enterprise Feature Access two-stage dialing features when making outbound calls to route these calls through the enterprise and anchor them in the enterprise PSTN gateway.

In addition to configuring a mobility identity for the iPhone or Android dual-mode device, you can configure additional mobile phone numbers or off-system phone numbers as remote destinations and associate them to the Cisco Dual-Mode for iPhone or Cisco Dual-Mode for Android device within Unified CM. When associating the mobility identity and additional remote destinations to the dual-mode device, you do not have to configure a remote destination profile.

For more information about the Cisco Unified Mobility feature set as well as design and deployment considerations, see Cisco Unified Mobility, page 25-32.

Cisco Jabber for iPad

This section describes characteristics and deployment considerations for Cisco Jabber for iPad.

Cisco Jabber for iPad is a mobile client for the Apple iPad, and it provides voice and video calling capabilities as well as enterprise visual voicemail and directory access. The Cisco Jabber for iPad client also provides XMPP-based IM and presence when integrated to on-premises Cisco IM and Presence services or cloud-based collaboration services such as Cisco WebEx Messenger.

Once the client application is downloaded from the Apple Application Store and installed on the iPad device, it can connect to the enterprise network and register to Unified CM or Cisco TelePresence Video Communication Server (VCS) as a SIP enterprise endpoint. To provide registration and call control services to the Cisco Jabber iPad client, the device must be configured within Unified CM or VCS. When registering to Unified CM call control services, the client device is configured as a **Cisco Jabber for Tablet** device type. When registering to VCS call control services, the client device is configured and provisioned using the *jabbertablet* provisioning template and Cisco TelePresence Management Suite (TMS).

Next, the client device must be configured to access the enterprise WLAN for connectivity based on the enterprise WLAN infrastructure and security policies. Alternatively the device can be connected to the enterprise network through the mobile data network (if the device supports mobile provider data and if mobile data services are enabled) or over non-enterprise WLANs. Once the client device has been configured to access the enterprise network, when the Cisco Jabber for iPad client is launched, it will register the device to Unified CM or VCS for voice and video call control services.

The Cisco Jabber for iPad client is supported on the Apple iOS iPad 2 or the new iPad (third generation). WLAN interfaces of Apple iPad devices support 802.11a, 802.11b, 802.11g, and 802.11n.

For details on the latest specific device and firmware version support, refer to the product release notes at

http://www.cisco.com/en/US/products/ps12430/prod_release_notes_list.html

For more information about the Cisco Jabber for iPad client, additional feature details, and supported hardware and software versions, refer to the Cisco Jabber for iPad documentation at

http://www.cisco.com/en/US/products/ps12430/tsd_products_support_series_home.html

For more information about deploying Cisco Jabber for iPad, refer to the *Cisco Jabber for iPad Deployment Guide*, available at

http://www.cisco.com/en/US/products/ps12430/product_solution_overview_list.html

Cisco Jabber IM

This section describes characteristics and deployment considerations for Cisco Jabber IM.

Cisco Jabber IM is an XMPP-based instant messaging (IM) and presence mobile client for Android, BlackBerry, iPhone, and other Apple iOS mobile devices. Once the client application is downloaded from the appropriate application store or download site (Apple Application Store, Google Play Store, or web) and installed on the Android, Apple iOS, or BlackBerry device, it can connect to the enterprise network for on-premises enterprise IM and presence services as provided by Cisco IM and Presence or to the Cisco WebEx Messenger service for off-premises enterprise IM and presence services. The mobile device must be configured to access the enterprise WLAN for connectivity based on the enterprise WLAN infrastructure and security policies. Alternatively the mobile device can be connected to the enterprise network through the mobile data network or over non-enterprise WLANs. Once the mobile device has been configured to access the enterprise network, when the Cisco Jabber IM client is launched, it will connect and register with the appropriate IM and presence service.

A Cisco Jabber IM client is available for most Android devices running Android 2.3, 4.0, or 4.1. A Cisco Jabber IM client is available for Apple iOS devices, including iPhone and iPad, with a minimum Apple iOS version of 4.2. A Cisco Jabber IM client is also available for a wide range of BlackBerry devices running BlackBerry OS 4.6 and later versions.

For more information on Cisco Jabber IM clients, refer to the following documentation:

- *Cisco Jabber IM for BlackBerry* data sheet

  http://www.cisco.com/en/US/products/ps11763/products_data_sheets_list.html
- *Cisco Jabber IM for Android* data sheet

  http://www.cisco.com/en/US/products/ps11678/products_data_sheets_list.html
- *Cisco Jabber IM for iPhone* data sheet

  http://www.cisco.com/en/US/products/ps11596/products_data_sheets_list.html

Interactions Between Cisco Jabber IM and Cisco Jabber

The Cisco Jabber IM client enables escalation or cross-launch of the Cisco Jabber client (if installed) from the contact screen to an enterprise voice or video over IP call. In turn, the Cisco Jabber client is able to cross-launch Cisco Jabber IM (if installed) from the contact screen to an enterprise IM or chat.

Cisco Cius

This section describes characteristics and deployment considerations for Cisco Cius.

Cisco Cius is an Android-based enterprise tablet that provides native voice and video over WLAN or mobile data network when registered to Unified CM as a SIP enterprise device. To provide registration and call control services to the Cisco Cius native phone client, the device must be configured within Unified CM as a **Cisco Cius** device type. Next, the mobile device must be configured to access the enterprise WLAN for connectivity based on the enterprise WLAN infrastructure and security policies. Alternatively the mobile device can be connected to the enterprise network via the mobile data network if a mobile data network interface is available or over non-enterprise WLANs. Once the mobile device has been configured to access the enterprise network, when the Cius device is powered on, it will register the device to Unified CM.

The Cisco Cius supports not only voice calling but also video calling when connected with other Cius or other video-capable endpoints such as the Cisco Unified IP Phone 9971. As with other mobile clients and devices capable of voice and video over WLAN, Cisco recommends associating the Cius to the 5 GHz WLAN band (802.11a/n), which will provide better throughput and less interference for voice and video calls.

For more information about deploying Cisco Cius wirelessly, including WLAN voice and video call capacity, refer to the *Cisco Cius Deployment Guide*, available at

http://www.cisco.com/en/US/products/ps11156/products_implementation_design_guides_list.html

The Cisco Cius also provides native support for enterprise directory access, enterprise visual voicemail, and XMPP-based IM and presence.

For more information about the Cisco Cius mobile device, additional feature and functionality details, and supported hardware and software versions, refer to the Cisco Cius product documentation available at

http://www.cisco.com/en/US/products/ps11156/tsd_products_support_series_home.html

Cisco AnyConnect Mobile Client

The Cisco AnyConnect mobile client provides secure remote connectivity capabilities for Cisco Cius and Cisco Jabber mobile device clients, enabling connectivity over mobile data networks and non-enterprise WLANs. The Cisco AnyConnect mobile client can be downloaded from the Apple Application Store or Google Play (formerly Android Market). Cisco AnyConnect is a native client application on the Cius. This client application provides SSL VPN connectivity for Apple iOS and Android mobile devices through the Cisco AnyConnect VPN solution available with the Cisco Adaptive Security Appliance (ASA) head-end.

When employing VPN network connectivity for connections over the mobile data network or public or private Wi-Fi hot spots, it is important to deploy a high-bandwidth secure VPN infrastructure that adheres to the enterprise's security requirements and policies. Careful planning is needed to ensure that the VPN infrastructure provides high bandwidth, reliable connections, and appropriate session or connection capacity based on the number of users and devices using this connectivity.

For more information on secure remote VPN connectivity using Cisco AnyConnect, refer to the Cisco AnyConnect Secure Mobile Client documentation available at

http://www.cisco.com/en/US/products/ps10884/tsd_products_support_series_home.html

High Availability for Cisco Mobile Clients and Devices

Although mobile devices and in particular dual-mode phones by their nature are highly available with regard to network connectivity (when the WLAN network is unavailable, the mobile voice and data networks can be used for voice and data services), enterprise WLAN and IP telephony infrastructure high availability must still be considered.

First, the enterprise WLAN must be deployed in a manner that provides redundant WLAN access. For example, APs and other WLAN infrastructure components should be deployed so that the failure of a wireless AP does not impact network connectivity for the mobile device. Likewise, WLAN management and security infrastructure must be deployed in a highly redundant fashion so that mobile devices are always able to connect securely to the network. Controller-based wireless LAN infrastructures are recommended because they enable centralized configuration and management of enterprise APs, thus allowing the WLAN to be adjusted dynamically based on network activity and AP failures.

Next, VPN infrastructure components, including the Cisco ASA head-end VPN or AnyConnect session terminator, should be deployed in a highly redundant fashion so that loss of a VPN session terminator does not impact or prevent remote secure enterprise connectivity for the mobile client.

Next, Unified CM call processing and registration service high availability must be considered. Just as with other devices within the enterprise that leverage Unified CM for call processing services, mobile client devices must register with Unified CM. Given the redundant nature of the Unified CM cluster architecture, which provides primary and backup call processing and device registration services, mobile device registration as well as call routing are still available even in scenarios in which a Unified CM server node fails.

Similar considerations apply to PSTN access. Just as with any IP telephony deployment, multiple PSTN gateways and call routing paths should be deployed to ensure highly available access to the PSTN. This is not unique to mobile client device deployments, but is an important consideration none the less.

Capacity Planning for Cisco Mobile Clients and Devices

Capacity planning considerations for Cisco mobile clients and devices, including dual-mode phones, are the same as for other IP telephony endpoints or devices that rely on the IP telephony infrastructure and applications for registration, call processing, and PSTN access services.

When deploying Cisco mobile clients and devices, it is important to consider the registration load on Unified CM as well as the Unified Mobility limits. A single Unified CM cluster is capable of handling a maximum of 40,000 device configurations and registrations. When deploying mobile clients and devices, you must consider the per-cluster maximum device support, and you might have to deploy additional call processing clusters to handle the added load.

In addition, as previously mentioned, the maximum number of remote destinations and mobility identities within a single Unified CM cluster is 15,000. Because most dual-mode mobile client devices will likely be integrated with Unified Mobility to take advantage of features such as Mobile Connect, single enterprise voicemail box mobile voicemail avoidance, desk phone pickup, and two-stage dialing, the mobile phone number of each of these dual-mode mobile devices must be configured as a mobility identity within the Unified CM cluster. This is necessary to facilitate integration to Unified Mobility as well as to facilitate handoff in some cases. Therefore, when integrating these dual-mode devices with Unified Mobility, it is important to consider the overall remote destination and mobility identity capacity of the Unified CM cluster to ensure sufficient capacity exists. If additional users or devices are already integrated to Unified Mobility within the system, they can limit the amount of remaining remote destination and mobility identity capacity available for dual-mode devices.

CTI capacity must also be considered when deploying the Cisco Jabber client for iPhone with desk phone integration. Because this feature relies on CTI monitoring of the primary line of the user's desk phone, each Cisco Jabber for iPhone mobility user enabled for desk phone integration will consume a CTI connection on the Unified CM system. This load must be considered in relation to the overall CTI capacity of the system.

Overall call processing capacity of the Unified CM or VCS system and PSTN gateway capacity must also be considered when deploying mobile client devices. Beyond handling the actual mobile device configuration and registration, these systems must also have sufficient capacity to handle the added BHCA impact of these mobile devices and users. Likewise, it is critical to ensure sufficient PSTN gateway capacity is available to accommodate mobile devices. This is especially the case for dual-mode mobile devices that are integrated to Unified Mobility because the types of users that would have dual-mode devices are typically highly mobile. Highly mobile users typically generate more enterprise PSTN gateway load from mobility features such as Mobile Connect, where an incoming call to a mobile user's enterprise number generates one or more calls to the PSTN, or from two-stage dialing, where a user makes a call through the enterprise by leveraging the enterprise PSTN gateway.

Finally, just as with enterprise mobility deployments, 802.11 WLAN call capacity must be considered when deploying Cisco mobile clients and device. As previously mentioned, a maximum of 27 VoWLAN calls or a maximum of 8 VVoWLAN calls are possible per 802.11 channel cell. This assumes no Bluetooth when devices are deployed on the 2.4 GHz band, 24 Mbps or higher data rates for VoWLAN calls, and 720p video resolution with bit rates up to 1 Mbps for VVoWLAN calls. Actual call capacity could be lower depending on the RF environment, wireless endpoint type, and WLAN infrastructure. See Capacity Planning for Campus Enterprise Mobility, page 25-9, for more details regarding 802.11 WLAN call capacity.

The above considerations are certainly not all unique to mobile clients and devices. They apply to all situations in which devices and users are added to Unified CM, resulting in additional load to the overall Unified Communications System.

For more information on general system sizing, capacity planning, and deployment considerations, see the chapter on Unified Communications Design and Deployment Sizing Considerations, page 29-1.

Design Considerations for Cisco Mobile Clients and Devices

Observe the following design recommendations when deploying Cisco mobile clients and devices:

- Dual-mode mobile devices must be capable of dual transfer mode (DTM) in order to be connected simultaneously to both the mobile voice and data network and the WLAN network so that the device is reachable and able to make and receive calls on both the cellular radio and WLAN interface of the device. In some cases, proper dual-mode client operation might not be possible if mobile voice and data networks do not support dual-connected devices.
- WLAN APs should be deployed with a minimum cell overlap of 20%. This overlap ensures that a mobile device can successfully roam from one AP to the next as the device moves around within a location, while still maintaining voice and data network connectivity.
- WLAN APs should be deployed with cell power level boundaries (or channel cell radius) of -67 dBm in order to minimize packet loss. Furthermore, the same-channel cell boundary separation should be approximately 19 dBm. A same-channel cell separation of 19 dBm is critical for ensuring that APs or clients do not cause co-channel interference to other devices associated to the same channel, which would likely result in poor voice and video quality.
- Whenever possible rely on the 5 GHz WLAN band (802.11a/n) for connecting mobile clients and devices capable of generating voice and video traffic. 5 GHz WLANs provide better throughput and less interference for voice and video calls.
- The enterprise wired and wireless LAN should be deployed and configured to support the necessary end-to-end QoS classes of service, including priority queuing for voice media and dedicated video and signaling bandwidth, to ensure the quality of client application voice and video calls and the appropriate behavior of all features. While most clients mark traffic appropriately at Layer 3 based on Cisco QoS recommendations, appropriate Layer 2 WLAN UP marking is dependent on the client device and vendor implementation. For this reason, Layer 2 marking is not consistent across platforms and as such cannot be relied upon.
- Because mobile devices are similar to desktop computers and can generate a large variety of data and real-time traffic, these devices are typically considered untrusted. For this reason, the network should be configured to re-mark all traffic from these client devices based on port number and/or protocol. Likewise, rate limiting and policing on ingress to the network is recommended.
- Cisco recommends using only an enterprise-class voice and video optimized WLAN network for connecting mobile devices and clients. While most mobile client devices are capable of attaching to public or private WLAN access points or hot spots for connecting back to the enterprise through the Internet for call control and other Unified Communications services, Cisco cannot guarantee voice and video quality for these types of connections.

- When deploying Cisco collaboration mobile clients and devices on a Cisco Bring Your Own Device (BYOD) infrastructure, administrators should consider a network attachment method that does not require user intervention and which maximizes utilization of the IP telephony infrastructure. Further, for remote connectivity scenarios, all relevant ports must be opened in the corporate firewall in order for Cisco mobile clients and devices to be able to access collaboration services.
- If corporate policy dictates that the BYOD infrastructure must remotely wipe or factory-reset lost or stolen mobile devices, employees using personal mobile devices should be aware of the policy and should regularly back up personal data.
- The Unified Mobility Mobile Connect feature will not extend incoming calls to the dual-mode device's configured mobility identity if the dual-mode device is inside the enterprise and registered to Unified CM. This is by design in order to reduce utilization of enterprise PSTN resources. Because the dual-mode device registers to Unified CM, the system knows whether the device is reachable inside the enterprise; and if it is, there is no reason to extend the call to the PSTN in order to ring the dual-mode device's cellular voice radio. Only when the dual-mode device is unregistered will Mobile Connect extend incoming calls to the user's enterprise number out to the mobility identity number on the PSTN.
- When you deploy mobile devices, Cisco recommends normalizing required dialing strings so that users are able to maintain their dialing habits, whether the mobile device is connected to the enterprise or not. Because dialing on the mobile network is typically done using full E.164 (with or without a preceding '+') and mobile phone contacts are typically stored with full E.164 numbers, Cisco recommends configuring the enterprise dial plan to accommodate full E.164 or full E.164 with preceding '+' for mobile client devices. By configuring the enterprise dial plan in this manner, you can provide the best possible end-user dialing experience so that users do not have to be aware of whether the device is registered to Unified CM.
- Cisco recommends that dual-mode phone users rely exclusively on the mobile voice network for making emergency calls and determining device and user location. This is because mobile provider networks typically provide much more reliable location indication than WLAN networks. To ensure that dual-mode phones rely exclusively on the mobile voice network for emergency and location services, configure the Emergency Numbers field of the dual-mode devices within Unified CM with emergency numbers such 911, 999, and 112 in order to force these calls over the mobile voice network. Dual-mode phone users should be advised to make all emergency calls over the mobile voice network rather than the enterprise network. Although making emergency calls over corporate WLANs or mobile data networks is not recommended, mobile devices that do not have cellular voice radios are capable of making calls only through these data interfaces. Mobile devices that do not have cellular voice radios should not be relied upon for making emergency calls.
- When deploying Cisco Jabber for iPhone with desk phone integration, the end-user account for the Cisco Jabber user must be enabled for CTI. In addition, call park should be configured at a system level so that the desk phone can auto-park the call and the Cisco Jabber client can retrieve it whenever a call is moved from the desk phone to the Cisco Jabber client. CTI overhead of this feature should be considered when sizing the overall Unified CM system.
- When deploying Cisco Jabber for iPhone or Android mobile clients, configure the WLAN network to accommodate the following deployment guidelines:
  - Minimize roaming of Cisco Jabber for iPhone and Android mobile devices at Layer 3 on the WLAN. Layer 3 roaming, where a device IP address changes, will result in longer roam times and dropped voice packets and could even result in dropped calls.
  - Configure the same SSID across all APs utilized by the Cisco Jabber mobile client devices within the WLAN to ensure the fastest AP-to-AP roaming.

  - Configure all enterprise WLAN APs to broadcast their SSIDs in order to prevent mid-call prompts to join other APs within the WLAN infrastructure, which could result in interrupted calls.

- Provide sufficient wireless voice and video call capacity on the enterprise wireless network for Cisco mobile clients and devices by deploying the appropriate number of wireless APs to handle the desired call capacity based on mobility-enabled user BHCA rates. Each 802.11g/n (2.4 GHz) or 802.11a/n (5 GHz) channel cell can support a maximum of 27 simultaneous voice-only calls with 24 Mbps or higher data rates. Each 802.11g/n (2.4 GHz) or 802.11a/n (5 GHz) channel cell can support a maximum of 8 simultaneous video calls assuming 720p video resolution at up to 1 Mbps bit rate. For 2.4 GHz WLAN deployments, Bluetooth must be disabled to achieve this capacity. Actual call capacity could be lower depending on the RF environment, wireless endpoint type, and WLAN infrastructure.

CHAPTER 26

Cisco Unified Contact Center

Revised: June 28, 2012; OL-27282-05

This chapter describes the Cisco Unified Contact Center solutions available with the Cisco Unified Communications System. It includes information on Cisco products such as Cisco Unified Contact Center Express, Cisco Unified Contact Center Enterprise, and Cisco Unified Customer Voice Portal. It also covers the design considerations for deploying these Cisco Unified Contact Center products with Cisco Unified Communications Manager and other Unified Communications components.

This chapter covers the following topics:

This chapter starts with a high-level overview of the main Cisco Unified Contact Center Portfolio. Then it covers the various Unified Communications deployment models for contact centers. Finally, it discusses design considerations on topics such as bandwidth, latency, Cisco Unified Communications Manager integration, and sizing.

The intent of this chapter is not to provide details on each contact center product and their various components but rather to discuss the design considerations for their integration with the Cisco Unified Communications System. Detailed design guidance for each Unified Contact Center product is covered in specific Solution Reference Network Design (SRND) guides for the Cisco Unified Contact Center Express, Cisco Unified Contact Center Enterprise, and Cisco Unified Customer Voice Portal products. These product-specific SRNDs are available at

http://www.cisco.com/go/ucsrnd

What's New in This Chapter

Table 26-1 lists the topics that are new in this chapter or that have changed significantly from previous releases of this document.

Table 26-1 New or Changed Information Since the Previous Release of This Document

New or Revised Topic	Described in:	Revision Date
Minor updates for Cisco Unified Communications System Release 9.0	Various sections throughout this chapter	June 28, 2012

Cisco Contact Center Architecture

This chapter discusses the following main Cisco Contact Center products:

- Cisco Unified Contact Center Enterprise (Unified CCE)
- Cisco Unified Customer Voice Portal (Unified CVP)
- Cisco Unified Contact Center Express (Unified CCX)

This chapter also discusses Cisco MediaSense, which can be deployed with a Cisco Unified Contact Center application or even in a non-contact center deployment.

For customers who need a basic contact center with limited functionality, the hunt pilot queuing in Cisco Unified CM is an available option. With this option enabled, callers to the hunt pilot can be put in queue to wait for an available agent, and periodic announcements can be played while the callers are in queue. However, unlike the full-featured contact center products, the hunt pilot queuing option lacks many of the contact center functionality such as agent desktop, supervisor, and reporting capabilities. If customers require complete contact center functionality, Cisco Unified Contact Center Enterprise or Cisco Unified Contact Center Express product should be used. For more information on the hunt pilot queuing option, refer to the Cisco Unified CM documentation available at

http://www.cisco.com/en/US/products/sw/voicesw/ps556/tsd_products_support_series_home.html

Cisco Unified Contact Center Enterprise

Cisco Unified Contact Center Enterprise (Unified CCE) provides a VoIP contact center solution that enables you to integrate inbound and outbound voice applications with Internet applications, including real-time chat, Web collaboration, and email. This integration provides for unified capabilities, helping a single agent support multiple interactions simultaneously, regardless of the communications channel the customer has chosen. Because each interaction is unique and may require individualized service, Cisco provides contact center solutions to manage each interaction based on virtually any contact attribute. The Unified CCE deployments are typically used for large size contact centers and can support thousands of agents.

Unified CCE employs the following major software components:

- Call Router

  The Call Router makes all the decisions on how to route a call or customer contact.

- Logger

  The Logger maintains the system database that stores contact center configurations and temporarily stores historical reporting data for distribution to the data servers. The combination of Call Router and Logger is called the *Central Controller*.

- Peripheral Gateway

  The Peripheral Gateway (PG) interfaces to various "peripheral" devices, such as Unified CM, Cisco Unified IP Interactive Voice Response (Unified IP IVR), Unified CVP, or multichannel products. A Peripheral Gateway that interfaces with Unified CM is also referred to as an *Agent PG*.

- CTI Server and CTI Object Server (CTI OS)

  The CTI Server and CTI Object Server interface with the agent desktops. Agent desktops can be based on the Cisco Agent Desktop (CAD) solution, Cisco CTI Desktop Toolkit, or customer relationship management (CRM) connectors to third-party CRM applications.

- Administration & Data Server

  The Administration & Data Server provides a configuration interface as well as real-time and historical data storage.

The Cisco Unified CCE solution is based on the integration with Cisco Unified Communications Manager (Unified CM), which controls the agent phones. For deployments without Unified CM but with traditional ACD, use Cisco Unified Intelligent Contact Management Enterprise (Unified ICME) instead of Unified CCE.

The queuing and self-service functions are provided by Cisco Unified IP Interactive Voice Response (Unified IP IVR) or Cisco Unified Customer Voice Portal (Unified CVP) and are controlled by the Unified CCE Call Router.

Most of the Unified CCE servers are required to be redundant, and the redundant instances are referred to as side A and side B instances. For example, Call Router A and Call Router B are redundant instances of the Call Router component running on two different servers.

Cisco Unified Customer Voice Portal

Cisco Unified Customer Voice Portal (Unified CVP) provides carrier-class IVR services on Voice over IP (VoIP) networks. It can perform basic prompt-and-collect or advanced self-service applications with CRM database integration and with automated speech recognition (ASR) and text-to-speech (TTS) integration. Unified CVP also provides IP-based call switching services by routing and transferring calls between voice gateways and IP endpoints.

Unified CVP is based on the Voice Extension Markup Language (VXML), which is an industry standard markup language similar to HTML and which is used to develop IVR services that leverage the power of web development and content delivery.

Unified CVP can be deployed standalone or integrated with Unified CCE to offer self-service and queuing functions. It supports voice calls as well as video calls.

The Unified CVP solution employs the following main components:

- Unified CVP Call Server

  The Unified CVP Call Server provides call control capabilities for SIP and H.323 through the SIP and H.323 services. The Unified CVP Call Server can also integrate with the Unified CCE Call Router through the Intelligent Contact Management (ICM) service. The IVR service allows the server to run VXML Micro applications and to create VoiceXML pages.

- Unified CVP VXML Server

  This component executes complex IVR applications by exchanging VoiceXML pages with the VoiceXML gateway's built-in voice browser. Unified CVP VXML Server applications are written using Cisco Unified Call Studio and are deployed to the Unified CVP VXML Server for execution. Note that there is no RTP traffic going through the Unified CVP Call Server or the Unified CVP VXML Server.

- Cisco Voice Gateway

  The Cisco Voice Gateway is the point at which a call enters or exits the Unified CVP system. The Cisco Voice Gateway could have a TDM interface to the PSTN. Alternatively, Cisco Unified Border Element could be used when the interface to the PSTN is an IP voice trunk.

- Cisco VoiceXML Gateway

  The VoiceXML Gateway hosts the Cisco IOS Voice Browser. This component interprets VoiceXML pages from either the Unified CVP Server IVR Service or the Unified CVP VXML Server. The VoiceXML Gateway can play prompts based on .wav files to the caller and can accept input from the caller through DTMF input or speech (when integrated with Automatic Speech Recognition). It then returns the results to the controlling application and waits for further instructions.

  The Cisco VoiceXML Gateway can be deployed on the same router as the Cisco Voice Gateway. This model is typically desirable in deployments with small branch offices. But the VoiceXML Gateway can also run on a separate router platform, and this model might be desirable in large centralized deployments with multiple voice gateways.

For more information, refer to the latest version of the *Cisco Unified Customer Voice Portal SRND*, available at

http://www.cisco.com/go/ucsrnd

Cisco Unified Contact Center Express

Cisco Unified Contact Center Express (Unified CCX) meets the needs of departmental, enterprise branch, or small to medium-sized companies that need easy-to-deploy, easy-to-use, highly available and sophisticated customer interaction management for up to 400 agents. It is designed to enhance the efficiency, availability, and security of customer contact interaction management by supporting a highly available virtual contact center with integrated self-service applications across multiple sites.

Unified CCX can integrate with Unified CM by means of JTAPI or with Unified CME by means of SIP.

All the Unified CCX components, including the Unified CCX engine, Unified CCX database, CAD Server, Unified CCX Outbound Dialer, and Express E-mail Manager, are installed on a single server. When Unified CCX is integrated with Unified CM, a second Unified CCX server can be added to provide system redundancy.

Unified CCX has built-in email, outbound dialer, and agent silent monitoring and recording capabilities. It can integrate with video endpoints such as Cisco TelePresence and can support advanced features such as Automated Speech Recognition (ASR) and Text to Speech (TTS), HTTP, and VXML. It also supports products such as Cisco Unified Workforce Optimization to optimize performance and quality of the contact center.

Cisco Unified IP IVR shares the same software architecture as Unified CCX. It provides prompting, collecting, and queuing capability for the Unified CCE solution. It could also be used as a standalone self-service application.

Administration and Management

Cisco Contact Center products have built-in administration and management capabilities. For example, Unified CCE can be administered with the Configuration Manager tool that is installed with Unified CCE, and Unified CVP can be administered with the Unified CVP Operations Console, also known as Operations, Administration, Maintenance, and Provisioning (OAMP).

In addition, Cisco Unified Contact Center Management Portal (Unified CCMP) can be deployed to simplify the operations and procedures for performing basic administrative functions such as managing agents and equipment. Unified CCMP is a browser-based management application designed for use by contact center system administrators, business users, and supervisors. It is a dense multi-tenant provisioning platform that overlays the Cisco Unified CCE, Unified ICM, Unified CM, and Unified CVP equipment.

Reporting

Cisco Unified Intelligence Center (Unified IC) is the main reporting tool for the Cisco Contact Center solutions. It is supported by Unified CCE, Unified CCX, and Unified CVP. This platform is a web-based application offering many Web 2.0 features, high scalability, performance, and advanced features such as the ability to integrate data from other Cisco Unified Communications products or third-party data sources.

Cisco Unified Intelligence Center gets source data from a database, such as an Unified CCE Administration & Data Server database or the Unified CVP Reporting Informix database. Reports are then generated and provided to a reporting client.

Multichannel Support

The Cisco Unified Enterprise solution supports web interaction and email interaction for multichannel support. Cisco Unified Web Interaction Manager (Unified WIM) technology helps ensure that communication can be established from nearly any web browser. Cisco Unified E-Mail Interaction Manager (Unified EIM) provides inbound email routing, automated or agent assisted email responses, real-time and historical reporting, and role-based hierarchical rights management for agents, supervisors, administrators, and knowledge base administrators.

For more design information on these products, refer to the *Cisco Unified Web and E-Mail Interaction Manager Solution Reference Network Design Guide*, available at

http://www.cisco.com/en/US/products/ps7236/products_implementation_design_guides_list.html

Recording and Silent Monitoring

Cisco Unified Contact Center solutions provide recording and silent monitoring capabilities based on the following options:

- The SPAN feature in Cisco switches

  This feature replicates the network traffic to a destination port to which a Cisco contact center server is connected.

- The ability of the phone to span the voice stream to the PC that is connected to it

  In this case, the agent desktop receives the voice packets and sends them to a recording server or to a supervisor desktop for silent monitoring.
- Unified CM and media replication by the built-in-bridge (BIB) in Cisco IP Phones

  With this option, Unified CM is involved in setting up the recording flows and can perform call admission control for those flows.

Cisco MediaSense

Cisco MediaSense is an IP-based media (voice and video) recording and playback system that implements the Open Recording Architecture (ORA) open interfaces. Cisco MediaSense is integrated into the Cisco Unified Communications architecture and provides a recording solution for both contact center deployments and non-contact center deployments. Recording can be accomplished by media forking in Cisco Unified IP Phones, where the built-in bridge (BIB) is used to replicate media to the Cisco MediaSense recording server. Recording can also be accomplished at the Cisco Unified Border Element, thus allowing all media flowing to or from the caller to get recorded, including possible interaction between the caller and an Interactive Voice Response (IVR) system. In addition, an IP phone user or SIP endpoint device may call the Cisco MediaSense system directly in order to leave a recording consisting only of media generated by that user. Such recordings may include video as well as audio, offering a simple and easy method for recording video blogs and podcasts.

Cisco MediaSense supports a redundant, highly available architecture. It can be deployed as a non-redundant, single server or as a highly available, redundant system with two recording servers in active/active mode. Additional servers can be added to expand storage capacity.

For more details on the Cisco MediaSense recording system, refer to the *Solution Reference Network Design for Cisco MediaSense*, available at

http://www.cisco.com/en/US/products/ps11389/products_implementation_design_guides_list.html

Contact Center Deployment Models

This section describes the various design models used for deploying Cisco Unified Contact Center solutions. For more details on these deployment models, refer to the Cisco Unified Contact Center SRNDs, available at http://www.cisco.com/go/ucsrnd

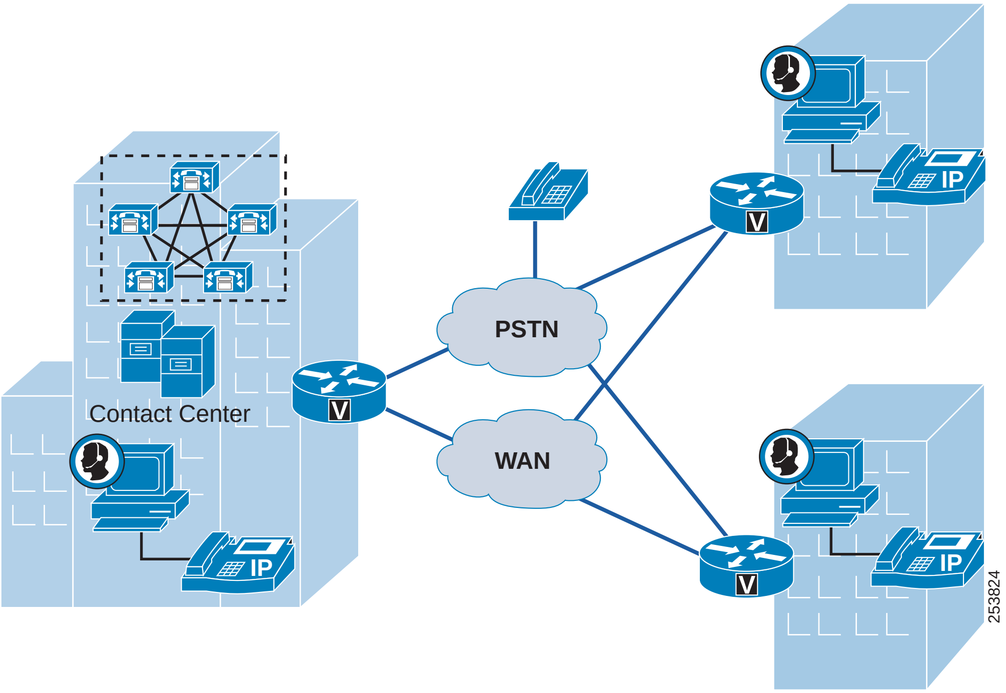
Single-Site Contact Center

In this deployment, all the components such as call processing servers, voice gateways, and contact center servers are in the same site. Agents and supervisors are also located at that site. The main benefit of the single-site deployment model is that there is no WAN connectivity required and, therefore, no need to use a low-bandwidth codec such as G.729, transcoders, compressed Real-Time Transport Protocol (cRTP), or call admission control.

Multisite Contact Center with Centralized Call Processing

A multisite deployment with centralized call processing consists of a single call processing cluster that provides services for many remote sites and uses the IP WAN. Cisco Contact Center applications (Unified CCE, Unified CCX, and Unified CVP) are also typically centralized to reduce the overall costs of management and administration. Figure 26-1 illustrates this type of deployment.

Figure 26-1 Multisite Contact Center with Centralized Call Processing



Because the agents or the voice gateways in this type of deployment are located in remote sites, it is important to consider the bandwidth requirements between the sites. It is also important to carefully configure call admission control, Quality of Service (QoS), codecs, and so forth. For more information on the general design considerations for Unified Communications solutions, refer to the chapter on Unified Communications Deployment Models, page 5-1.

Contact center deployments in a Unified Communications system typically have the following additional bandwidth requirements:

- The traffic volume handled by the agents is higher than that of typical users, and therefore voice and signaling traffic is also higher for agents.
- Agents and supervisors use desktops with screen popup, reports and statistics, and so forth. This causes data traffic between the agent or supervisor desktops and the contact center servers. In addition, bandwidth calculations must account for reporting information if, for example, an agent or supervisor is remote and pulls data from a server in a central location. For more information and guidance, refer to the design guides for the individual Cisco Contact Center products, available at http://www.cisco.com/go/ucsrnd.
- Depending on type of IVR solution, there could be traffic between the voice gateway and the IVR system. For example, if the voice gateways are distributed and calls arrive at a voice gateway located in a remote site with Unified IP IVR, there would be voice traffic across the WAN between the voice

gateway and Unified IP IVR. With Unified CVP, the call could be queued at the remote site, with the VXML Gateway providing call treatment and queuing and therefore avoiding voice traffic across the WAN for IVR and reducing overall WAN bandwidth requirements.

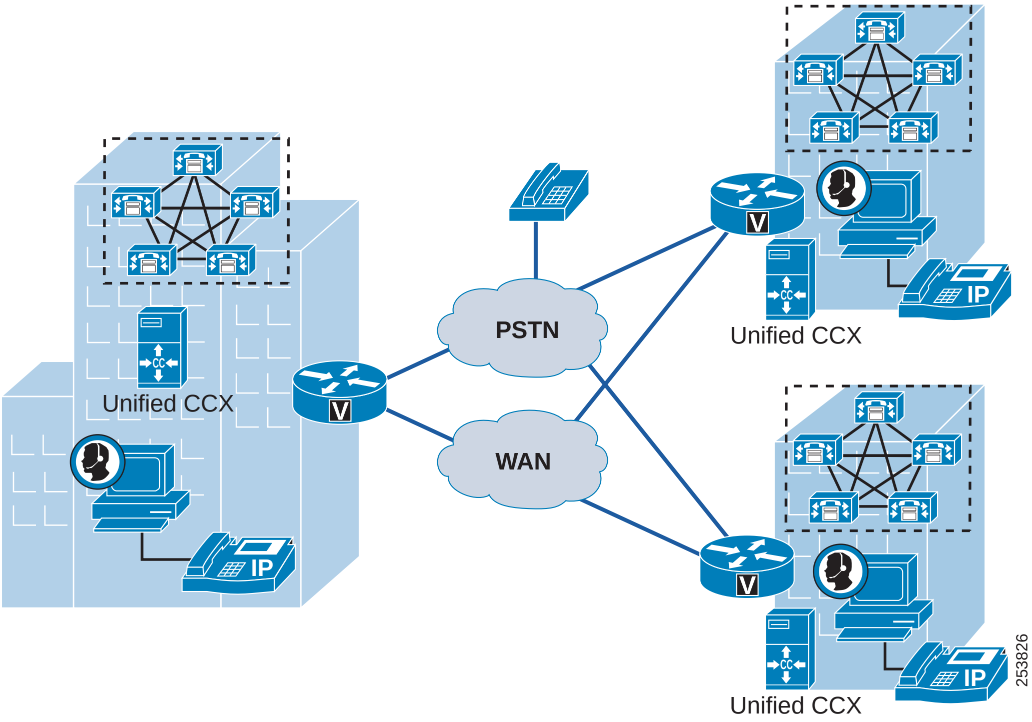
Remote agents (for example, agents working from home) are also supported with Cisco Unified Contact Center. There are mainly two solutions. The first one requires the agent to use an IP phone that is connected to the central site by a broadband internet connection. In this solution, the phone is CTI controlled by the Cisco Unified Contact Center application. The second solution is based on Cisco Unified Mobile Agent, which enables an agent to participate in a call center with any PSTN phone such as cell phone.

Multisite Contact Center with Distributed Call Processing

The model for a multisite deployment with distributed call processing consists of multiple sites, each with its own call processing cluster connected to an IP WAN. This section assumes that each Unified CM cluster has agents registered to it.

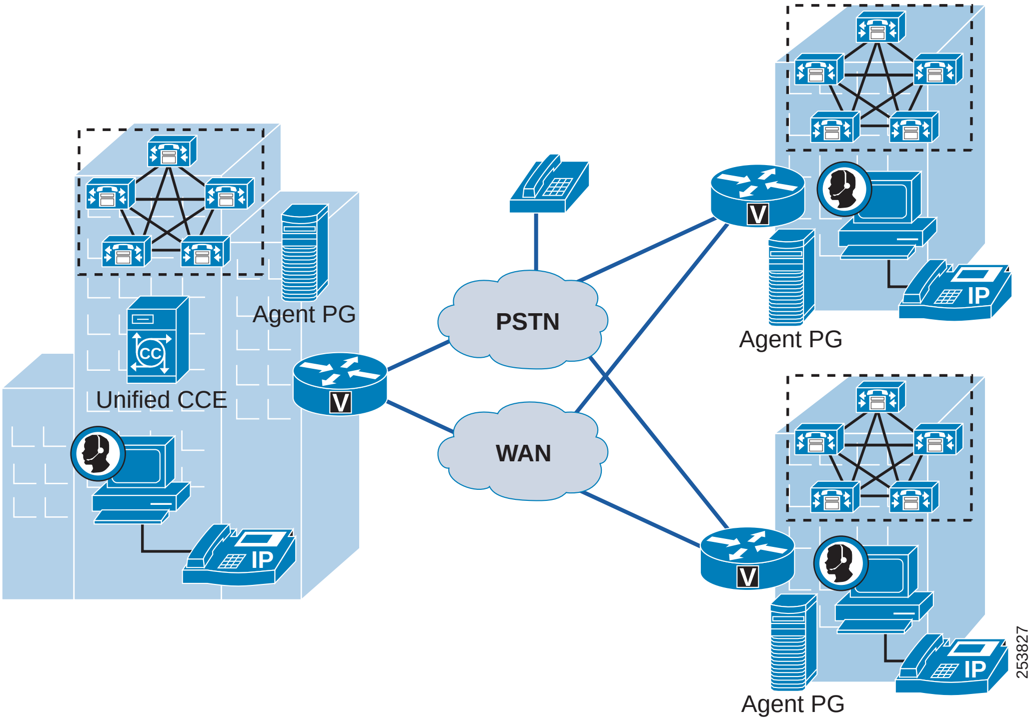
A Unified CCX deployment cannot be shared across multiple Unified CM clusters. Each Unified CM cluster requires its own Unified CCX deployment, as illustrated in Figure 26-2.

Figure 26-2 Multisite Unified CCX Deployment with Distributed Call Processing



Requirements for Unified CCE differ from Unified CCX. A single Unified CCE system can span across multiple Unified CM clusters distributed across multiple geographic locations. A Unified CCE Agent PGs must be installed in each Unified CM cluster location and could be physically remote from the Unified CCE Central Controller (Call Router + Logger). Figure 26-3 illustrates this type of deployment and highlights the placement of the Agent PG.

Figure 26-3 Multisite Unified CCE Deployment with Distributed Call Processing



If you require multiple contact center deployments, you could connect those deployments through Unified ICM by using the parent/child deployment model to form a single virtual contact center. The parent/child model provides several benefits, such as enterprise queuing and enterprise reporting across all the contact center deployments. It also provides complete site redundancy and higher scalability. For more details on the parent/child model, refer to the following documents:

- *Cisco Unified Contact Center Enterprise SRND*, available at

  http://www.cisco.com/go/ucsrnd
- *Cisco Contact Center Gateway Deployment Guide for Cisco Unified ICME/CCE/CCX*, available at

  http://www.cisco.com/en/US/products/sw/custcosw/ps1001/prod_installation_guides_list.html

Similarly to the multisite model with centralized call processing, multisite deployments with distributed call processing require careful configuration of QoS, call admission control, codecs, and so forth.

Clustering Over the IP WAN

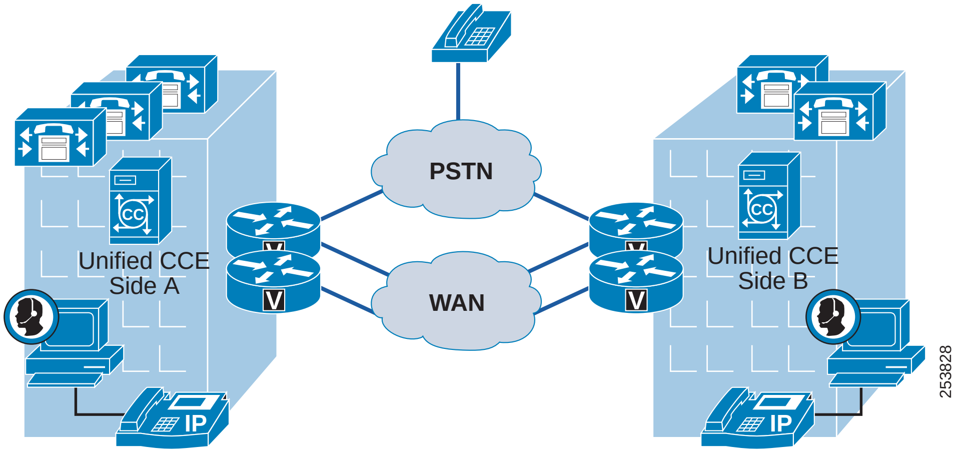
In this deployment model, a single Unified CM cluster is deployed across multiple sites that are connected by an IP WAN with QoS features enabled. Cisco Unified Contact Center solutions can be deployed with this model. In fact, the Cisco Unified Contact Center components themselves can also be clustered over the WAN.

For example, with Unified CCE, the side A servers could be remote from the Unified CCE side B servers and separated from them by an IP WAN connection. (For more details on Unified CCE high availability, see High Availability for Contact Centers, page 26-11.) The following design considerations apply to this type of deployment:

- The IP WAN between the two sites must be highly available, with no single point of failure. For example, the IP WAN links, routers, and switches must be redundant. WAN link redundancy could be achieved with multiple WAN links or with a SONET ring, which is highly resilient and has built-in redundancy For more details, refer to the Unified CCE SRND, available at http://www.cisco.com/go/ucsrnd.
- The Agent Peripheral Gateway (PG) must be co-located with the CTI Manager server to which it is connected. Because of the large amount of redirect and transfer traffic and additional CTI traffic, the Intra-Cluster Communication Signaling (ICCS) bandwidth requirements between the Unified CM servers are higher when deploying Unified CCE. For more details, refer to the Unified CCE SRND, available at http://www.cisco.com/go/ucsrnd.
- If the primary Unified CCE and Unified CM servers are located in one site and the secondary Unified CCE and Unified CM servers are in another site, the maximum latency between the two sites is dictated by the Unified CM latency requirement of 80 ms round trip time (RTT). However, if the Unified CCE servers are in different locations than the Unified CM servers, it is possible to have a higher latency between the redundant Unified CCE servers. For more information, refer to the Unified CCE SRND, available at http://www.cisco.com/go/ucsrnd.

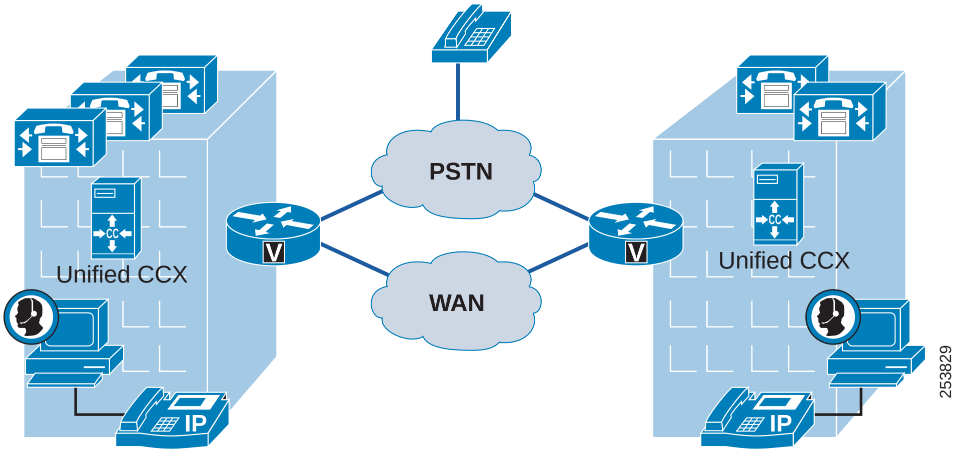
Figure 26-4 illustrates a deployment of Unified CCE using clustering over the WAN.

Figure 26-4 Unified CCE Deployment with Clustering Over the WAN



With Unified CCX and Unified IP IVR solutions, the primary Unified CCX or Unified IP IVR server could also be remote from the backup server. The requirements for Unified CCX deployments are different than the ones for Unified CCE deployments. For example, redundant WAN links are not required with Unified CCX. Also, the maximum latency between the primary and backup Unified CCX servers is 80 ms RTT. Figure 26-5 illustrates this type of deployment. For more details, refer to the Unified CCX SRND, available at http://www.cisco.com/go/ucsrnd.

Figure 26-5 Unified CCX Deployment with Clustering Over the WAN



Design Considerations for Contact Center Deployments

This section summarizes the following major design considerations for contact center deployments:

- High Availability for Contact Centers, page 26-11
- Bandwidth, Latency, and QoS Considerations, page 26-12
- Call Admission Control, page 26-13
- Integration with Unified CM, page 26-13
- Other Design Considerations for Contact Centers, page 26-14

High Availability for Contact Centers

All Cisco Unified Contact Center products provide high availability. For example, when you integrate Unified CCX or Unified IP IVR with Unified CM, you could add a second Unified CCX or Unified IP IVR server to provide high availability. One of the servers would be the active server and would handle all the call processing. The other server would be in standby mode and become active only if the primary server fails. Unified CVP also supports high available deployments with multiple Unified CVP servers, voice gateways, VXML gateways, SIP proxies, and so forth.

With Unified CCE, most of the servers are required to be redundant, and the redundant instances are referred to as side A and side B instances. For example, Call Router A and Call Router B are redundant instances of the Call Router module (process) running on two different servers. This redundant configuration is also referred to as *duplex mode*. The Call Routers run in synchronized execution across the two servers, which means both sides of the duplex servers process every call. Other components, such as the Peripheral Gateways, run in hot-standby mode, meaning that only one of the Peripheral Gateways is actually active at any given time.

In addition to the redundancy of the Unified Contact Center components themselves, their integration with Unified CM can also be redundant. For example, each Unified CCX or Unified IP IVR server can connect to a primary CTI Manager and also to a backup CTI Manager in case the primary CTI Manager fails. With Unified CCE, a PG side A would connect to a primary CTI Manager, while the redundant PG side B connects to the secondary CTI Manager, thus providing high availability if one CTI Manager fails.

For more details, refer to the Cisco Unified Contact Center SRNDs, available at http://www.cisco.com/go/ucsrnd.

Bandwidth, Latency, and QoS Considerations

This section describes how to provision WAN bandwidth in a multisite contact center deployment, taking into account different types of call control traffic and real-time voice traffic. It is important to understand the latency and QoS parameters because adequate bandwidth provisioning and implementation of QoS are critical components in the success of contact center deployments.

Bandwidth Provisioning

Contact center solutions require sufficient WAN bandwidth to accommodate the following main types of traffic:

- Voice traffic between the ingress gateway and the IVR system. With Unified IP IVR, if the Unified IP IVR servers are in a central location and PSTN gateways are in remote locations, there will be voice traffic over the WAN. With Unified CVP, it is possible to queue the call at the edge and therefore keep the voice traffic local to the remote site to avoid voice traffic across a WAN link.
- Voice traffic between the ingress gateway and the agent.
- Voice signaling traffic. This is typically for the signaling traffic between the ingress gateway and Unified CM, and between the agent phone and Unified CM.
- VXML Gateway traffic if Unified CVP is deployed. The traffic includes media file retrieval from the media server and VXML documents exchanged with the VXML server.
- Data traffic between the agent or supervisor desktop and the Unified Contact Center server (CAD or CTI-OS traffic).
- Reporting traffic between the reporting user and the Unified Contact Center Reporting server.
- Traffic between Unified Contact Center servers if they are remote from each other. For example, this type of traffic occurs with clustering over the IP WAN or with multisite and distributed call processing with PGs remote from the Unified CCE Central Controller.
- Additional Intra-Cluster Communication Signaling (ICCS) traffic between the Unified CM subscribers due to the large amount of redirect and transfer traffic and additional CTI traffic.
- Voice traffic due to recording and silent monitoring. Depending on the solution, one or two RTP streams could be sent in order to silently monitor or record the conversation with an agent.

Bandwidth calculations and guidelines are provided in the Cisco Unified Contact Center SRNDs, available at http://www.cisco.com/go/ucsrnd.

Latency

Agents and supervisors can be located remotely from the call processing server and contact center server. Technically, the delay between the CTI OS server and CTI OS client, as well as between the CAD server and CAD or CSD desktop, could be very high because of high time-out values. Long latency will affect the user experience and might cause confusion or become unacceptable from the user perspective. For example, the phone could start ringing but the desktop might not be updated until later.

Latency requirements between the contact center components and the call processing servers, and between the contact center components themselves, depend on the contact center solutions. For example, the Unified CCX redundant servers can be located remotely from each other, with a maximum latency of 80 ms RTT. With Unified CCE, the maximum latency between the Unified CCE servers and Unified CM, or between the Unified CCE servers themselves, is higher than 80 ms RTT.

For more details, refer to the Cisco Unified Contact Center SRNDs, available at http://www.cisco.com/go/ucsrnd.

QoS

Similar to deployments with other Unified Communications components, contact center deployments require the configuration of Quality of Service (QoS) to prioritize time-sensitive or critical traffic. QoS marking for voice and voice signaling in a contact center environment is the same as with other Unified Communications deployments. Traffic specific to the contact center must be marked with specific QoS markings. For example, some of the traffic for the Unified CCE private network must be marked as AF31, while other traffic must be marked as AF11. The QoS marking recommendations and QoS design guidance are documented for each Unified Contact Center solution in their respective Cisco Unified Contact Center SRNDs, available at http://www.cisco.com/go/ucsrnd.

Call Admission Control

Similar to deployments with other Unified Communications components, contact center deployments require careful provisioning of call admission control. The same mechanisms described in the chapter on Call Admission Control, page 11-1, also apply to contact center environments.

Voice traffic associated with silent monitoring and recording might not be accounted for in the call admission control calculation. For example, voice traffic from silent monitoring and recording by Unified CM (voice traffic forked at the phone) is properly accounted for, but voice traffic from desktop-based silent monitoring (desktop connected to the back of the agent IP phone) is not counted in call admission control calculations.

Call admission control for Mobile Agent and Unified CVP involves special considerations. For more details, refer to the Cisco Unified Contact Center SRNDs, available at http://www.cisco.com/go/ucsrnd.

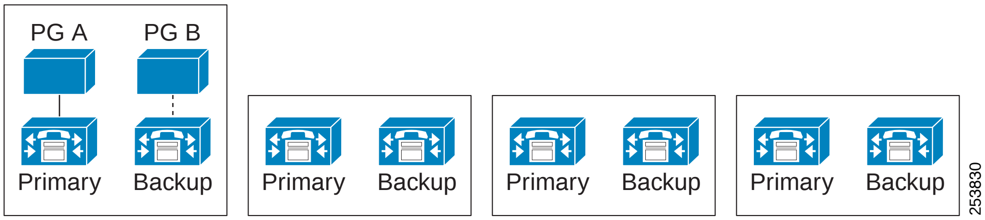
Integration with Unified CM

Observe the following design considerations apply when integrating Cisco Unified Contact Center components with Unified CM:

- For administration and upgrade purposes, Cisco recommends separate Unified CM clusters for contact center and non-contact center deployments. If separate clusters are not possible, then Cisco recommends separate Unified CM subscriber servers for contact center and non-contact center applications. For more details, refer to the Unified CCE SRND, available at http://www.cisco.com/go/ucsrnd.
- With contact center deployments, Cisco recommends that you do not use a 2:1 redundancy scheme for the Unified CM servers. Use 1:1 redundancy to provide higher resiliency and faster upgrades. For more details, refer to the Unified CCE SRND, available at http://www.cisco.com/go/ucsrnd.

- The integration between Unified CM and Unified CCX, Unified IP IVR, or Unified CCE is done through JTAPI. The Unified CCX server connects to a primary CTI Manager. It also has a backup connection to a secondary CTI Manager. With Unified CCE, the Agent PG connects to only one CTI Manager. The redundant Agent PG connects to the backup CTI Manager only. If the primary CTI Manager fails, the primary Agent PG will also fail and trigger the failover.
- There are several ways to deploy CTI Manager with the Unified CCE PG. For example, in a Unified CCE deployment that requires four Unified CM subscriber pairs, four Agent PGs could be deployed and each Agent PG could be connected to a separate Unified CM subscriber pair that is also running the CTI Manager Service. Alternatively, a single PG could connect to only one of the Unified CM subscriber pairs that is running the CTI Manager Service, and through this Unified CM pair, the PG would be able to control/monitor agent phones on all four Unified CM subscriber pairs. This configuration is common in centralized deployments and is illustrated in Figure 26-6. For more details, refer to the Unified CCE SRND, available at http://www.cisco.com/go/ucsrnd.
- It is possible to integrate multiple Unified CCX deployments with a single Unified CM cluster. For more details, refer to the Unified CCX SRND, available at http://www.cisco.com/go/ucsrnd.

Figure 26-6 Deployment with One Agent PG and Four Unified CM Subscriber Pairs



Other Design Considerations for Contact Centers

The following additional design considerations apply in the situations indicated:

- Because Unified CVP allows queuing at the edge, deploying Unified CVP instead of Unified IP IVR could lower the bandwidth requirements for multisite deployments.
- Most of the Cisco Unified Contact Center products and components can be installed in a virtualized environment based on VMware. For details, consult the respective Cisco Unified Contact Center SRNDs, available at http://www.cisco.com/go/ucsrnd.
- Media termination point (MTP) resources might be required in some scenarios. For example, with Mobile Agents, MTPs are required for the associated CTI ports when RFC 2833 is negotiated. MTPs are also required in some scenarios with Unified CVP. For details, consult the respective Cisco Unified Contact Center SRNDs, available at http://www.cisco.com/go/ucsrnd.
- Some third-party contact center products are also supported with Unified CM. The integration with Unified CM could be based on JTAPI and could use CTI ports for call treatment and queuing and CTI route points. To size Unified CM correctly, it is important to have a good understanding of the call flows and their impact on Unified CM. It is also important to understand how the redundancy is implemented and whether or not it impacts Unified CM or CTI scalability.

Capacity Planning for Contact Centers

All deployments must be sized with the Cisco Unified Communications Sizing Tool (Unified CST). This tool performs sizing of the contact center products such as Unified CCE, Unified IP IVR, Unified CVP, and Unified CCX. It determines the contact center resources required for your deployment, such as number of agents, number of IVR ports, and number of gateway ports. In addition to performing sizing for the contact center components themselves, the tool also sizes the rest of the Unified Communications solution, including Unified CM and voice gateways. This tool is available to Cisco employees and partners only (with proper login authentication) at http://tools.cisco.com/cucst.

In general, sizing of the contact center depends heavily on the busy hour call attempts (BHCA) for calls coming into the contact center. It also depends on other parameters such as the Service Level Goal and Target Answer Time. For example, a deployment where 90% of the calls must be answered within 30 seconds will require more contact center resources than a deployment where 80% of the calls must be answered within 2 minutes. Another parameter that impacts the sizing is whether CAD or CTI OS is used, which could result in different Agent PG scalability. Use the Unified CST for sizing, and consult the respective Cisco Unified Contact Center SRNDs, available at http://www.cisco.com/go/ucsrnd, for more details.

The contact center design also impacts Unified CM sizing. The following considerations apply to sizing Unified CM when it is deployed in contact center solutions:

- The maximum number of Unified CCE agents in a single Unified CM cluster depends on the IVR solution. With Unified IP IVR, CTI route points and CTI ports are used during the call treatment queuing, which consume Unified CM resources. With Unified CVP, the call treatment and queuing are typically handled by the VXML Gateway, Unified CVP VXML server, and Unified CVP call server, with no impact on Unified CM. Therefore, a single Unified CM cluster can support more agents with Unified CVP than with Unified IP IVR.
- The Unified CCE Mobile Agent feature relies on CTI ports and therefore needs additional resources from Unified CM subscribers. Therefore, Unified CM scalability is reduced when Mobile Agents are deployed.
- With Unified CCE deployments, two types of outbound dialers are available. With the SCCP dialer, the dialer ports are registered to Unified CM, and each outbound call involves Unified CM even if the outbound call does not reach a live customer. With the SIP dialer, each outbound call is placed directly from the SIP dialer port to the egress voice gateway. With the SIP dialer, the call reaches Unified CM only when the call is transferred to an agent. Therefore, Unified CM capacity is much higher when the SIP dialer is used.
- When sizing Unified CM, it is also important to account for any additional CTI applications. For example, some PC clients can control a phone remotely through CTI. Some call recording applications can also integrate directly with Unified CM through the CTI Manager and can monitor agent phones, which could require additional resources from Unified CM. For more details, refer to Computer Telephony Integration (CTI), page 8-30, and to the Cisco Unified Contact Center SRNDs available at http://www.cisco.com/go/ucsrnd.
- Some silent monitoring and recording solutions (such as the silent monitoring and recording feature based on Unified CM) consume resources from Unified CM, whereas other solutions such as SPAN or desktop silent monitoring and recording do not.
- Again, due to the complexity associated with sizing, all deployments must be sized with the Cisco Unified Communications Sizing Tool, available to Cisco employees and partners only (with proper login authentication) at http://tools.cisco.com/cucst

For more details, refer to the Cisco Unified Contact Center SRNDs, available at http://www.cisco.com/go/ucsrnd.

Network Management Tools

Unified CCE is managed with the Simple Network Management Protocol (SNMP). Unified CCE devices have a built-in SNMP agent infrastructure that supports SNMP v1, v2c, and v3, and it exposes instrumentation defined by the CISCO-CONTACT-CENTER-APPS-MIB. This MIB provides configuration, discovery, and health instrumentation that can be monitored by standard SNMP management stations. Moreover, Unified CCE provides a rich set of SNMP notifications that alert administrators of any faults in the system. Unified CCE also provides a standard syslog event feed (conforming to RFC 3164) for those administrators who want to take advantage of a more verbose set of events.

For more information about configuring the Unified CCE SNMP agent infrastructure and the syslog feed, refer to the *SNMP Guide for Cisco ICM/IPCC Enterprise & Hosted Editions*, available at

> http://www.cisco.com/en/US/products/sw/custcosw/ps1001/products_installation_and_configuration_guides_list.html

Unified CVP health monitoring can be performed by using any SNMP standard monitoring tool to get a detailed visual and tabular representation of the health of the solution network. All Unified CVP product components and most Unified CVP solution components also issue SNMP traps and statistics that can be delivered to any standard SNMP management station or monitoring tool.

Unified CCX can also be managed with SNMP and a syslog interface.

PART 5

Unified Communications Operations and Serviceability

CHAPTER 27

Overview of Cisco Unified Communications Operations and Serviceability

Revised: February 29, 2012; OL-27282-05

Once the network, call routing, call control infrastructure, and applications and services have been put in place for your Cisco Unified Communications System, network and application management components can be added or layered on top of that infrastructure. There are numerous operations and serviceability applications and services that can be deployed in an existing Cisco Unified Communications infrastructure. These applications and services can be classified into four basic areas:

- User and Device Provisioning Services — Provide the ability to centrally provision and configure users and devices for unified communications applications and services.
- Voice Quality Monitoring and Alerting — Provide the ability to monitor on an ongoing basis various call flows occurring within the system to determine whether voice quality is acceptable and to alert administrators when the voice quality is not acceptable.
- Operations and Fault Monitoring — Provides the ability to centrally monitor all application and service operations and to issue alerts to administrators regarding network and application failures.
- Network and Application Probing — Provides the ability to probe and collect network and application traffic information at various locations throughout the deployment and to allow administrators to access and retrieve this information from a central location.

This part of the SRND covers the applications and services mentioned above. It provides an introduction to the various network management applications and services, followed by discussions surrounding architecture, high availability, capacity planning, and design considerations. The discussions focus on design-related aspects of the applications and services rather than product-specific support and configuration information, which is covered in related product documentation.

This part of the SRND includes the following chapters:

- Network Management, page 28-1

  This chapter examines unified communications network and application management services, a common and prevalent set of services within most unified communications deployments, which allows administrators to provision and configure users and devices, monitor network and application operations as well as voice quality, and receive alerts and alarms when issues arise. This chapter also examines the impact of these management applications and services on deployment models and provides design and deployment best practices for network and application management services and applications.

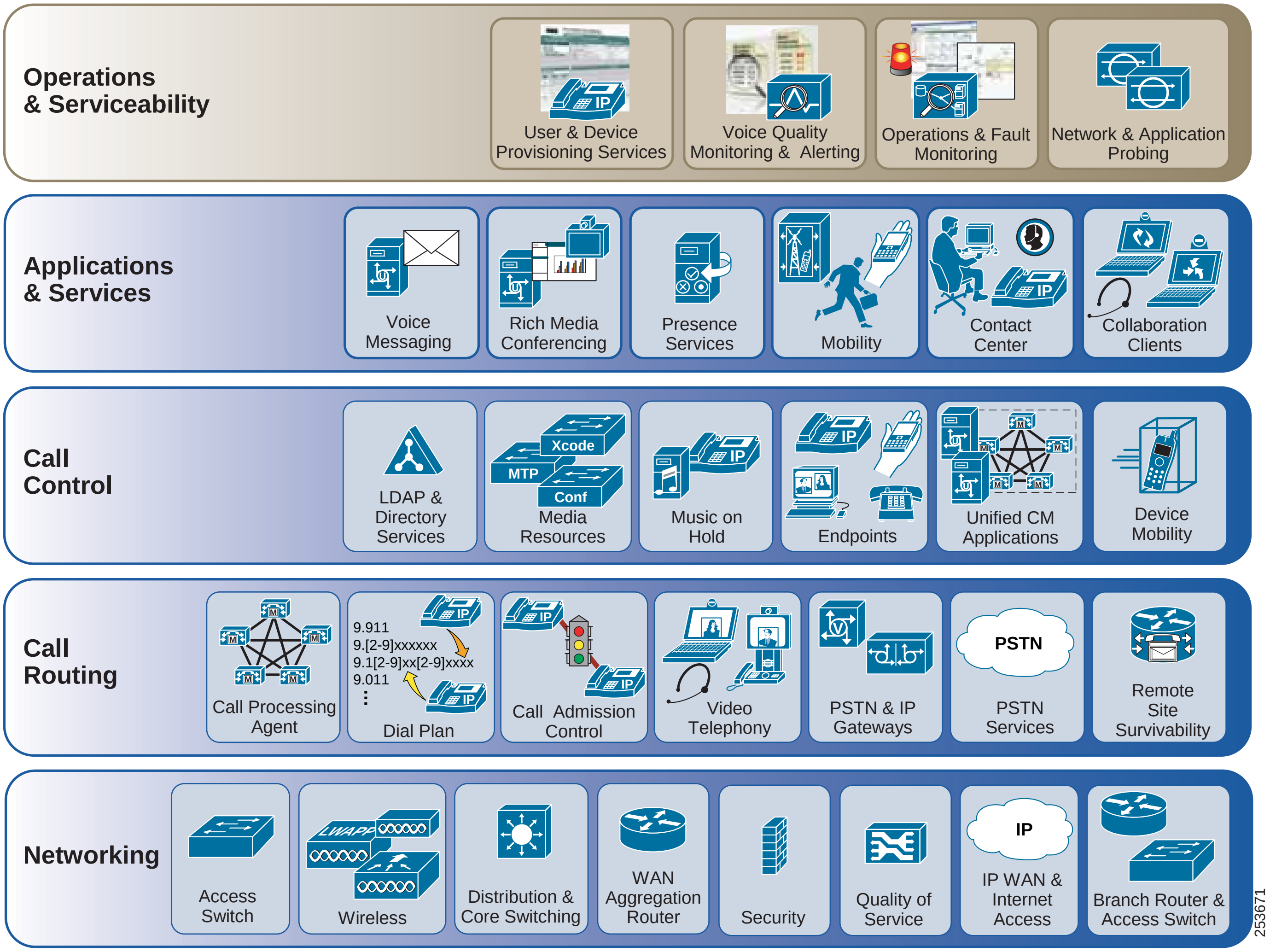
- Unified Communications Design and Deployment Sizing Considerations, page 29-1

  This chapter discusses the sizing of individual Unified Communications components as well as systems consisting of several components communicating with each other. This chapter also discusses the performance impact of the different functions that the various Unified Communications products support, and it explains why "designing by datasheets" is not the preferred way to deploy a complex Unified Communications network. In addition, this chapter provides insights on how to work with the various sizing tools available, mostly notably the Cisco Unified Communications Sizing Tool.

Architecture

As with other network and application technology systems, operations and serviceability applications and services must be layered on top of the underlying network, system, and application infrastructures in order to be able to monitor and control these infrastructures. Figure 27-1 shows the logical location of unified communications operations and serviceability in the overall Cisco Unified Communications System architecture.

Figure 27-1 Cisco Unified Communications Operations and Serviceability Architecture



Unified communications operations and serviceability services such as user and device provisioning, voice quality monitoring and altering, operations and fault monitoring, and network and application probing, all rely on the underlying network infrastructure for network connectivity for various operations and serviceability applications and probes. While there is no direct reliance on the unified communications call routing, call control infrastructure, or unified communications applications and services, these infrastructures and applications are what the various operational and management services actually manage and configure. For example, user and device provisioning services as well as various monitoring and alerting services leverage the network infrastructure for connectivity to various unified communications applications and service nodes in order to configure and monitor various components and operations. These same services also communicate directly with, and in some cases change configurations on or receive alerts from, components such as call processing agents, PSTN and IP gateways, media resources, endpoints, and various unified communications applications for voice messaging, rich media conferencing, and collaboration clients. In addition to relying on these infrastructure layers and basic unified communications services and applications, services pertaining to operations and serviceability are also often dependent upon each other for full functionality.

High Availability

As with network, call routing, and call control infrastructures and critical unified communications applications and services, unified communications operations and serviceability services should be made highly available to ensure that required provisioning, monitoring, and altering will continue even if failures occur in the network or applications. It is important to understand the various types of failures that can occur as well as the design considerations around those failures. In some cases, the failure of a single operations and management application or server can impact multiple services because the unified communications operations and serviceability components are dependent on other components or services. For example, while the various application service components of a network management deployment might be functioning properly, the loss of network connectivity to, or a failure of, a network probe would effectively eliminate the ability to monitor network health or voice quality unless redundant network probes had been deployed along with alternate paths of connectivity.

For operations and serviceability functions such as user and device provisioning, high availability considerations include temporary loss of functionality due to network connectivity or application server failures resulting in the inability of administrators to provision users and devices or to make changes to these user account or device configurations. In addition, failover considerations for these types of operations include scenarios in which portions of the functionality can be handled by a redundant operation or management application that allows administrators to continue to facilitate some configuration changes in the event of certain failures.

High availability considerations are also a concern for operations and serviceability applications that provide services such as voice quality monitoring or application and operations fault monitoring. Interrupted network connectivity or server or application failures will typically result in a reduced ability to monitor and/or alert, and in some cases complete loss of such functionality. For voice quality monitoring, this can mean that voice quality measurements for some call flows or devices will be unavailable. For operations and fault monitoring services, high availability considerations include the potential for loss of operational change tracking data or fault alerts and indications.

Capacity Planning

Network, call routing, and call control infrastructures as well as unified communications applications and services must be designed and deployed with an understanding of the capacity and scalability of the individual components and the overall system. Similarly, deployments of operations and serviceability components and services must also be designed with attention to capacity and scalability considerations. When deploying various operations and serviceability applications and components, not only is it important to consider the scalability of these applications themselves, but you must also consider the scalability of the underlying infrastructures. Certainly the network infrastructure must have available bandwidth and be capable of handling the additional traffic load these operations will create. Likewise, the call routing and control infrastructure must be capable of handling required inputs and outputs as facilitated by the various operations and serviceability components in use. For example, with operational applications and services such as voice quality monitoring and alerting and operations and fault monitoring, there are capacity implications for each of these individual applications or services in terms of the number of devices and call flows that can be monitored at a given time, but just as important is the scalability of the underlying infrastructure and monitored applications to handle the added network traffic and connections required for monitoring and alerting. While the monitoring and alerting application or service itself may be able to support the monitoring of many network devices and call flows, the underlying network or devices might not have available capacity to handle the probing connections or the alarm messaging load generated by these monitoring and alerting services.

For operation applications or services that provide user or device provisioning capabilities, capacity planning considerations include things such as ensuring that the provisioning application can handle the requested load and also that user or device provisioning operations not only do not exceed the number of support devices or users for a particular underlying unified communications application or service, but also that provisioning or configuration change transactions do not exceed either the capacity of the underlying network or the rate at which a particular application can handle transactions. In most cases additional capacity can be added by increasing the number of operational provisioning application servers or by increasing the size or number of underlying unified communications applications or service instances, assuming the underlying network and call routing and control infrastructures are capable of handling this additional load.

For a complete discussion of system sizing, capacity planning, and deployment considerations related to sizing, refer to the chapter on Unified Communications Design and Deployment Sizing Considerations, page 29-1.

CHAPTER 28

Network Management

Revised: September 28, 2012; OL-27282-05

Network management is a service consisting of a wide variety of tools, applications, and products to assist network system administrators in provisioning, operating, monitoring and maintaining new and existing network deployments. A network administrator faces many challenges when deploying and configuring network devices and when operating, monitoring, and reporting on the health of the network infrastructure and components such as routers, servers, switches and so forth. Network management helps system administrators monitor each network device and network activity so that they can isolate and investigate problems in a timely manner for better performance and productivity.

With the convergence of rich media and data, the need for unified management is greater than ever. The Cisco Prime Collaboration (Prime Collaboration) offers a set of integrated tools that help to test, deploy, and monitor Cisco Unified Communications and TelePresence systems. Prime Collaboration implements the various management phases to strategically manage the performance and availability of Cisco Unified Communications applications including voice, video, contact center, and rich media applications. The network management phases typically include: plan, design, implement, and operate (PDIO). Table 28-1 lists the PDIO phases and the major tasks involved with each phase.

Table 28-1 Network Management Phases and Tasks

Plan & Design	Implement	Operate
Assess the network infrastructure for Cisco Unified Communications capability. For example, predict overall call quality. Prepare the network to support Cisco Unified Communications. Analyze network management best practices.	Deploy and provision Cisco Unified Communications. For example, configure the dial plan, partitioning, user features, and so forth. Enable features and functionality on the existing infrastructure to support Cisco Unified Communications. For example, configure voice ports, gateway functionality on routers, and so forth.	Manage changes for users, services, IP phones, and so forth. Generate reports for operations, capacity planning, executive summaries, and so forth. Track and report on user experiences. For example, use sensors to monitor voice quality. Monitor and diagnose problems such as network failures, device failures, call routing issues, and so forth.

This chapter provides the design guidance for the following management tools and products that fit into the implementation and operation phases of Cisco Unified Communications Management:

- Cisco Prime Collaboration manages provisioning of initial deployments and ongoing operational activation for Unified Communications and TelePresence services. Cisco Prime Collaboration provides comprehensive monitoring with proactive and reactive diagnostics for the entire Cisco Unified Communications system. It also provides a reliable method of monitoring and evaluating voice quality in Cisco Unified Communications systems. For details, refer to the related product documentation available at

  http://www.cisco.com/en/US/products/ps11480/index.html
- Cisco Unified Service Statistics Manager (Unified SSM) provides advanced statistics analysis and reporting capabilities for Cisco Unified Communications deployments. For details, refer to the related product documentation available at

  http://www.cisco.com/en/US/products/ps7285/index.html

For information on which software versions are supported with Cisco Unified Communications Manager (Unified CM), refer to the *Cisco Unified Communications Manager Software Compatibility Matrix*, available at

http://www.cisco.com/en/US/docs/voice_ip_comm/cucm/compat/ccmcompmatr.html

What's New in This Chapter

Table 28-2 lists the topics that are new in this chapter or that have changed significantly from previous releases of this document.

Table 28-2 New or Changed Information Since the Previous Release of This Document

New or Revised Topic	Described in:	Revision Date
Cisco Prime Collaboration	Cisco Prime Collaboration, page 28-2, and other sections throughout this chapter	September 28, 2012
No changes for Cisco Unified Communications System Release 9.0		June 28, 2012

Cisco Prime Collaboration

Cisco Prime Collaboration (9.0 and later releases) integrates the following three products from the Cisco Unified Communications Management Suite:

- Cisco Prime Unified Operations Manager
- Cisco Prime Unified Provisioning Manager
- Cisco Prime Unified Service Monitor

The Cisco Unified Service Statistics Manager remains a separate product. Prime Collaboration was primarily developed for TelePresence applications, but it now also covers Unified Communications. Cisco Prime Collaboration supports the following installation combinations:

- Assurance only (Unified Communications and TelePresence)
- Assurance only (Unified Communications)

- Assurance only (TelePresence)
- Any combination of above assurance together with Provisioning (Unified Communications only)
- Provisioning only (Unified Communications)

Cisco Prime Collaboration provides comprehensive voice and video network monitoring with diagnostics for the Cisco Collaboration systems, including the underlying transport infrastructure. Prime Collaboration is a converged application that eliminates the need to manage the video deployments separately from voice. It is delivered as two separate applications, Assurance and Provisioning, that are installed on separate virtual machines. This converged application combines the benefits of Assurance and Provisioning.

The Assurance application provides:

- End-to-end visualization of video collaboration sessions
- End-to-end service monitoring for Cisco Collaboration applications.
- Real-time service troubleshooting and diagnostics for Cisco TelePresence systems and endpoints.
- Video service readiness assessment with Cisco medianet.
- Diagnostics tests using Cisco IP Service Level Agreements (IP SLA) and Video SLA Assessment Agent (VSAA).
- Service-level and inventory reports for voice and video systems.

The Provisioning application provides:

- Standard services (for example, phone, line, and voicemail) to be ordered for subscribers (the owner of the individual phone, voicemail, or other service).
- Configuration templates provide the ability to auto-configure the Cisco Unified Communications voice infrastructure in a consistent way.
- Easy addition of the Provisioning application to an existing Cisco Unified Communications network.
- Simplified policy-driven Day 2 provisioning interface to manage subscribes and users.
- A selfcare feature that enables end users to set up lines, manage services, and configure phone options quickly and easily.
- Batch provisioning for a large number of subscribers

You can run these applications either as:

- A converged application with single sign-on. This mode provides a converged user interface with launch points for both Assurance and Provisioning features.
- Standalone applications with separate login. This mode provides a separate user interface for Assurance and Provisioning features.

For information on the benefits and key features of Prime Collaboration, refer to the Cisco Prime Collaboration documentation available at

http://www.cisco.com/en/US/products/ps11480/index.html

Failover and Redundancy

Prime Collaboration does not currently support failover. However, it can support Network Fault Tolerance when deployed on server platforms with dual Ethernet network interface cards (NICs) that support NIC teaming. This feature allows a server to be connected to the Ethernet through two NICs and, therefore, two cables. NIC teaming prevents network downtime by transferring the workload from the failed port to the working port. NIC teaming cannot be used for load balancing or for increasing the interface speed.

Cisco Prime Collaboration Server Performance

Prime Collaboration runs only in a virtual environment and it requires a minimum of two virtual machines (one for Assurance and at least one for Provisioning). For specific system requirements and capacity information, refer to the *Cisco Prime Collaboration Quick Start Guide*, available at

http://www.cisco.com/en/US/products/ps11480/index.html

Network Infrastructure Requirements for Cisco Unified Network Management

You should enable Domain Name Service (DNS) in the network to perform a reverse lookup on the IP address of the device to get the hostname for the device. If DNS is not desired, then host files may be used for IP address-to-hostname resolution.

Network Time Protocol (NTP) must be implemented to allow network devices to synchronize their clocks to a network time server or network-capable clock. NTP is a critical network service for network operation and management because it ensures accurate time-stamps within all logs, traps, polling, and reports on devices throughout the network.

You should enable Cisco Discovery Protocol (CDP) within the network to ensure proper monitoring. Prime Collaboration's automated device discovery is based on a CDP table. Ping Sweep may be used instead of CDP, but IP phones discovered using Ping Sweep are reported in "unmanaged" state. Simple Network Management Protocol (SNMP) must also be enabled on network devices to allow Prime Collaboration to get information on network devices at configured polling intervals and to receive alerts and faults via trap notification sent by the managed devices.

Trivial File Transfer Protocol (TFTP) must be enabled in the network for deployments with Cisco 1040 Sensors. TFTP provides the Cisco 1040 Sensor with a TFTP-based process to download its configuration files.

For more information on Cisco Unified Communications network requirements, see the chapter on Network Infrastructure, page 3-1.

Assurance

Cisco Prime Collaboration provides a unified view of the entire Cisco Unified Communications infrastructure and presents the current operational status of each element of the Cisco Unified Communications network. Prime Collaboration also provides diagnostic capabilities for faster problem isolation and resolution. In addition to monitoring Cisco gateways, routers, and switches, Prime Collaboration continuously monitors the operational status of various Cisco Unified Communications elements such as:

- Cisco Unified Communications Manager (Unified CM)
- Cisco Unified Communications Manager Express (Unified CME)
- Cisco Unified Communications Manager Session Management Edition
- Cisco Unity and Unity Connection
- Cisco Unity Express
- Cisco Unified Contact Center Enterprise (Unified CCE), Unified Contact Center Express (Unified CCX), and Unified Customer Voice Portal (Unified CVP)

  Note Cisco Prime Collaboration Service Level View does not support multiple Cisco Unified System Contact Center Enterprise (SCCE) deployments.

- Cisco IM and Presence
- Cisco Emergency Responder
- Cisco Unified MeetingPlace and Unified MeetingPlace Express
- Cisco Unified Border Element
- Cisco Unified Endpoints

Note Cisco Prime Collaboration supports Unified Communications and TelePresence applications running in a virtualized environment but does not provide monitoring of VMware or hardware. Use vCenter for managing VMware hosts. For Unified Computing System (UCS) B-series Blade servers, UCS Manager provides unified, embedded management of all software and hardware components in the Cisco UCS. It controls multiple chassis and manages resources for thousands of virtual machines. For UCS C-series servers, the Cisco Integrated Management Controller provides the management service.

For more information on the supported products (particularly Cisco endpoints) and versions supported by Prime Collaboration, refer to the Cisco Prime Collaboration data sheet available at

http://www.cisco.com/en/US/products/ps11480/index.html

One protocol that Prime Collaboration uses to monitor the Unified Communications elements is Simple Network Management Protocol (SNMP). SNMP is an application-layer protocol using UDP as the transport layer protocol. There are three key elements in SNMP managed network:

- Managed devices — Network devices that have an SNMP agent (for example, Unified CM, routers, switches, and so forth).
- Agent — A network management software module that resides in a managed device. This agent translates the local management information on the device into SNMP messages.
- Manager — Software running on a management station that contacts different agents in the network to get the management information (for example, Prime Collaboration).

The SNMP implementation supports three versions: SNMP v1, SNMP v2c, and SNMP v3. SNMP v3 supports authentication, encryption, and message integrity. SNMP v3 may be used if security is desired for management traffic. Prime Collaboration supports all three versions of SNNP. SNMP v1 and v2c read/write community strings or SNMP v3 credentials must be configured on each device for agent and manager to communicate properly. Prime Collaboration needs only SNMP read access to collect network device information.

For more information on SNMP, refer to the Cisco Prime Collaboration documentation available at

http://www.cisco.com/en/US/products/ps11480/index.html

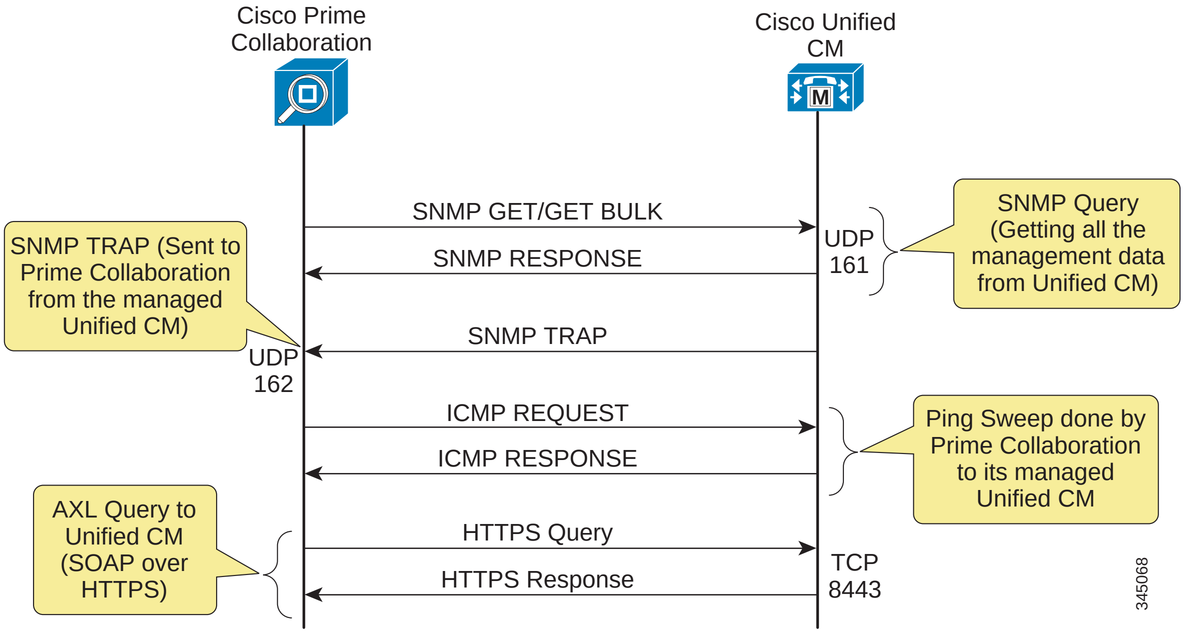
Assurance Design Considerations

Cisco Prime Collaboration interfaces with other devices in the network in the following ways:

- Simple Network Management protocol (SNMP) to manage all Cisco Unified Communications servers, gateways, and switches.
- Administrative XML Layer (AXL) to manage Unified CM. AXL is implemented as a Simple Object Access Protocol (SOAP) over HTTPS web service.
- HTTP to the IP phone to collect serial number and switch information. HTTP must be enabled on the IP phones.
- Enhanced event processing with Cisco Unified CM remote syslog integration, and leveraging the Cisco Real-Time Monitoring Tool (RTMT) interface for pre-collected Unified CM cluster-wide data
- Skinny Client Control Protocol (SCCP) and Session Initiation Protocol (SIP) to Cisco Unified IP Phones for synthetic tests.
- Internet Control Message Protocol (ICMP) or Ping Sweep for Cisco IOS routers and switches, and for other voice as well as non-voice devices.
- Windows Management Instrumentation (WMI) for Cisco Unity servers.

Figure 28-1 shows the system-level overview of how Prime Collaboration leverages multiple interfaces with Unified CM to gather performance counters and alarms.

Figure 28-1 Prime Collaboration and Unified CM System-Level Integration


Cisco Prime Collaboration
Cisco Unified CM
SNMP GET/GET BULK
SNMP RESPONSE
UDP 161
SNMP Query (Getting all the management data from Unified CM)
SNMP TRAP (Sent to Prime Collaboration from the managed Unified CM)
SNMP TRAP
UDP 162
ICMP REQUEST
ICMP RESPONSE
Ping Sweep done by Prime Collaboration to its managed Unified CM
AXL Query to Unified CM (SOAP over HTTPS)
HTTPS Query
HTTPS Response
TCP 8443
345068

Voice Quality Monitoring

Cisco Prime Collaboration monitors voice quality of calls on the Cisco Unified Communications network. It relies on Unified CM, Cisco 1040 Sensors, and Network Analysis Modules (NAMs) to monitor and gather voice quality statistics on real calls rather than simulated calls in the network. Then it compares the collected voice quality statistics against a predefined Mean Opinion Score (MOS) threshold. If the voice quality falls below the threshold, Prime Collaboration is also responsible for sending voice quality information to Cisco Unified Service Statistics Manager (Unified SSM) so that Unified SSM can perform call data analysis and generate reports.

Note A set of global call quality thresholds can be defined as one per supported codec type. Different thresholds can be grouped together based on the Cisco 1040 Sensor being implemented or the Unified CM cluster being monitored.

Voice Quality Measurement

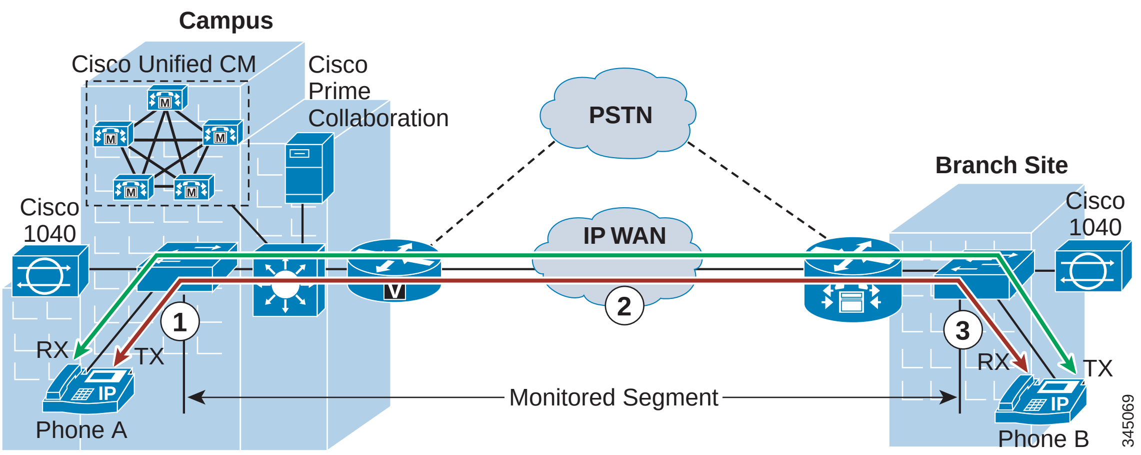
Voice quality is the qualitative and quantitative measure of the sound and conversational quality of the IP phone call. Voice quality measurement describes and evaluates the clarity and intelligibility of voice conversations. Prime Collaboration uses the Cisco 1040 Sensor, the Network Analysis Module (NAM), and Unified CM to monitor and report voice quality information.

Cisco 1040 Sensor Voice Quality Monitoring

The Cisco 1040 Sensor is a hardware device that predicts a subjective quality rating that an average listener might experience on the VoIP calls. It operates by measuring various quality impairment metrics that are included in the IP header of RTP streams, such as packet loss, delay, jitter, and concealment ratio. This computed quality rating is converted to a MOS value. The MOS value is included in syslog messages that are sent to Prime Collaboration every 60 seconds, thus the Cisco 1040 Sensor monitors the voice quality almost on a real-time basis.

The Cisco 1040 Sensor has two Fast Ethernet interfaces, one of which is used to manage the sensor itself and the other is connected to the Switch Port Analyzer (SPAN) port on the Cisco Catalyst switch to monitor the actual RTP streams. To monitor voice quality of calls across the WAN, you must deploy a pair of Cisco 1040 Sensors at both sides of the WAN cloud, as illustrated in Figure 28-2.

Figure 28-2 Voice Quality Monitoring with the Cisco 1040 Sensor



There are two call legs, transmitting and receiving, for each phone. Each call leg can be divided into three segments along the call path. For example, for the transmitting call leg of phone A in Figure 28-2, segment 1 runs between phone A and the campus access switch, segment 2 is between the two access switches, and segment 3 is between the access switch at the branch site and phone B. Segments 1 and 3 are within a local area network, which presents the fewest transmission impairments to voice quality. Therefore, it is reasonably safe to assume that voice quality degradation will not occur in these two segments, and it is unnecessary to monitor those RTP streams.

Segment 2 spans across the WAN circuit and several network devices along the call path. It is more likely to experience degradation of voice quality due to packet loss, delay, and jitter inherent in the WAN. Therefore, the RTP streams (from campus to branch) should be monitored by the Cisco 1040 Sensor at the branch site. By the same token, the sensor in the central site should monitor the incoming RTP streams in that segment across the WAN. These RTP streams provide important voice quality statistics, and their associated MOS values should be analyzed carefully.

Strategic vs. Tactical Monitoring

There are two strategies for deploying Cisco 1040 Sensors: strategic monitoring and tactical monitoring. With strategic monitoring, the Cisco 1040 Sensor is deployed to continuously monitor all or subsets of IP phones in the network. With tactical monitoring, the Cisco 1040 Sensor is deployed in a site where a voice quality issue has been identified. The Cisco 1040 Sensor complies with FCC Class-B standards, and it can be deployed easily in the enterprise environment.

In a small network, Cisco recommends deploying strategic monitoring to monitor all IP phones on a continuous basis. In a medium to large network, Cisco recommends deploying strategic monitoring to continuously monitor a subset of IP phones, while using tactical monitoring to troubleshoot any voice quality issues experienced by the rest of the IP phones.

Design Considerations for the Cisco 1040 Sensor

Consider the following design factors when deploying a Cisco 1040 Sensor:

- A Cisco 1040 Sensor can monitor 100 simultaneous RTP streams. By monitoring the incoming RTP stream only, as illustrated in Figure 28-2, the Cisco 1040 Sensor can provide the full benefit of monitoring 100 (instead of 50) simultaneous voice calls. An environment with a high call volume tends to require the use of more Cisco 1040 Sensors.
- If there are more RTP streams than the Cisco 1040 Sensor can handle, the Cisco 1040 Sensor will randomly select RTP streams.
- The Cisco 1040 Sensor utilizes the SPAN port on the Cisco Catalyst Switch to monitor the actual RTP streams. Different types of Catalyst switches have different quantities of SPAN ports that can be configured. For example, a maximum of two SPAN ports can be configured on the Cisco Catalyst 6500 and 4500 switches, while the maximum limit for Cisco Catalyst 3550 switch is only one. Therefore, the types of Catalyst switches that have been deployed in the network will determine how many Cisco 1040 Sensors can be deployed.
- If there is a trunking connection between multiple Cisco Catalyst switches and if the call volume is low, there is no need to deploy a Cisco 1040 Sensor for every Catalyst switch. Remote Switch Port Analyzer (RSPAN) can be used so that a single Cisco 1040 Sensor can monitor IP phones on other switches within the same VLAN.
- It is inefficient to deploy a Cisco 1040 Sensor at every site that has just a few IP phones and a small call volume. In such cases, Cisco Enhanced Switched Port Analyzer (ESPAN) can be used so that one Cisco 1040 Sensor can monitor voice streams across multiple networks.

Unified CM Voice Quality Monitoring

Unified CM utilizes the Cisco Voice Transmission Quality (CVTQ) algorithm to monitor voice quality. CVTQ is based on the Klirrfaktor (K-factor) method to estimate the MOS value of voice calls. At the end of each call, Unified CM stores the MOS value in Call Management Records (CMRs). The CMRs and Call Detail Records (CDRs) are transferred to Prime Collaboration via Secure File Transfer Protocol (SFTP) every 60 seconds. To integrate with Unified CM, Prime Collaboration must be configured as a Billing Application Server in the Unified CM Unified Serviceability configuration web page. Up to three Billing Application Servers can be configured per Unified CM cluster. The following settings must be configured for the Billing Application Server:

- Hostname or IP address of the Prime Collaboration Assurance virtual machine
- Username and password for SFTP file transfer

- Protocol: SFTP
- Directory path on the Prime Collaboration virtual machine to which CDR and CMR files are transferred

CVTQ is supported natively by Unified CM 7.*x* and Cisco Unified IP Phones running in both SCCP and SIP modes. The phone models that support CVTQ are listed in the compatibility information at

http://www.cisco.com/en/US/products/ps6535/products_device_support_tables_list.html

As a comparison to the Cisco 1040 Sensor, which performs a full-depth inspection on various quality impairment metrics, the K-factor method inspects only one dimension of quality impairments, packet loss, which is really a network effect. Thus, CVTQ is a less sophisticated algorithm than the one that the Cisco 1040 Sensor uses to monitor the quality of calls. Cisco recommends using CVTQ to flag a voice quality issue and using the Cisco 1040 Sensor to validate and troubleshoot the issue.

Cisco Network Analysis Module (NAM)

Cisco NAM is a traffic analysis module that leverages Remote Monitoring (RMON) and some SNMP Management Information Bases (MIBs) to enable network administrators to view all layers of the Unified Communications infrastructure to monitor, analyze, and troubleshoot applications and network services such as QoS for voice and video applications. Voice instrumentation added in Cisco NAM 4.0 enables NAM integration with Prime Collaboration for call metrics through NAM-embedded data collection and performance analysis.

The Cisco NAM complements Prime Collaboration to deliver an enterprise-wide voice management solution. Cisco NAMs are available in different configurations for Cisco Catalyst 6000 Series, 7600 Series, and Integrated Services Routers. The NAM Appliances come with a graphical user interface for troubleshooting and analysis, and they provide a rich feature set for voice quality analysis with RTP and voice control and signaling monitoring. Table 28-3 lists the maximum number of concurrent RTP streams (single direction) that each type of NAM can support.

Table 28-3 Number of Supported Concurrent RTP Streams per NAM Type

Cisco NAM Type	1040 Sensor	NME-NAM	NAM-2	NAM 2204 Appliance	NAM 2220 Appliance
Number of concurrent RTP streams supported	100	100	400	1500	4000

Cisco Prime Collaboration polls the NAM every 60 seconds for voice quality metrics. It then consolidates the data from both the Cisco 1040 Sensor and NAM, and it uses the same method for MOS calculation on both the Cisco 1040 Sensor and NAM. This enables Prime Collaboration to correlate CDR and call stream reports from the Cisco 1040 Sensor and NAM for enhanced analysis.

For more information on Cisco NAM, refer to the following site:

http://www.cisco.com/go/nam

Comparison of Voice Quality Monitoring Methods

Cisco 1040 Sensors, CVTQ, and NAM complement each other and provide a total solution for voice quality measurement. The following list notes key differences between voice quality monitoring with the Cisco 1040 Sensor, CVTQ, and Cisco NAM:

- The Cisco 1040 Sensor monitors voice calls based on packet loss, delay, jitter, and concealment ratio. CVTQ monitors voice calls based on packet loss only.
- The Cisco 1040 Sensor and Cisco NAM provide voice quality statistics every 60 seconds. CVTQ provides voice quality statistics after the call is completed.
- The Cisco 1040 Sensor is compatible with all Cisco Unified CM releases and all types of endpoints connecting to the Cisco Catalyst switch. CVTQ supports only Unified CM 4.2 and later releases.
- For intercluster calls, the Cisco 1040 Sensor monitors the end-to-end call segment. CVTQ monitors only the call segment within its own cluster.
- Cisco recommends using the Cisco 1040 Sensor to monitor key IP phone devices, gateway devices, and application servers in the network and to investigate and troubleshoot voice quality issues. CVTQ-based voice quality monitoring should be used to gauge the overall voice call quality in the network.

Even if CVTQ is not used, Prime Collaboration uses CDR information to correlate with the NAM report for the following metrics:

- Source and/or destination extension number
- Device types
- Interface through which the call flowed in the case of a call to or from a gateway
- Call disconnect reason, where possible
- Exact Unified CM server (not just the Unified CM cluster) to which the phone is connected

Trunk Utilization

Cisco Prime Collaboration provides real-time trunk utilization performance graphs. It is also tightly integrated with Cisco Unified Service Statistics Manager (Unified SSM) in order to provide the call information it collects to Unified SSM for long-term trending and reporting purposes. The call information is provided from the CDR and CMR records Prime Collaboration gathers from Unified CM.

Failover and Redundancy

The Unified CM publisher server is responsible for transferring CDR and CMR files to Prime Collaboration via SFTP. If the publisher server is unavailable, there is no failover mechanism for Prime Collaboration to obtain the new CDR and CMR files that contain MOS values of calls in the Unified CM cluster.

Voice Monitoring Capabilities

Cisco Prime Collaboration supports the following voice quality monitoring capacities:

- Up to 50 Cisco 1040 Sensors
- Any of the following scenarios:
  - 5,000 sensor-based RTP streams per minute (with Cisco 1040 Sensors or NAM modules)
  - 1,600 CVTQ-based calls per minute
  - 1,500 RTP streams and 666 CVTQ calls per minute
- Prime Collaboration automatically selects and gathers voice quality information (via CDR and CMR files) for all Cisco Unified IP Phones configured in a given Unified CM cluster. There is no configuration option to monitor only certain IP phones in the cluster.

Note When Cisco Prime Collaboration is operating at full capacity, its projected database growth (for Syslog, CDR, and CMR files) is estimated to be about 2.4 GB per day.

Assurance Ports and Protocol

Table 28-4 lists the ports used by the various protocol interfaces for Cisco Prime Collaboration for Assurance. Cisco recommends opening these ports in the corporate internal firewalls (if applicable) to allow communication between Prime Collaboration and other devices in the network

Table 28-4 Cisco Prime Collaboration Port Utilization for Assurance

Protocol	Port	Service
UDP	161	SNMP Polling
UDP	162	SNMP Traps
TCP	80	HTTP
TCP	443	HTTPS
TCP	1741	CiscoWorks HTTP server
UDP	22	SFTP
TCP	43459	Database
UDP	5666	Syslog[1]
TCP	2000	SCCP[2]
UDP	69	TFTP[3]
UDP	514	Syslog
TCP	8080	Determining status of Unified CM web service
TCP	8443	SSL port between Unified CM and Prime Collaboration

[1] Prime Collaboration receives Syslog messages from the Cisco 1040 Sensor.

[2] Prime Collaboration communicates with the Cisco 1040 Sensor via SCCP.

[3] The Cisco 1040 Sensor downloads its configuration file via TFTP.

Note The Cisco NAM is accessed remotely over HTTPS with a non-default port. Prime Collaboration will authenticate with each Cisco NAM and maintain the HTTP/S session.

All the management traffic (SNMP) originating from Prime Collaboration or managed devices is marked with a default marking of DSCP 0x00 (PHB 0). The goal of network management systems is to respond to any problem or misbehavior in the network. To ensure proper and reliable monitoring, network management data must be prioritized. Implementing QoS mechanisms ensures low packet delay, low loss, and low jitter. Cisco recommends marking the network management traffic with an IP Precedence of 2, or DSCP 0x16 (PHB CS2), and providing a minimal bandwidth guarantee. The DSCP value must be configured in the Windows Operating System.

If managed devices are behind a firewall, the firewall must be configured to allow management traffic. Prime Collaboration has limited support in a network that uses Network Address Translation (NAT). It must have IP and SNMP connectivity from the Prime Collaboration server to the NAT IP addresses for the devices behind the NAT. Prime Collaboration contains static NAT support.

Bandwidth Requirements

Prime Collaboration polls the managed devices for operational status information at every configured interval, and it has the potential to contain a lot of important management data. Bandwidth must be provisioned for management data, especially if you have many managed devices over a low-speed WAN. The amount of traffic varies for different types of managed devices. For example, more management messages may be seen when monitoring Unified CM as compared to monitoring a Cisco Voice Gateway. Also, the amount of management traffic will vary if the managed devices are in a monitored or partially monitored state and if any synthetic tests are performed. Prime Collaboration has a Bandwidth Estimator that is available at

http://www.cisco.com/web/applicat/ombwcalc/OMBWCalc.html

Assurance Analysis and Reports: Cisco Unified Service Statistics Manager

The Cisco Unified Service Statistics Manager (Unified SSM) performs advanced call statistics analysis and generates reports for executives, operations, and capacity planning functions. Unified SSM is fully dependent on Cisco Prime Collaboration to obtain call statistics information; therefore, Prime Collaboration must be implemented and operating before you deploy Unified SSM. Unified SSM provides both out-of-the-box reports as well as customizable reports that provide visibility into key metrics such as call volume, service availability, call quality, resource and trunk utilization, and capacity across the Cisco Unified Communications system. For the detailed information on feature support and functionality, refer to the Cisco Unified Service Statistics Manager product documents available at http://www.cisco.com.

Integration with Prime Collaboration

Unified SSM integrates with Cisco Prime Collaboration in order to extract call statistics data from its database. The data extraction process is performed by the Unified SSM agent.

The Unified SSM agent facilitates communication between Unified SSM and Prime Collaboration, and it is responsible for transmitting call statistics data from Prime Collaboration to Unified SSM. Unified SSM then stores the extracted data in its own SQL database.

There are two different data collection approaches within Unified SSM. The first approach is called *raw data collection*. With this approach, Unified SSM instructs the Unified SSM agent to retrieve all call statistics data directly from the Prime Collaboration databases. All retrieved data is then saved in Unified SSM's database for up to 30 days. The advantage of this approach is that it provides Unified SSM with a comprehensive data source to perform detailed analysis and report generation.

The second approach is called *monitor-based data collection*. With this approach, Unified SSM instructs the Unified SSM agent to transfer the processed call statistics data only. The advantage of this approach is fewer traffic loads over the network, and the processed data can be stored in the Unified SSM database for up to three months. To process the original call statistics data in the Prime Collaboration databases, a specific monitor instance must be created in the Unified SSM Administration console and that monitor instance must be associated with the appropriate Unified SSM agent. The monitor instance extracts only the data based on predefined attributes. For example, for Call Volume Monitor, the attributes include number of completed calls on-net, number of failed calls on-net, average duration per call on-net, and so forth. Each monitor instance has a unique list of predefined attributes. The monitor instance then polls and extracts the data every 15 minutes, and the Unified SSM agent aggregates the processed data from its associated monitor instance(s) and sends it to Unified SSM every 30 minutes.

For a comprehensive list on all attributes of each monitor type and its configuration guidelines, refer to the Cisco Unified Service Statistics Manager product documents available at http://www.cisco.com.

Note Currently there is no redundancy or failover support with Unified SSM. Unified SSM can still provide reports for more than three months because data is not completely purged but is summarized or aggregated and kept in its database.

Unified SSM Server Performance

Unified SSM operates only in a single-server mode. For hardware requirements and information about Unified SSM, refer to the *Cisco Unified Service Statistics Manager Data Sheet*, available at

http://www.cisco.com/en/US/products/ps7285/products_data_sheets_list.html

Ports and Protocol

Table 28-5 lists the ports used by the various protocol interfaces for Cisco Unified Service Statistics Manager. Cisco recommends opening these ports in the corporate internal firewalls (if applicable) to allow communication between Unified SSM and other devices in the network.

Table 28-5 Unified SSM Port Utilization

Protocol	Port	Service
TCP	48101	HTTP
TCP	48443	HTTPS
TCP	12123	Unified SSM Agent Controller Listener
TCP	12124	Unified SSM Agent Listener[1]
TCP	12125	Unified SSM and Unified SSM agent communication[2]

[1] Unified SSM connects all distributed Unified SSM agents.

[2] Unified SSM agents send call statistics data back to Unified SSM.

Provisioning

Cisco Prime Collaboration provides a simplified web-based provisioning interface for both new and existing deployments of Cisco Unified Communications Manager (Unified CM), Cisco Unified Communications Manager Express (Unified CME), Cisco Unity, Cisco Unity Connection, and Cisco Unity Express. Prime Collaboration provides provisioning for both the infrastructure and subscribers (or phone users) for Day 1 and Day 2 needs. Day 1 needs include configuring new deployments and adding more sites or locations; Day 2 needs include services for ongoing moves, adds, and changes on various components of the Cisco Unified Communications solution.

Cisco Prime Collaboration also exposes northbound APIs to allow Cisco and third parties to integrate with external applications such as HR systems, custom or branded user portals, other provisioning systems, and directory servers.

For details on Prime Collaboration system requirements and installation steps, provisioning users and the infrastructure of supported components, and capacity information, refer to the Cisco Prime Collaboration documentation available at

http://www.cisco.com/en/US/products/ps11480/index.html

To provide a better understanding of how Prime Collaboration can be used as a network management solution for provisioning various Cisco Unified Communications components, the next section presents some of the basic concepts of Prime Collaboration.

Provisioning Concepts

Cisco Prime Collaboration serves as a provisioning interface for the following components of a Cisco Unified Communications system:

- Call processors
  - Cisco Unified Communication Manager (Unified CM)
  - Cisco Unified Communications Manager Express (Unified CME)

- Message processors
  - Cisco Unity
  - Cisco Unity Connection
  - Cisco Unity Express
- Presence processors
  - Cisco IM and Presence
  - Cisco Voice Gateways
  - Cisco VG224, VG204, and VG202 Analog Voice Gateways

Note For more information on component version compatibility, refer to the Prime Collaboration information at http://www.cisco.com/en/US/products/ps11480/index.html.

The following sections describe some of the Prime Collaboration concepts involved in configuring those components.

Domain

Domains are used for administrative purposes to create multiple logical groups within a system. Domains have the following characteristics:

- A domain can be mapped to a geographical location or an organization unit.
- One domain can contain multiple call processors and multiple optional message processors.
- A given call processor or message processor can be a member of multiple domains.
- A domain can partition subscribers so that they can be administered separately.

Service Area

Service areas represent offices. Service areas determine the dial plans and other voice-related configuration settings in the domain. In reality, each office may have multiple service areas. The service area determines attributes such as device group, route partition, and calling search space used within Unified CM. Service areas have the following characteristics:

- Each service area is assigned to a single call processor and one optional message processor.
- Each service area should be associated with one dial plan.

Users and Subscribers

A *user* is a person who is authorized to perform various tasks in Prime Collaboration, based on assigned user roles. When installed, Prime Collaboration creates a Prime Collaboration Admin (also called a Super Admin in Prime Collaboration) who has global administrative rights and complete authorization to perform all tasks in Prime Collaboration.

User roles determine the level of access within Prime Collaboration. Domain-specific users can be assigned more than one user role to have rights to specific tasks in a domain. Individual user roles are related to either policy or workflow tasks. A user can be an administrator or a phone user.

A *subscriber* in Prime Collaboration is an entity that uses IP telephony services provided by the underlying voice applications. A subscriber is the same as a phone user in Unified CM. Users in Prime Collaboration can also have services themselves; thus, a user (an administrator) can also be a subscriber (or a phone user). There can also be pseudo-subscribers (for example, conference rooms and lobby phones) in Prime Collaboration that are not present in Unified CM.

Work Flow and Managing Orders

When deploying a new site or making moves, adds and changes to an existing site, users make all changes to the underlying systems through a two-stage process of creating an order and then processing that order. You can set policies for both of these stages. For example, you can configure the system so that one group of users can only create and submit orders, while another group of users can view and perform processing-related activities. Prime Collaboration contains an automation engine that performs the order processing, including service activation and business flow, based on how Prime Collaboration is configured.

The workflow coordinates activities of the ordering process (approval, phone assignment, shipping, and receiving).

Configuration Templates

Prime Collaboration enables you to configure Unified CM, Unified CME, Cisco Unity, Cisco Unity Express, and Cisco Unity Connection in a consistent way through the use of configuration templates. You can use these templates to configure any of these products, to perform an incremental rollout on these existing products, and to deploy a new service across existing customers.

Batch Provisioning

Creating users and provisioning their services can also be done automatically through batch provisioning for rolling out a new office or transitioning from legacy systems.

Best Practices

The following best practices and guidelines apply when using Prime Collaboration to provision Cisco Unified Communications components for any new and/or existing deployments:

- Managed devices must be up and running before using Prime Collaboration for further day-one activities such as rolling out a new site and day-two activities such as moves, adds, and changes.
- Pre-configuration is required for Cisco Unified CM, Cisco Unity, Unified CME, Survivable Remote Site Telephony (SRST), Cisco Unity Express, and Cisco IM and Presence Service.
- Define the correct domains, service areas, and provisioning attributes.
- Modify only the workflow rules if necessary.
- Consider the use of Subscriber Types, Advanced Rule settings, and other configuration parameters.

The following basic tasks help support these best practices:

- Add call processors such as Unified CM, and/or Unified CME and message processors such as Cisco Unity, Unity Connection, and/or Unity Express.
- Create domains and assign call processors and message processors to the created domains.
- Provision the voice network by creating and using templates to configure Unified CMs or Unified CMEs, or import current voice infrastructure configurations from an existing deployment.
- Perform bulk synchronization of LDAP users into Prime Collaboration, if applicable.
- Set up the deployment by creating service areas for each domain (typically one per dial plan) and assigning subscriber (user) types to each service area.
- Create administrative users for each domain.
- Order, update, or change subscriber or user services.

Prime Collaboration Design Considerations

The following design considerations apply to Prime Collaboration for provisioning:

- Set up domains in one of the following ways:
  - Create a single domain for multiple sites, with multiple call processors and multiple message processors.
  - Create a domain for each site, consisting of one call processor and zero or more optional message processors.
  - Create multiple domains if different administrators are required to manage a subset of the subscribers.
- Create multiple service areas for multiple dial plans.
- Add only the Unified CM publisher as the call processor for Prime Collaboration. Any changes made to the Unified CM publisher through Prime Collaboration will be synchronized to all the Unified CM subscriber servers.
- Use configuration templates for Unified CM, Unified CME, or Cisco Unity Express.
- Use Cisco IOS commands for Unified CME and Cisco Unity Express configuration templates.
- Add Cisco Unified CM infrastructure data objects for Unified CM configuration templates.
- Change and modify the existing configuration templates for batch provisioning for large quantities of phones and lines (DNs).
- Create multiple domains if you want different domain administrators to manage different sets of subscribers for Day 2 moves, adds, and changes of services (such as phones, lines, and voicemail), even for a single-site deployment.
- Create one service area for one dial plan.
- Create multiple service areas if multiple dial plans are required for the device pools, location, calling search space, and phones.
- Prime Collaboration is an IPv6-aware application with the following characteristics:
  - Prime Collaboration communicates with Unified CM over an IPv4 link. The Prime Collaboration user configuration interface allows users to enter only IPv4 IP addresses because Unified CM has SOAP AXL interfaces in IPv4 only. Therefore, Prime Collaboration must use IPv4 addresses to communicate with the AXL interfaces on Unified CM.
  - Prime Collaboration handles the IPv6 addresses contained in SIP trunk AXL response messages.
  - Support of IPv6-aware functions does not affect support for current Cisco Unified Communications Manager Express, Cisco Unity, Cisco Unity Express, and Cisco Unity Connection devices.

Redundancy and Failover

If Prime Collaboration fails in the middle of the configuration process, changes made to the configured devices from the Prime Collaboration GUI might not be saved and cannot be restored. Administrators must use manual steps to continue the configuration process by using other tools such as telnet or login (HTTP) to the managed devices until Prime Collaboration comes back live. Manually added configuration changes to the managed device will not automatically show up in the Prime Collaboration dashboard or database unless you also perform synchronization from Prime Collaboration for the call processors (Unified CM and/or Unified CME), message processors (Cisco Unity, Unity Connection, and/or Unity Express), and domains.

Provisioning Ports and Protocol

Table 28-6 lists the ports used by the various protocol interfaces for Prime Collaboration. Cisco recommends opening those ports in the corporate internal firewalls (if applicable) to allow communication between Prime Collaboration and other devices in the network.

Table 28-6 Prime Collaboration Port Utilization for Provisioning

Protocol	Port	Service
TCP	80	HTTP[1] [2]
TCP	8443	HTTPS[2]
TCP	22	SSH[3]
SSH	23	Telnet[3]
TCP	1433	Database[4]

[1] To access the Prime Collaboration Administration web page.

[2] Prime Collaboration provisions Unified CM via Administrative XML Layer (AXL) Simple Object Access Protocol (SOAP).

[3] For Prime Collaboration to communicate with Unified CME and Cisco Unity Express.

[4] For Prime Collaboration to connect to the database of Cisco Unity and Cisco Unity Connection.

Additional Tools

In addition to the network management tools mentioned above, the following tools also provide troubleshooting and reporting capabilities for Cisco Unified Communications systems:

- Cisco Unified Analysis Manager, page 28-19
- Cisco Unified Reporting, page 28-20

Cisco Unified Analysis Manager

Cisco Unified Analysis Manager is included with the Cisco Unified Communications Manager Real-Time Monitoring Tool (RTMT). Unlike the other RTMT functions, Unified Analysis Manager is unique in that it supports multiple Unified Communications elements instead of just one. When the Unified Analysis Manager is launched, it collects troubleshooting information from your Unified Communications system and provides an analysis of that information. You can use this information to perform your own troubleshooting operations, or you can send the information to Cisco Technical Assistance Center (TAC) for analysis.

Unified Analysis Manager supports the 8.*x* and later versions of the following Unified Communications elements:

- Cisco Unified Communications Manager
- Cisco Unified Contact Center Enterprise
- Cisco Unified Contact Center Express

- Cisco IOS Voice Gateways (3700 Series, 2800 Series, 3800 Series, 5350XM, and 5400XM)
- Cisco Unity Connection
- Cisco IM and Presence

Unified Analysis Manager provides the following key features and capabilities:

- Supports collection of Unified Communications application hardware, software, and license information from Unified Communications elements.
- Supports setting and resetting of trace level across Unified Communications elements.
- Supports collection and export to a define FTP server of log and trace files from Unified Communications elements.
- Supports analysis of the call path (call trace capability) across Unified Communications elements.

For more details on the report options, refer to the information about the Cisco Unified Analysis Manager in the *Cisco Unified Real-Time Monitoring Tool Administration Guide*, available at

http://www.cisco.com/en/US/docs/voice_ip_comm/cucm/service/8_5_1/rtmt/RTMT.html

Cisco Unified Reporting

The Cisco Unified Reporting web application generates reports for troubleshooting or inspecting Cisco Unified Communications Manager cluster data. It is a convenient tool that you can access from the Unified Communications Manager console. The tool facilitates gathering data from existing sources, comparing the data, and reporting irregularities. For example, you can view a report that shows the hosts file for all servers in the cluster. The application gathers information from the publisher server and each subscriber server. Each report provides data for all active cluster nodes that are accessible at the time the report is generated.

For example, the following reports can be used for general management of a Unified CM cluster:

- Unified CM Cluster Overview — Provides an overview of the cluster, including Unified CM version, hostname, and IP address of all servers, a summary of the hardware details, and so forth.
- Unified CM Device Counts Summary — Provides the number of devices by model and protocol that exist in the Cisco Unified Communications Manager database.

The following report can be used for debugging a Unified CM cluster:

- Unified CM Database Replication Debug — Provides debugging information for database replication.

The following report can be used for maintenance of a Unified CM cluster:

- Unified CM Database Status - Provides a snapshot of the health of the Unified CM database. This report should be generated before an upgrade to ensure the database is healthy.

For more information on the report options, refer to the latest version of the *Cisco Unified Reporting Administration Guide*, available at

http://www.cisco.com/en/US/products/sw/voicesw/ps556/prod_maintenance_guides_list.html

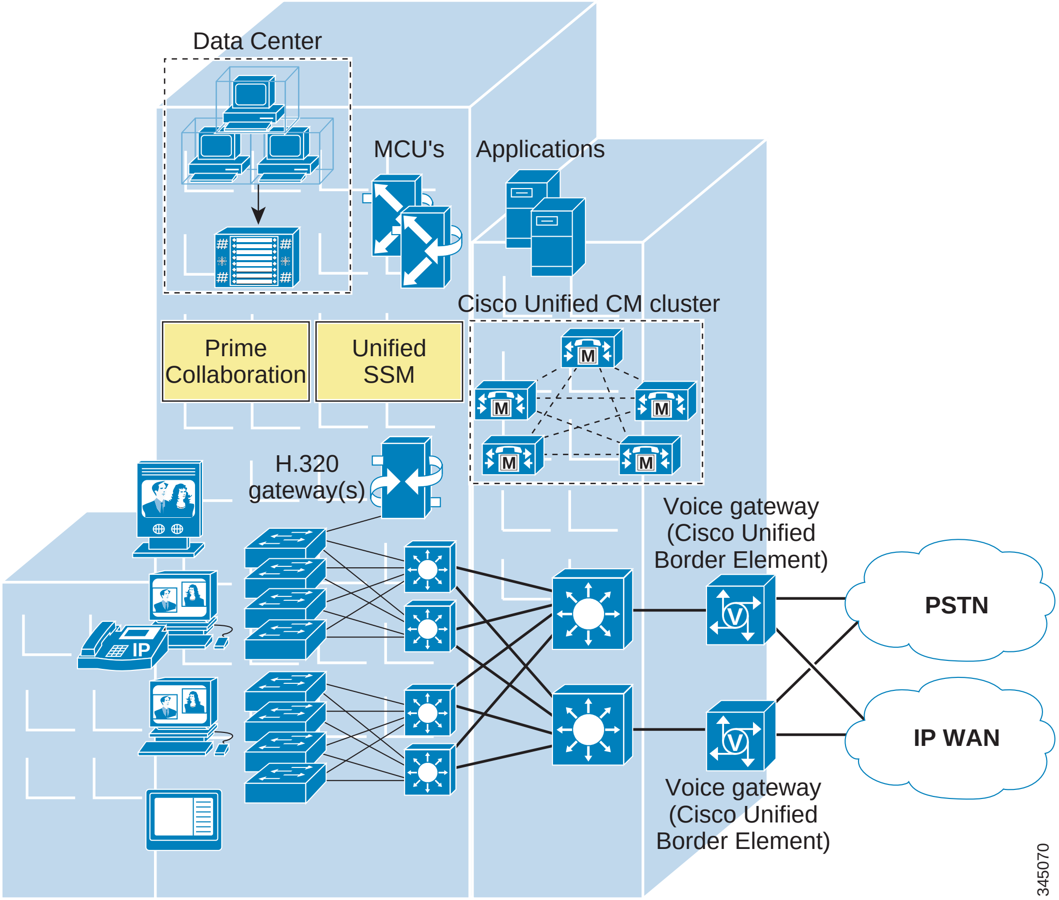
Integration with Cisco Unified Communications Deployment Models

This section discusses how to deploy Cisco Unified Network Management applications in various Cisco Unified Communications deployment models. For detailed information on the deployment models, see the chapter on Unified Communications Deployment Models, page 5-1.

Campus

In the campus model, Cisco Unified Network Management applications, along with call processing agents, are deployed at a single site (or campus) with no telephony services provided over an IP WAN. An enterprise would typically deploy the single-site model over a LAN or metropolitan area network (MAN). Figure 28-3 illustrates the deployment of Cisco Unified Network Management applications in the single-site model.

Figure 28-3 Campus Deployment



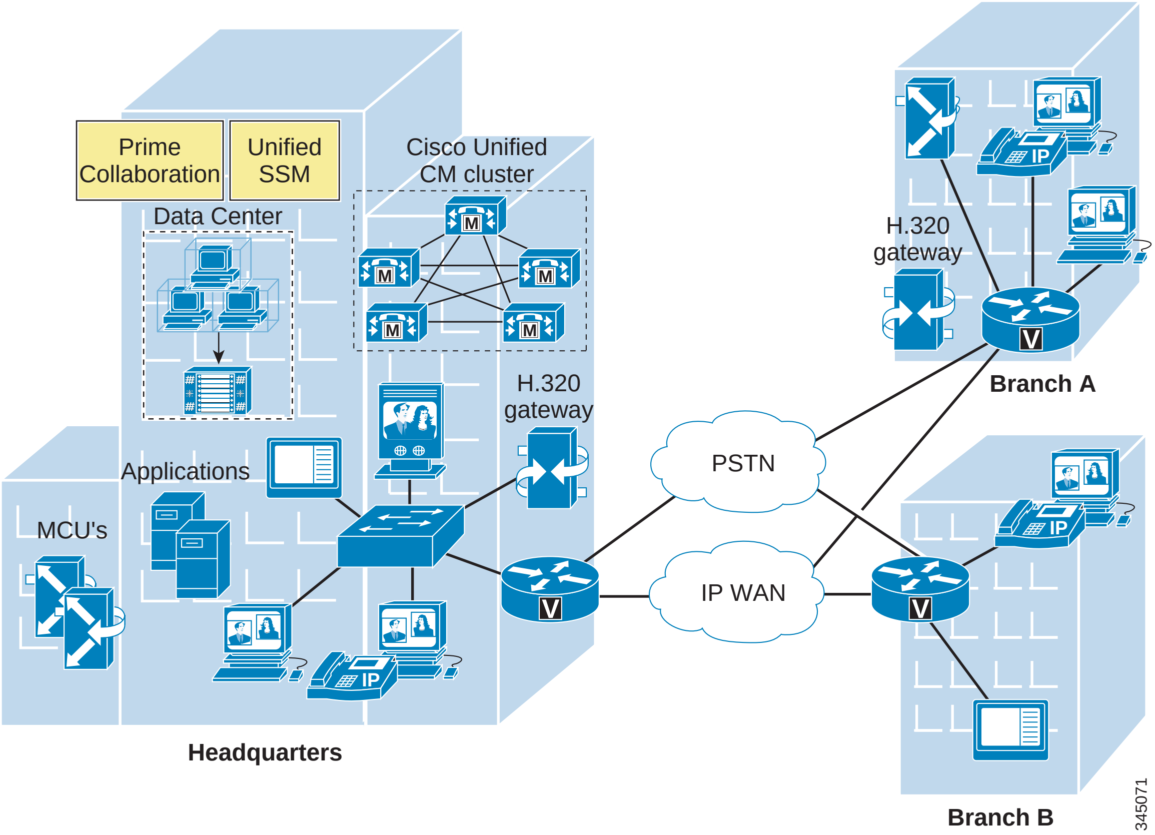
The following design characteristics and recommendations apply to the single-site model for deploying Prime Collaboration and Unified SSM:

- Cisco recommends deploying CVTQ-based voice quality monitoring to monitor overall voice quality in the network.
- Cisco recommends deploying the Cisco 1040 Sensor or NAM to monitor key IP phone devices, gateway devices, and application servers in the network and to investigate and troubleshoot voice quality issues.
- Each Prime Collaboration can support a maximum of 150,000 IP phones and 30 Unified CM clusters.
- Prime Collaboration can support, concurrently, a maximum of 90,000 RTP streams per hour being monitored by the Cisco 1040 Sensor and 15,000 CVTQ-based calls per hour being monitored by Unified CM.

Multisite WAN with Centralized Call Processing

The multisite WAN model with centralized call processing is really an extension of single-site model, with an IP WAN between the central site and remote sites. The IP WAN is used to transport voice traffic between the sites and call control signaling between the central site and the remote sites. Figure 28-4 illustrates the deployment of Cisco Unified Network Management applications in a multisite WAN model with centralized call processing.

Figure 28-4 Multisite WAN Deployment with Centralized Call Processing



The following design characteristics and recommendations apply to the multisite model for deploying Prime Collaboration and Unified SSM with centralized call processing:

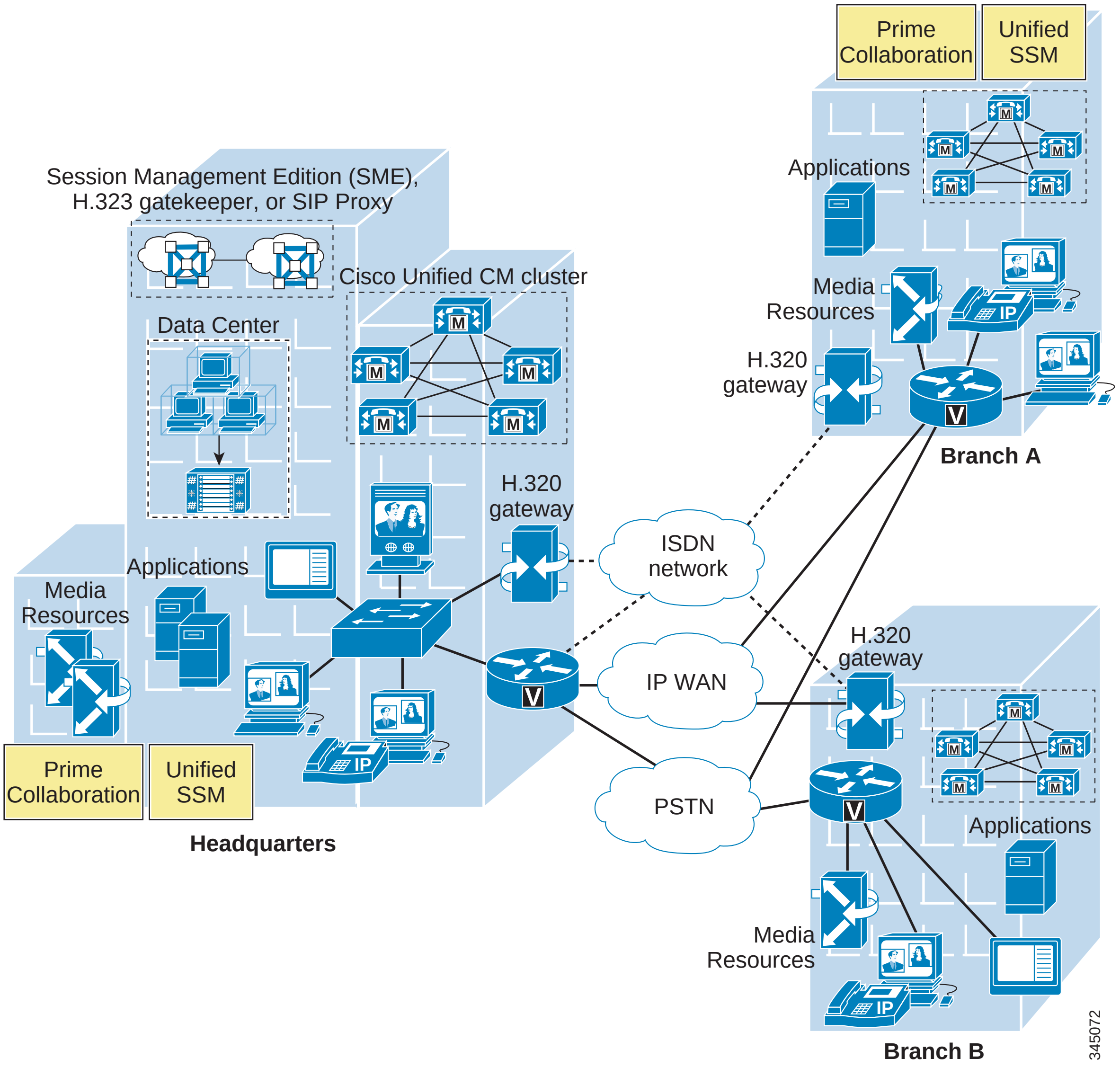
- Cisco recommends deploying all network management applications (including Prime Collaboration and Unified SSM) in the central site to locate them with the call processing agent. The benefit of such an implementation is that it keeps the network management traffic between call processing agent and network management applications within the LAN instead of sending that traffic over the WAN circuit.
- Multiple Prime Collaborations can be deployed, with each instance managing multi-site and multi-cluster Unified Communications environments. In this deployment scenario, Cisco recommends that you deploy a Manager of Managers (MoM). Each Prime Collaboration can provide real-time notifications to the higher-level MoM using SNMP traps, syslog notifications, and email to report the status of the network being monitored.
- Each Prime Collaboration can support a maximum 150,000 IP phones.

- Cisco recommends deploying CVTQ-based voice quality monitoring to monitor overall voice quality in the network.
- Cisco recommends deploying the Cisco 1040 Sensor or NAM to monitor key IP phone devices, gateway devices, and application servers in the network and to investigate and troubleshoot voice quality issues.
- Prime Collaboration can support, concurrently, a maximum of 90,000 RTP streams per hour being monitored by the Cisco 1040 Sensor and 15,000 CVTQ-based calls per hour being monitored by Unified CM.
- Each Unified SSM can support a maximum of 45,000 IP phones.

Multisite WAN with Distributed Call Processing

The multisite WAN model with distributed call processing consists of multiple independent sites, each with its own call processing agent connected to an IP WAN. Figure 28-5 illustrates the deployment of Cisco Unified Network Management applications in a multisite WAN model with distributed call processing.

Figure 28-5 Multisite WAN Deployment with Distributed Call Processing



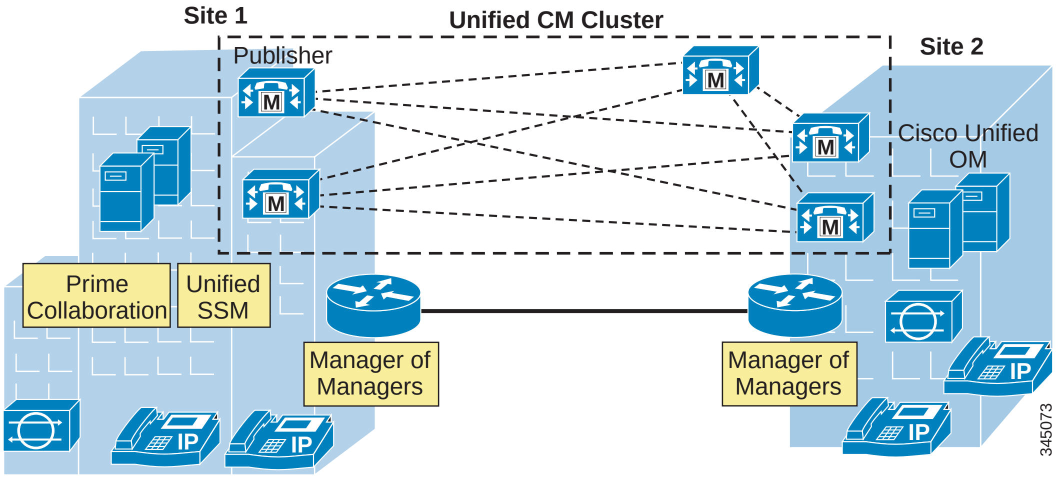
A multisite WAN deployment with distributed call processing has many of the same requirements as a single site or a multisite WAN deployment with centralized call processing in terms of deploying Prime Collaboration and Unified SSM. Follow the best practices and recommendations from these other models in addition to the ones listed here for the distributed call processing model:

- If only one Cisco Unified Network Management system is deployed to manage multiple Unified CM clusters, Cisco recommends deploying Prime Collaboration and Unified SSM along with the Unified CM cluster that has the highest call volume and the most endpoints.
- Multiple Prime Collaborations can be deployed, with each instance managing multi-site and multi-cluster Unified Communications environments. In this deployment scenario, Cisco recommends that you deploy a Manager of Managers (MoM). Each Prime Collaboration can provide real-time notifications to the higher-level MoM using SNMP traps, syslog notifications, and email to report the status of the network being monitored.
- Each Prime Collaboration can support a maximum 150,000 IP phones.
- Cisco recommends deploying CVTQ-based voice quality monitoring to monitor overall voice quality in the network.
- Cisco recommends deploying the Cisco 1040 Sensor or NAM to monitor key IP phone devices, gateway devices, and application servers in the network and to investigate and troubleshoot voice quality issues.

Clustering over the WAN

Clustering over the WAN refers to a single Cisco Unified CM cluster deployed across multiple sites that are connected by an IP WAN with QoS features enabled. This deployment model is designed to provide call processing resiliency if the IP WAN link fails. Figure 28-6 illustrates the deployment of Cisco Unified Network Management applications with clustering over the WAN.

Figure 28-6 Clustering over the WAN



Note There is no native high-availability or redundancy support for Prime Collaboration or Unified SSM with this model.

The following design characteristics and recommendations apply when deploying Prime Collaboration and Unified SSM with clustering over the WAN:

- Cisco recommends deploying Prime Collaboration and Unified SSM in the headquarter site where Unified CM publisher is located.
- Multiple Prime Collaborations can be deployed, with each instance managing multi-site and multi-cluster Unified Communications environments. In this deployment scenario, Cisco recommends that you deploy a Manager of Managers (MoM). Each Prime Collaboration can provide real-time notifications to the higher-level MoM using SNMP traps, syslog notifications, and email to report the status of the network being monitored.
- Cisco recommends deploying CVTQ-based voice quality monitoring to monitor overall voice quality in the network.
- Cisco recommends deploying the Cisco 1040 Sensor or NAM to monitor key IP phone devices, gateway devices, and application servers in the network and to investigate and troubleshoot voice quality issues.
- Each Prime Collaboration can support a maximum 150,000 IP phones.
- Prime Collaboration can support, concurrently, a maximum of 90,000 RTP streams per hour being monitored by the Cisco 1040 Sensor and 15,000 CVTQ-based calls per hour being monitored by Unified CM.
- Each Unified SSM can support a maximum of 45,000 IP phones.

CHAPTER 29

Unified Communications Design and Deployment Sizing Considerations

Revised: October 31, 2012; OL-27282-05

An accurate estimation of the type and quantity of hardware platforms is a prerequisite to a successful deployment of Unified Communications products. Adequate computing and network resources must be provided so that the expected service goals are met.

Each Unified Communications product publishes its capacity limits on each hardware server platform where it runs. These published limits are obviously an important part of determining the needed amount of hardware resources. Individual products, however, may publish only their best-case performance numbers, or may publish numbers for a typical deployment. Both of these numbers are very useful but insufficient for a real-world sizing exercise. For example, Cisco Unified Communications Manager (Unified CM) publishes the maximum number of endpoints that a cluster consisting of Cisco MCS-7845-I3 servers can support. This number may assume average call rates and the absence of any other major activity in the cluster. In an actual usage scenario the call rate might be higher than that assumed, or there might be a requirement to support other services, and even though the nominal number of phones is not exceeded, a single cluster might be inadequate.

Another complexity arises from the fact that each product is rarely used just by itself. Most products are used as part of a larger deployment containing other Unified Communications products. For example, Cisco Unity Connection is likely used with Unified CM and gateways. Larger, more complex deployments may consist of several Unified Communications products, including those from the Cisco Contact Center portfolio (Cisco Unified Contact Center Enterprise, Unified Customer Voice Portal, Unified Intelligence Center, and others), which must work with Unified CM, gateways, Cisco Unified MeetingPlace, Cisco Unity Connection voice messaging, and network management applications. Interaction of each of these components on the others must be taken into account. For example, Unified CM might have to manage not only its regular phones but also the ones assigned to agents who can experience much higher call volumes. Also, gateways might have to handle VXML calls in addition to the regular voice calls. All of these interactions must be taken into account for an accurate sizing estimation.

This chapter discusses the sizing of individual Unified Communications components as well as systems consisting of several components communicating with each other. This chapter also discusses the performance impact of the different functions that the various Unified Communications products support, and it explains why "designing by datasheets" is not the preferred way to deploy a complex Unified Communications network. In addition, this chapter provides insights on how to work with the various sizing tools available, mostly notably the Cisco Unified Communications Sizing Tool.

Note This chapter should be read in conjunction with the product descriptions and design and deployment considerations also covered in other chapters of this document. A good understanding of both of these aspects is required for a successful deployment.

What's New in This Chapter

Table 29-1 lists the topics that are new in this chapter or that have changed significantly from previous releases of this document.

Table 29-1 New or Changed Information Since the Previous Release of This Document

New or Revised Topic	Described in:	Revision Date
Collaborative conferencing	Collaborative Conferencing, page 29-47	October 31, 2012
CTI resources	Applications and CTI, page 29-30	August 31, 2012
Sizing information for Cisco Unified Mobility	Cisco Unified Mobility for Cisco Business Edition, page 29-18	June 28, 2012
Sizing information for Unified CM with Cisco Collaboration Clients and Applications	Cisco Collaboration Clients and Applications, page 29-23	June 28, 2012
Sizing information for LDAP directory integration	LDAP Directory Integration, page 29-37	June 28, 2012
Other minor updates for Cisco Unified Communications System Release 9.0	Various sections throughout this chapter	June 28, 2012

Factors That Affect Sizing

Unified Communications products are designed to be scalable. Capacity of a particular service can generally be increased either by stepping up to a higher-capacity server or by increasing the number of servers. Each product lists the servers it supports and its scalability model. The products also list their tested limits on the servers they support. In theory, one can simply follow these limits and models and come up with the required number of servers for a particular deployment.

In practice, however, sizing is not so simple. For one thing, there are several limits that apply to any deployment. For example, a Unified CM server may be qualified to register 2,500 users and define up to 500 regions. The Unified CM cluster composed of such servers will need more servers if either of these limits is exceeded while the other values are still within limits. Moreover, some of these limits are not absolute but change dynamically based on what else has been configured in the system.

The other major challenge in a sizing exercise is the interaction among components. Unified CM plays a central role in almost all Unified Communications deployments, and it is affected by how customers choose to use other systems. For example, the addition of Cisco Unified MeetingPlace to enable conferencing would tend to concentrate a large number of call setups into a short period (at the beginning of conferencing sessions) and thereby increase the stress on Unified CM during that short period, and this must be accounted for in Unified CM sizing.

Server variations also need to be considered. For example, Unified CM running on a Cisco MCS-7815 or MCS-7816 server is only a standalone entity and may not be clustered. Similarly, different models of Cisco Integrated Service Routers (ISR) have restrictions on the number and types of network modules or Services Ready Engine (SRE) modules they can host.

From a customer perspective, the sizing exercise consists of itemizing all of the functions that are expected in the proposed deployment. Some of these performance factors are obvious, but others are not. For example, one may correctly surmise that the busy hour call attempts (BHCA) that the system is expected to handle is a key factor of performance expectations. But there are nuances even in BHCA that need consideration, such as the types of calls. There are variations in resources consumed by each call type: calls between phones in the same server, calls between two servers in the same cluster, calls between two clusters, and calls that flow to and from the PSTN. Even calls from different types of phones and gateways are different, depending on the protocol and services such as video. The expected number of phones and users is another example of an obvious factor that would affect sizing. Here again, the type of phones, the number of lines that they are configured with, and whether they are in secure mode, among other things, have an impact on Unified CM sizing.

Because of all these factors and possible variations, a proper sizing exercise is complex and must be well understood, especially for large deployments. This chapter provides guidance on the significant factors that consume resources, and their impact on the system, which must be estimated accurately in order to do a complete and accurate sizing.

Cisco Unified Communications Sizing Tools

You should not expect to be able to perform sizing for complex systems after reading this chapter. On the contrary, a manual calculation of all the sizing factors is not practical. However, this chapter will enable you to gain an appreciation of the factors that significantly affect the performance of the system as a whole and that must be accounted for in any sizing effort.

To assist in accurate sizing, Cisco provides several tools that do the calculations based on these significant performance factors. These tools take into account data from testing, individual server performance, advanced and new features in product releases, design recommendations from this SRND, and other factors. The tools allow you to enter specific deployment information, and they apply their sizing algorithms on your supplied data to recommend a set of hardware resources. Obviously, for a desired deployment, this recommendation is only as good as the accuracy of the input data. User guides for the tools contain an explanation of the inputs and how they can best be collected from an existing system or estimated for a system still in the design stage.

The sizing tools are available at http://tools.cisco.com/cucst and they include the following:

- Cisco Unified Communications Sizing Tool — Guides users through a complete system deployment consisting of Cisco Unified CM, voice messaging, conferencing, gateways, Cisco Intercompany Media Engine (IME), Cisco Unified Communications Management Suite, and Cisco Unified Contact Center components.
- Cisco Unified Communications Manager Session Management Edition (SME) Sizing Tool — A specialized tool that focuses on the specific functions of a Unified CM Session Management Edition deployment.
- Cisco VXI Sizing and Configuration Tool — A specialized tool for sizing the Cisco Virtual Experience Infrastructure (VXI).

Access to the sizing tools is limited to users who have a qualified Cisco login account. For more information on these tools and their access privileges, refer to the *Unified Communications Sizing Tool Frequently Asked Questions (FAQ)*, available at

http://tools.cisco.com/cucst/help/ucst_faq.pdf

Caution If any parameter of your system design exceeds the range of values that the above sizing tools allow you to enter, you must consult a Cisco Certified Systems Engineer (CCSE) about your design before proceeding any further with it.

Unified Communications Sizing Compared with PBX Sizing

PBX sizing in the past has mostly been about PSTN access trunks. Consequently, the processes for determining how many trunk circuits are required for a given user base and the desired level of service are well documented. Well known models such as the Erlang B, Extended Erlang B, Erlang C, and other models are used for that purpose. However, sizing a Unified Communications system is inherently more complex for the following reasons:

- Unified Communications is not a monolithic system. Rather, it is composed of several servers doing different things but communicating with each other.
- Unified CM performs many more functions and provides many more services than a PBX.

Definition of Terms

The following terms are used throughout this chapter:

Simultaneous Calls

The number of calls that are all active in the system at the same time.

Maximum Simultaneous Calls

The maximum number of simultaneous calls in active (talk) state that the system can handle at one time.

Calls per Second

The call arrival rate, described as the number of calls that arrive (that is, new call setup attempts) in one second. Call arrival rates are also often quoted in calls per hour, but this metric is looser in the sense that 100 calls arriving in the last five seconds of an hour provides an average call arrival rate of 100 calls per hour (which is an extremely low rate for a communications system), while it also provides an arrival rate of 20 calls per second (which is a high rate). Sustaining 20 calls per second for an entire hour would result in 72,000 calls per hour. Therefore calls-per-hour is not a very useful metric for ascertaining a system's ability to handle bursty call arrival traffic patterns.

Busy Hour

The busiest hour of the day when people are most likely to use their phones. This hour varies from organization to organization and from industry to industry. But for most it is likely to be either in the morning (for example, 9AM to 10AM) or in the afternoon (for example, 2PM to 3PM).

Busy Hour Call Attempts (BHCA)

The number of calls attempted during the busiest hour of the day (the peak hour). This is the same as the calls-per-second (cps) rating for the busiest hour of the day, but it is expressed over a period of an hour rather than a second. For example, 10 cps would be equal to 36,000 calls per hour. There is also a metric for Busy Hour Call Completions (BHCC), which can be lower than the BHCA (call attempts) under the assumption that not all calls are successful (as when a blocking factor exists). This chapter assumes 100% call completions, so that BHCA = BHCC.

Blocking Factor

The maximum percentage or fraction of call attempts that may be blocked during the busy hour. A blocking factor or 0.0 would mean that the number of circuits is equal to the number of callers, which is unrealistic for most deployments.

Average Hold Time

This is the period of "talk time" on a voice call; that is, the period of time between call setup and tear-down when there is an open speech path between the two parties. A hold time of 3 minutes (180 seconds) is an industry average used for traffic engineering of voice systems. The shorter the hold time on the average call, the greater the percentage of system CPU time spent on setting up and tearing down calls compared to the CPU time spent on maintaining the speech path.

Bursty Traffic

Steady arrival means the call attempts are spaced more or less equally over a period of time. For example, 60 calls per hour at a steady arrival rate would present one call attempt roughly every minute (or approximately 0.02 cps). With bursty arrival, the calls arriving over a given period of time (such as an hour) are not spaced equally but are clumped together in one or more spikes. In the worst case, an arrival rate of 60 calls per hour could offer all 60 calls in a single second of the hour, thus averaging 0 cps for most of the hour with a peak of 60 cps for that one second. This kind of traffic is extremely stressful to communications systems.

Erlang

An Erlang is a unit of measure for communications traffic. It is used to represent the utilization of a resource over a one-hour period. One Erlang means that one resource was used 100% of the hour. This could be due to a single call of one-hour duration or multiple sequential calls whose durations total to one hour. Therefore, if 10 Erlangs are required, it is necessary to have 10 resources to ensure that all traffic is serviced.

Designing for Performance

After analyzing the functional requirements and determining the appropriate products for a Unified Communications system, the next major question is how to design the network so that it is able to adequately deliver acceptable performance as measured by availability, reliability, response time, and quality of service. Can the system cope with the real-time performance requirements, support the desired number of users, and still scale up to meet the increasing needs of the foreseeable future?

To aid the Unified Communications network designers with answers to these questions, Cisco tests each of the products for its performance characteristics. The results are published and broad recommendations are made regarding the size and number of clusters, servers, and other components that should be deployed for supporting the given number of users. To a large extent these test results, combined with the design recommendations in this document, provide sufficient information for most Unified Communications deployments. For others, however, the system designers will need a deeper understanding of how each product works and how users will use it before a viable hardware set can be selected. The selection of such a set can also be complicated by the following concerns that should be addressed:

- System release
- Complexity of the configuration
- Utilization of options such as trace compression, call detail recording (CDR), call management record (CMR) generation, and so forth

- Interaction between individual products
- Anticipated growth
- Use of external applications
- Average and peak usage

Quantitative Analysis of Performance

Testing for performance analyzes the product under test for a set of basic functions it is designed to perform. For example, Unified CM performs many functions and each function requires a finite amount of CPU and memory. Unified CM handles endpoint registrations, user initiated calls, database queries, and many other functions. Performance testing involves testing of each of these basic functions in isolation, measuring the computing resources that are utilized as these functions are executed in an increasing volume.

A quantitative analysis of the performance characteristics of a software system given the hardware platform is done in a series of tests that aim to determine the linear range of the system operations. A linear range is where the amount of resources used and the throughput achieved vary in direct proportion with each other. This range is critical because, if the system does not exhibit linear behavior, its performance is unpredictable. Most systems exhibit linearity within a certain range, beyond which the system's performance becomes unpredictable. Therefore, the design must ensure that the system operates within the parameters of the linear range.

Conversely, putting together a system for deployment consists of decomposing the requirements into sets of basic functions, comparing them against the published test results, and determining the set of servers that meets the performance needs in their linear range of operation.

Performance Modeling of Computer Systems

The first step in determining how much a computer system can accomplish is to itemize the various tasks it is called upon to perform. For example, Unified CM may be required to do all of the following tasks:

- Initialize configured values such as those for endpoints, directory numbers, dial plans, and so forth.
- Perform endpoint registrations, which requires handling the initial registration messages, looking up databases to find their configuration information, and creating configuration files for the endpoints to download.
- Maintain endpoint registrations by handling periodic registration messages
- Handle new call requests, which can be a fairly complex process consisting of ensuring user entitlement, analyzing dialed digits, determining the destination (either another phone, gateway, or trunk), assembling the correct signaling based on rules stored in the database, and transmitting and receiving call signaling messages.
- Provide mid-call feature requests such as transfer and conference.
- Offer user management and requests for functions such as Do Not Disturb, Call Forward, and so forth.

Each of the functions that a computer system performs requires it to spend some of its resources consisting of CPU, memory, and disk I/O.

The linear operating range of the system under test is determined by subjecting the system to a battery of tests. Some tests that attempt to find this range are described in this section. From this linear range of operation, the cost incurred in terms of CPU, memory, and disk I/O can be determined for each incremental unit of the operation.

For example, memory utilization of each additional endpoint of a certain type can be determined from the slope of the line depicting the amount of memory used for a range of endpoints. Similarly, memory utilization for each registering endpoint and for each additional call can be quantified by using the same techniques.

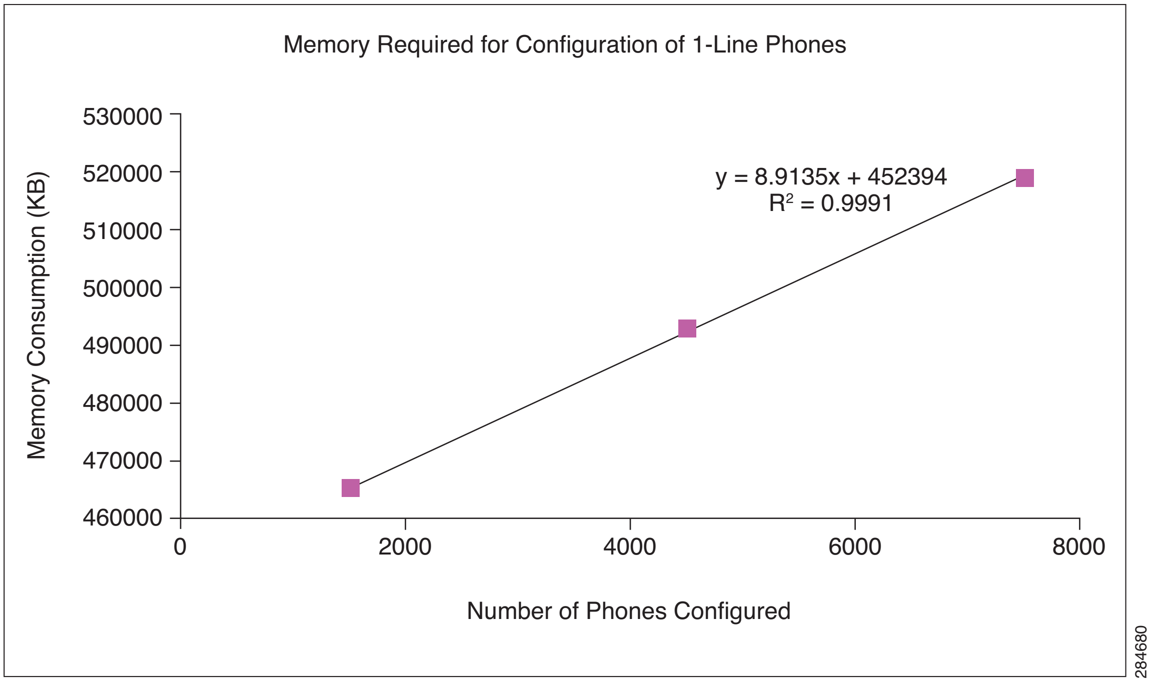
Memory Usage Analysis

Two types of memory usage are identified in the system: static and dynamic. Static memory is defined as the amount of memory that is in use even when there is zero call traffic. This usage of memory arises from configuration data, registration of endpoints, and other factors. Dynamic usage of memory results from call activity. Each active call requires its context to be saved, which results in a certain amount of memory being utilized for the duration of the active call. Thus, whereas static memory is a function of the number of endpoints, dynamic memory is a function of the number of concurrent calls, which itself is a function of the call rate (calls per second) and the average hold time (AHT) per call.

In practice, system memory is also required by the operating system (OS) and by other processes, so the net memory available for operations (static and dynamic memories) is somewhat less than the total memory available on the platform. In addition, some memory is needed for other processes and services running in the system and for any unforeseen spikes in usage.

Figure 29-1 shows the results of a test conducted to determine the memory requirements for configuring one-line phones. It shows the memory consumed by simply configuring 1500, 4500, and 7500 IP phones in Unified CM. Linear regression techniques are used to draw a trend line through the data points. The equation of this trend line is then determined, as is the correlation coefficient R^2. A correlation coefficient of 1 or very close to it (at least 0.99) indicates that the trend is linear and that the equation of the trend line is valid and may be used to predict the dependent variable (in this case memory) based on the control variable (the number of phones).

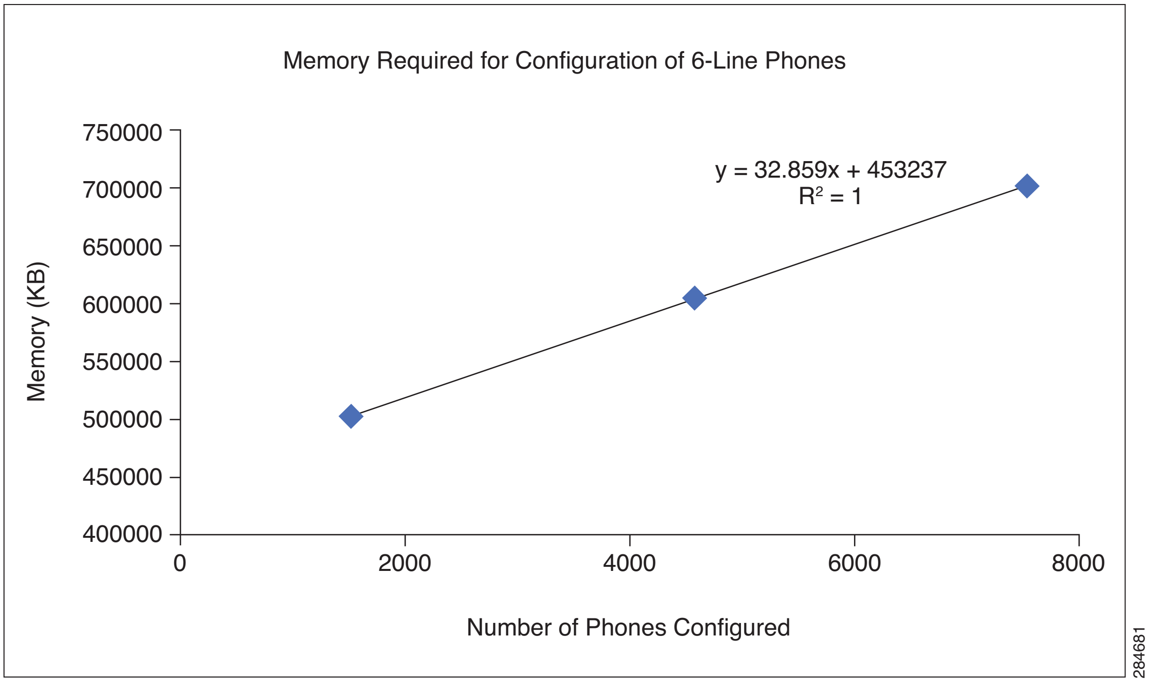
Figure 29-1 Memory Required for Configuration of One-Line Phones



In this particular experiment the R^2 value is extremely close to 1 (discounting small errors in measurement) and the equation for the trend line is valid. From the equation we can derive that the memory consumed with no phones is 452,394 Kbytes (the Y-intercept) and that each additional one-line phone configured in the system consumes 8.91 Kbytes.

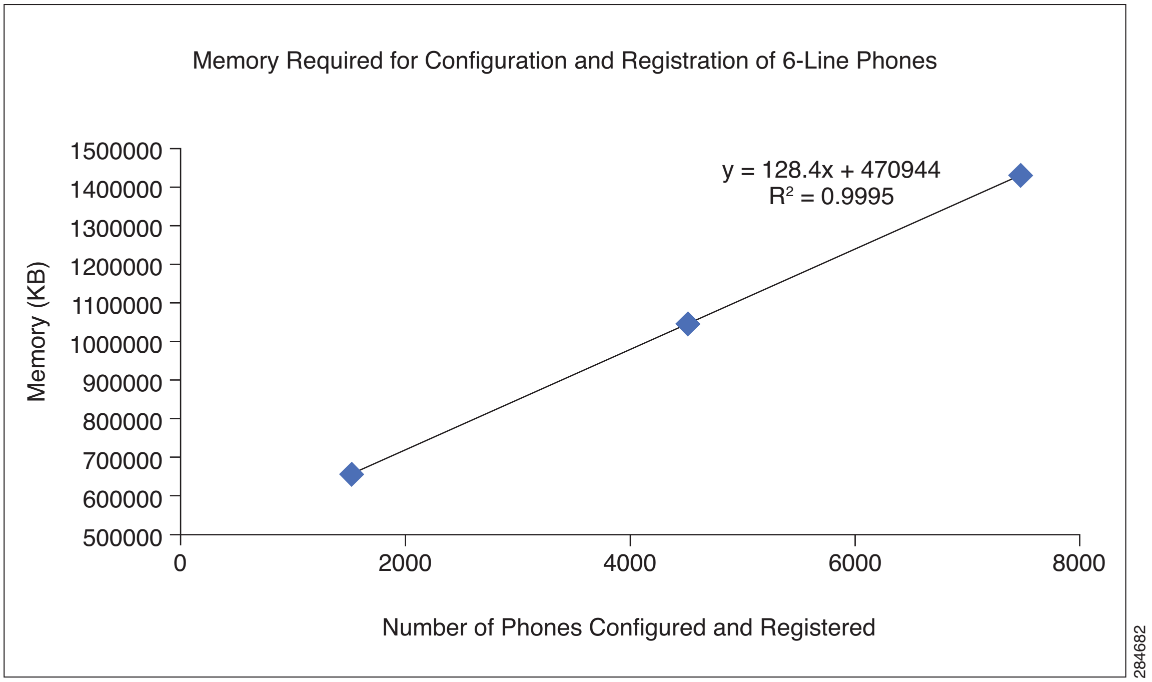
Figure 29-2 depicts the memory requirements for configuring six-line phones. In this chart R^2 is actually equal to 1, indicating that the trend line is a valid model. From the equation we can determine that configuring each six-line IP phone consumes approximately 33 Kbytes of memory.

Figure 29-2 Memory Required for Configuring Six-Line Phones



The other component that makes up the static memory – the memory required for registration of phones – can also be estimated in the same manner. Figure 29-3 shows the tested, measured, and plotted memory requirements for configuring and registering 1500, 4500, and 7500 phones, each with six lines. Note that R^2 is close enough to 1 to make the trend line a valid model. From the equation we can determine that registration of each six-line IP phone consumes approximately 128 Kbytes of memory.

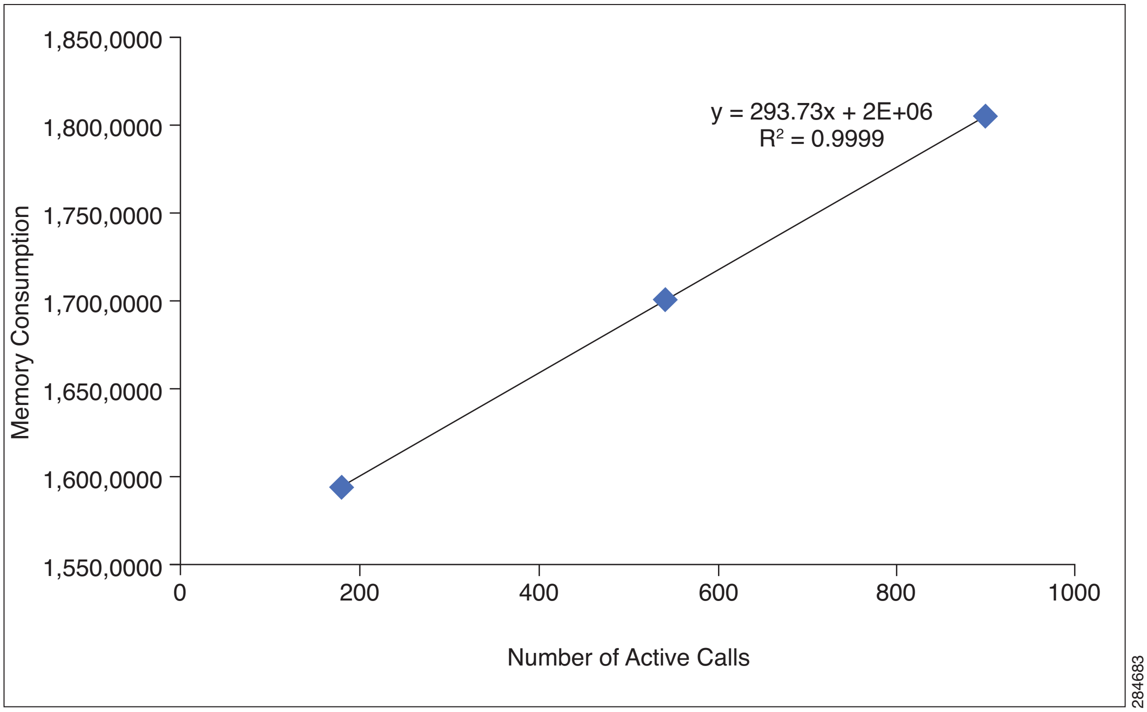
Figure 29-3 Memory for Configuration and Registration of Six-Line Phones



Static memory also includes other configuration items such as partitions, translation patterns, route lists and groups, as well as memory used for CTI and other applications.

Another type of memory called the dynamic memory is defined as the memory used for active calls. In contrast to static memory, which stays allocated all the time, dynamic memory is allocated for each call attempt and remains only until the end of the call. Figure 29-4 shows how the memory is utilized for 180, 540, and 900 active calls on one subscriber node of the Unified CM cluster. The graph shows that the trend line is a good fit and that approximately 294 Kbytes are used for each active call.

Figure 29-4 Memory Consumption Per Active Call



The preceding graphs and analysis are indicative of how memory is measured in the system. From a set of these observations, data may be interpolated that can start to build a memory model for various activities going on in the system. For example, we can estimate:

- Incremental memory required for configuring each additional line
- The maximum number of calls that can exist in the system before it runs out of memory and starts paging

A major determinant of dynamic memory usage is the average call holding time (ACHT), which is the average duration of each call. A longer ACHT means that more memory will be used in the system because there will be a larger number of active calls present at any time.

The description provided in this section has been simplified. Further complexities arise from the variety of phones that can be configured on Unified CM with different protocols, capabilities, security status, and other variables. Each of these variants is tested and analyzed. Furthermore, each of these variables depends on the software release, which could add improvements and new features. For active call measurements, the various types of calls that can be made between different destinations, such as between two SCCP phones or between a SIP phone and an MGCP gateway, is also considered.

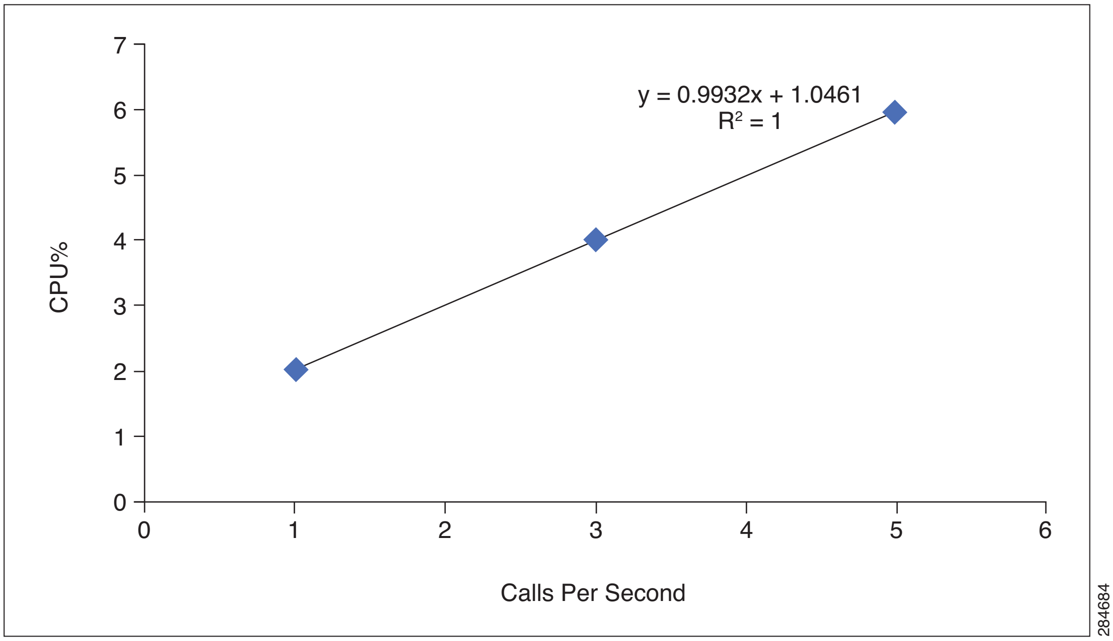
CPU Usage Analysis

Analysis of CPU usage follows the methodology used for memory analysis. While there is some CPU activity even when there are no calls being initiated or terminated, most of the CPU utilization occurs during the process of setting up or tearing down calls. Therefore, one of the key determinants of CPU usage is the call rate.

There are significant differences between the types of calls being made. Calls can originate and terminate within the same server, or they can be made between two servers. Calls can also originate from the Unified CM cluster and travel across a gateway or a trunk. All of these different call activity types impact the CPU differently, so it is important to consider them carefully.

Figure 29-5 shows CPU utilization as measured at 1, 3, and 5 calls per second. Because the trend line is linear, we can conclude that the CPU processing cost required to process one incoming call each second is about 1%.

Figure 29-5 CPU Consumption Per Call Setup



As with memory analysis, CPU usage involves many complexities that must also be considered. For example, CPU usage analysis must account for different costs of terminating and originating calls, different protocols, whether the calls are secure or not, and so forth. CPU usage also depends on whether or not the configuration database is complex or relatively simple, whether CDRs and/or CMRs are being generated, and so forth.

Whereas incremental memory usage is fairly independent of the actual server platform, CPU usage will vary substantially with the actual hardware being tested. Therefore, the same tests must be repeated on all servers that are supported.

Other CPU-intensive call operations such as call transfers, conferences, media resource functions such as MTP or music on hold, and so forth, should also be considered when sizing CPU resources.

Shared lines also consume CPU resources. Not only do shared lines count as extra lines on the phones that share DNs, but each call from or to any of the shared line phones is reflected on all of the other phones as well.

Fundamentals of Voice Traffic Engineering

Traffic engineering is the science of determining an optimum number of resources given the key usage data. In telephony this user data includes the busy hour call attempts (BHCA) and the average hold time (AHT). The BHCA measures all the calls that an average user initiates or receives during the busiest hour of the day. The AHT measures the time that the user spends on the phone for each initiated or received call. An individual's BHCA, when multiplied by the number of users, gives the volume of calls that the system must be able to handle. Once we have the total BHCA and the AHT, we can calculate the Erlang value that the system should be able to handle. One Erlang is a full hour of telephone conversation. For example, if the system BHCA is 10 and the calls last for 3 minutes each, then the system is being used for a total of 30 minutes and the equivalent Erlang value is 0.5.

Note This document assumes that traffic follows the Extended Erlang B model with random arrival pattern and that blocked callers make multiple attempts to complete their calls. For a more thorough discussion of the various Erlang models used in the industry, refer to the information at http://www.erlang.com/calculator/.

While this analysis reveals the total BHCA that the system must be able to handle at the given AHT, another key piece required for analysis is the blocking factor. It is well understood that deploying a telephony system that has enough capacity for all of the users to be on the phone all of the time would be prohibitively expensive, especially for larger systems. It follows, therefore, that if more than a certain number of callers try to access the system at the same time, some callers will necessarily be blocked. A key decision in system deployment is how many over-the-limit callers may be blocked during peak calling times. The amount of resources required for providing a smaller probability of being blocked, say 0.01 or 1%, would be more than the amount required to provide a blocking factor of 0.1 or 10%.

The Erlang value and the blocking factor are useful for calculating the amount of shared resources that must be provisioned in the system. For example, with these pieces of information one can figure out how many DS0s will be required on gateways for a system that has a given number of Erlangs of through traffic with the required blocking factor. This is generally done through an Erlang calculator or lookup tables. The number of required DS0s would increase with the number of Erlangs and decrease with an increase in the blocking probability.

Table 29-2 illustrates the relationship between number of circuits, blocking probability, and busy hour traffic.

Table 29-2 Erlang C Traffic Table (Maximum Offered Load)

Number of Circuits	Blocking Probability							
	0.01%	0.05%	0.1%	0.5%	1.0%	2.0%	5.0%	10.0%
1	0.0001	0.0005	0.0010	0.0050	0.0100	0.0200	0.0500	0.1000
2	0.0142	0.0319	0.0452	0.1025	0.1465	0.2103	0.3422	0.5000
3	0.0860	0.1490	0.1894	0.3339	0.4291	0.5545	0.7876	1.0400
4	0.2310	0.3533	0.4257	0.6641	0.8100	0.9939	1.3190	1.6530
5	0.4428	0.6289	0.7342	1.0650	1.2590	1.4970	1.9050	2.3130

From Table 29-2 we can determine the following information:

- The number of Erlangs that the system can handle increases with the number of circuits and with the blocking factor. Whereas the first relationship is obvious, the second can be understood by realizing that a greater number of calls are being blocked.
- Given an Erlang requirement of 0.50 and a blocking factor of 0.1%, the system would need 5 circuits.
- Assuming we have 5 circuits and a blocking factor of 1%, there would be 1.259 Erlangs available. It then follows that if we have 10 users, each user can talk for (1.259∗3600/10) = 453.24 seconds during the busy hour.

Note Specifically for Cisco Unified Contact Center deployments, there might be other resources that have to be sized according to the same principles. For example, requirements for the number of interactive voice response (IVR) ports and agents are modeled using similar quantitative analysis. Some of the considerations here besides average hold time and BHCA include time waiting in queues and other factors, which means that a higher number of DS0 circuits will be required. For a full description, refer to the *Cisco Unified Contact Center Enterprise SRND*, available at http://www.cisco.com/en/US/products/sw/custcosw/ps1844/products_implementation_design_guides_list.html.

Sizing by Product

This section discusses significant factors that influence sizing of the following individual products and describes how these individual products influence the sizing considerations of other products in the system deployment:

- Cisco Unified Communications Manager Express, page 29-15
- Cisco Business Edition, page 29-15
- Cisco Unified Communications Manager, page 29-19
- Cisco Unified CM Megacluster Deployment, page 29-38
- Cisco Unified CM Session Management Edition, page 29-38
- Cisco Intercompany Media Engine, page 29-39
- Emergency Services, page 29-40
- Media Resources, page 29-35
- Gateways, page 29-40
- Voice Messaging, page 29-46
- Collaborative Conferencing, page 29-47
- Cisco IM and Presence, page 29-53
- Cisco Unified Communications Management Suite, page 29-55

Cisco Unified Communications Manager Express

Cisco Unified Communications Manager Express (Unified CME) runs on one of the Cisco IOS Integrated Services Router (ISR) platforms, from the low-end Cisco 1861 ISR to the high-end Cisco 3945E ISR 2. Each of these routers has an upper limit on the number of phones that it can support. The actual capacity of these platforms to do call processing may be limited by the other functions that they are performing, such as IP routing, Domain Name System (DNS), Dynamic Host Control Protocol (DHCP), and so forth.

Unified CME can support a maximum of 450 endpoints on a single Cisco IOS platform; however, each router platform has a different endpoint capacity based on the size of the system. Because Unified CME is not supported within the Cisco Unified Communications Sizing Tool, it is imperative to follow the capacity information provided in the Unified CME product data sheets available at

http://www.cisco.com/en/US/products/sw/voicesw/ps4625/products_data_sheets_list.html

Cisco Business Edition

The three models of Cisco Business Edition – Business Edition 3000, 5000, and 6000 – offer different capacities measured in terms of number of users, number of endpoints, and maximum call volumes. Table 29-3 describes the pertinent performance characteristics of the three models.

Table 29-3 Capacities of Cisco Business Edition Models

Model	Maximum Number of Users	Maximum Number of Endpoints	Maximum BHCA
Business Edition 3000 (MCS 7816)	300	400	2,200
Business Edition 5000 (MCS 7828)	500	575	3,600
Business Edition 6000 (UCS C200)	1,000	1,200	5,000

Busy Hour Call Attempts (BHCA) for Cisco Business Edition

As shown in Table 29-3, Business Edition 3000 supports a maximum of 2,200 BHCA, Business Edition 5000 supports a maximum of 3,600 BHCA, and Business Edition 6000 supports a maximum of 5,000 BHCA. When calculating your system usage, stay at or below the BHCA maximum shown in Table 29-3 to avoid oversubscribing Cisco Business Edition.

The BHCA consideration becomes significant when the usage for any phone is above 4 BHCA. A true BHCA value can be determined only by taking a baseline measurement of usage for the phone during the busy hour. Extra care is needed when estimating this usage without a baseline.

Device Calculations for Cisco Business Edition

Devices can be grouped into two main categories for the purpose of this calculation: phone devices and trunk devices.

A phone device is a single callable endpoint. It can be any single client device such as a Cisco Unified IP Phone 7900 Series, a software client such as Cisco IP Communicator, an analog phone port, or an H.323 client. While Cisco Business Edition supports a maximum number of endpoints as indicated in Table 29-3, actual endpoint capacity depends on the total system BHCA.

Note Business Edition 3000 supports a limited set of endpoints. For a list of the supported endpoints, refer to the *Administration Guide for Cisco Business Edition 3000*, available at http://www.cisco.com/en/US/products/ps11370/prod_maintenance_guides_list.html.

A trunk device carries multiple calls to more than one endpoint. It can be any trunk or gateway device such as a SIP trunk, a gatekeeper-controlled H.323 trunk, or in the case of Business Edition 3000 an MGCP backhauled PRI trunk.

Business Edition 5000 and 6000 both support intercluster trunking as well as H.323, SIP, and MGCP trunks or gateways and analog gateways. However, Business Edition 3000 does not support intercluster trunking. Business Edition 3000 trunk and gateway support is limited to the Cisco 2901 Integrated Services Router (ISR) for MGCP PSTN connectivity over a maximum of two E1/T1 PRIs. Business Edition 3000 also supports the Cisco VG224 Analog Voice Gateway for analog phones.

The method for calculating BHCA is much the same for both types of devices, but trunk devices typically have a much higher BHCA because a larger group of endpoints is using them to access an external group of users (PSTN or other PBX extensions).

You can define groups of devices (phone devices or trunk devices) with usage characteristics based on BHCA, and then you can add the BHCA for each device group to get the total BHCA for the system, always ensuring that you are within the supported BHCA maximum specified in Table 29-3.

For example, you can calculate the total BHCA for 100 phones at 4 BHCA each and 80 phones at 12 BHCA each as follows:

100 phones at 4 BHCA is 100∗4 = 400

80 phones at 12 BHCA is 80∗12 = 960

Total BHCA = (100∗4) + (80∗12) = 1,360 BHCA for all phones

For trunk devices, you can calculate the BHCA on the trunks if you know the percentage of calls made by the devices that are originating or terminating on the PSTN. For this example, if 50% of all device calls originate or terminate at the PSTN, then the net effect that the device BHCA (1360 in this case) would have on the gateways would be 50% of 1360, or 680 BHCA. Therefore, the total system BHCA for phone devices and trunk devices in this example would be:

Total system BHCA = 1,360 + 680 = 2,040 BHCA

If you have shared lines across multiple phones, the BHCA should include one call leg (there are two call legs per each call) for each phone that shares that line. Shared lines across multiple groups of devices will affect the BHCA for that group. That is, one call to a shared line is calculated as one call leg per line instance, or half (0.5) of a call. If you have different groups of phones that generate different BHCAs, use the following method to calculate the BHCA value:

Shared line BHCA = 0.5∗(Number of shared lines)∗(BHCA per line)

For example, assume there are two classes of users with the following characteristics:

100 phones at 8 BHCA = 800 BHCA

150 phones at 4 BHCA = 600 BHCA

Also assume 10 shared lines for each group, which would add the following BHCA values:

10 shared lines in the group at 8 BHCA = 0.5∗10∗8 = 40 BHCA

10 shared lines in the group at 4 BHCA = 0.5∗10∗4 = 20 BHCA

The total BHCA for all phone devices in this case is the sum of the BHCA for each phone group added to the sum of the BHCA for the shared lines:

800 + 600 + 40 + 20 = 1,460 total BHCA

Note that the total BHCA in each example above is acceptable because it is below the system maximum BHCA as shown in Table 29-3.

If you are using Cisco Unified Mobility for Mobile Connect (also known as single number reach, or SNR) on Business Edition 5000 or 6000, or if you are using the Reach Me Anywhere feature (also SNR), keep in mind that calls extended to remote destinations or off-system phone numbers affect BHCA. In order to avoid oversubscribing the appliance, you have to account for this SNR remote destination or off-system phone BHCA. To calculate the BHCA for these SNR features, see Capacity Planning for Cisco Unified Mobility, page 25-58, and add that value to your total BHCA calculation.

Note Media authentication and encryption using Secure RTP (SRTP) impacts the system resources and affects system performance. If you plan to use media authentication or encryption, keep this fact in mind and make the appropriate adjustments. Typically, 100 IP phones without security enabled results in the same system resource impact as 90 IP phones with security enabled (10:9 ratio).

Note Cisco Business Edition 3000 does not support media authentication or encryption.

Another aspect of capacity planning to consider for Cisco Business Edition is call coverage. Special groups of devices can be created to handle incoming calls for a certain service according to different rules (top-down, circular hunt, longest idle, or broadcast). This is done through hunt or line group configuration within Cisco Business Edition. BHCA can also be affected by this factor, but only as it pertains to the line group distribution broadcast algorithm (ring all members). For Business Edition, Cisco recommends configuring no more than three members of a hunt or line group when a broadcast distribution algorithm is required. Depending on the load of the system, doing so could greatly affect the BHCA of the system and possibly oversubscribe the platform's resources. The number of hunt or line groups that have a distribution algorithm of broadcast should also be limited to no more than three.

Business Edition 5000 with Cisco Unified Contact Center

For this example, assume that Cisco Unified Contact Center Express (Unified CCX) is integrated with Business Edition 5000 and that the system has the following characteristics:

- The required specification is for 15 contact center agents with a maximum of 30 calls per hour during the busiest hour.
- There are 96 non-agent users with average usage of 4 BHCA, and each user has the ability to configure one remote destination for single number reach with Cisco Unified Mobility.
- There are 36 non-agent users with heavy usage of 10 BHCA, and each also has the ability to configure one remote destination for single number reach.
- There are 20 extra shared lines, 10 of which are shared across 10 users from the average usage pool as well as 10 in the heavy usage pool.
- There are 7 T1 trunks (allowing for up to 161 simultaneous calls) with a total of 1200 BHCA across all trunks.

Note Cisco Business Edition 5000 is not supported with Cisco Unified Contact Center Enterprise.

Note This example groups the BHCA for all gateway trunks into a single total trunk BHCA value. This method would be typical for a single-site deployment. However, in a multisite deployment, the various sites' trunks could have different BHCA requirements and thus require different BHCA groupings.

The BHCA calculations for this system are as follows:

15 contact center agents at 30 BHCA = 450 BHCA

96 average-usage users at 4 BHCA = 384 BHCA

36 heavy-usage users at 10 BHCA = 360 BHCA

10 shared lines in the 4 BHCA group = 20 BHCA

10 shared lines in the 10 BHCA group = 50 BHCA

Total of 1200 BHCA for all T1 trunks = 1200 BHCA

One remote destination for single number reach across each of the 96 average-usage users at 4 BHCA = 192 BHCA. (See Cisco Unified Mobility for Cisco Business Edition, page 29-18, for details on this calculation.)

One remote destination for single number reach across each of the 36 heavy-usage users at 10 BHCA = 180 BHCA. (See Cisco Unified Mobility for Cisco Business Edition, page 29-18, for details on this calculation.)

Total BHCA for all endpoint devices in this case is:

(450 + 384 + 360 + 20 + 50 + 192 + 180 + 1200) = 2,836 BHCA

This level of usage is acceptable because it is below the system maximum of 3,600 BHCA, and it allows for future growth of approximately 800 BHCA.

This sizing example applies exclusively to Business Edition 5000. Business Edition 3000 is not capable of trunking to Unified Contact Center deployments. Business Edition 6000 runs Unified Contact Center Express co-resident; and although sizing considerations are similar, this example is specifically related to Business Edition 5000.

Cisco Unified Mobility for Cisco Business Edition

The capacity for Cisco Unified Mobility users on Cisco Business Edition systems depends exclusively on both the number of remote destinations per user and the BHCA of the users enabled for Unified Mobility, rather than on server hardware. Thus, the number of remote destinations supported on Cisco Business Edition depends directly on the BHCA of these users. The guidelines for sizing Cisco Unified Mobility for Cisco Business Edition are as follows:

- No more than 4 remote destinations can be configured per user. Given a maximum of 500 users per Cisco Business Edition system, the theoretical limit in terms of remote destinations is 2,000. However, given the maximum BHCA per Cisco Business Edition is 3,600, it is possible that the system might not be able to support 2,000 remote destinations. Instead BHCA calculations should be used to properly size the number of remote destinations that can be handled by the system.
- Each configured remote destination has potential BHCA implications. For every remote destination configured for a user, one additional call leg is used. Because each call consists of two call legs, one remote destination ring is equal to half (0.5) of a call. Therefore, you can use the following formula to calculate the total remote destination BHCA:

Total remote destination BHCA = 0.5 ∗ (Number of users) ∗ (Number of remote destinations per user) ∗ (User BHCA)

For example:

> Assuming a system of 300 users at 5 BHCA each, with each user having one remote destination (total of 300 remote destinations), the calculation for the total remote destination BHCA would be:
>
> > Total remote destination BHCA = 0.5 ∗ (300 users) ∗ (1 remote destination per user) ∗ (5 BHCA per user) = 750 BHCA
>
> Total user BHCA in this example is [(300 users) ∗ (5 BHCA per user)], which is 1,500 total user BHCA. By adding the total remote destination BHCA of 750 to this value, we get a total system BHCA of 2,250 (1,500 total user BHCA + 750 total remote destination BHCA).

If other applications or additional BHCA variables are in use on the system in the example above, the capacity might be limited. (See the preceding sections for further details.)

For more information on Cisco Business Edition capacity planning as well as all other Business Edition product information, refer to the following product documentation:

- Cisco Business Edition 3000

  http://www.cisco.com/en/US/products/ps11370/tsd_products_support_series_home.html
- Cisco Business Edition 5000

  http://www.cisco.com/en/US/products/ps7273/tsd_products_support_series_home.html
- Cisco Business Edition 6000

  http://docwiki.cisco.com/wiki/Cisco_Unified_Communications_Manager_Business_Edition_6000

  http://www.cisco.com/en/US/products/ps11369/tsd_products_support_series_home.html

Cisco Unified Communications Manager

Cisco Unified Communications Manager (Unified CM) is the hub of any Unified Communications deployment. It performs the most basic functions and controls endpoints, routes calls, enforces policies, hosts applications, and in general anchors other Unified Communications products such as gateways, Cisco Unity Connection, Cisco Unified MeetingPlace, Cisco Unified Contact Center suite of products, and others. These applications depend on Unified CM to function and in turn affect Unified CM's performance, which must be accounted for in Unified CM sizing.

The following factors affect Unified CM performance and must be considered when sizing a Unified CM deployment:

- Server and cluster maximum capacities
- System-level settings such as database complexity, trace level, and so forth
- Number and types of endpoints that are registered on Unified CM
- Number of users
- Traffic mix
- Dial plan
- Applications within Unified CM (Extension Mobility, WebDialer, and other CTI-enabled applications)
- Media resources hosted by the subscribers using the Cisco IP Voice Media Streaming Application

Server and Cluster Maximums

Although it is not practical to list every minute detail needed to accurately determine the number of Unified CM servers required for a particular sizing calculation, there are certain server and cluster maximums that must be observed, and some of these values change with Unified CM software version:

- Each cluster can support configuration and registration for a maximum of 40,000 secured or unsecured SCCP or SIP phones with Unified CM 8.6(1) and later releases.
- Each cluster can support configuration and registration for a maximum of 30,000 secured or unsecured SCCP or SIP phones with Unified CM 8.5 and earlier releases.
- Two TFTP servers are required if the number of endpoints in the cluster exceeds 1,250.
- Support for CTI connections has improved over the last several releases, and each cluster can support a maximum of 40,000 CTI connections.
- The number of call processing subscribers in a cluster cannot exceed 4, plus 4 standby, for a total of 8 call processing servers. Also, the total number of servers in a cluster, including the publisher, TFTP, and media servers, may not exceed 20.

The following sections describe how each of these components of Unified CM affects its sizing and therefore must be considered in an analysis of a given system description.

Deployment Options

The following deployment options are overall settings that affect all operations in the system, and they are independent of how many endpoints are registered or how many calls are in progress.

Tracing Level

The system supports two tracing levels: default and detailed. With Unified CM 9.0 and later releases, detailed tracing is enabled by default with no appreciable performance impact on CPU resources. In prior releases, detailed tracing required about 20% more CPU resources compared to the default tracing option.

Database Complexity

There is really no one measurement to determine if the database of configuration information in Unified CM should be considered as simple or complex. As a general rule, if you have more than a few thousand endpoints and more than a few hundred dial plan elements such as translation and route patterns, hunt pilots, shared lines, and so forth, then the resulting database should be considered complex. The CPU usage is considerably higher when the underlying database is complex.

Call Detail and Call Management Records

Generation of call detail records (CDR) and call management records (CMR) places a heavier burden on the CPU.

Trace Compression

Beginning with Cisco Unified CM 8.0, traces are always compressed and the compression may not be turned on or off. For earlier releases, turning on compression saves disk space but adds to CPU utilization.

Number of Regions and Locations

Configuration of regions and locations in the Unified CM cluster requires both database and static memory. The number of gateways that can be defined in the cluster is also tied to the number of locations that can be defined. Table 29-4 lists these limits for some of the Unified CM server platforms.

Table 29-4 Maximum Number of Regions, Locations, Gateways, and Trunks

Server Platform	Maximum Number of Regions	Maximum Number of Locations	Maximum Number of Trunks and Gateways
MCS-7815 and 7816	100	100	110
MCS-7825	1,000	1,000	1,100
MCS-7835 or Open Virtualization Archive (OVA) equivalent	1,000	1,000	1,100
MCS-7845 or OVA equivalent	2,000	2,000	2,100

Whether or not you can actually define the maximum number of locations and regions in a cluster depends on how "sparse" your codec matrix is. If you have too many non-default values in the inter-region codec setting, you might not be able to scale the system to its full capacity for regions and locations. As a general rule, the change from default should not exceed 10% of the maximum number.

High Availability

Deploying redundant servers increases the number of total servers required in the solution. After figuring out the minimum number of servers required for the specified deployment, add the desired number of subscriber servers. Redundancy options are described in the chapter on Unified Communications Deployment Models, page 5-1. Note that some servers do not lend themselves well to redundancy.

Number of Servers per Cluster

A cluster may be configured to consist of from one to four subscriber pairs. Reducing the number of subscriber pairs per cluster may increase the number of clusters, and hence the number of total servers, required for a given sizing analysis. An increase in the number of clusters can sometimes be desirable if the deployment consists of geographically distributed equally large locations or if any cluster-wide limit is forcing a new cluster even if the per-server utilization is low.

Choice of Servers and UCS Platforms

Unified CM is supported on a variety of Cisco Media Convergence Server (MCS) and Unified Computing System (UCS) platforms. For defining the Unified CM Virtual Machine on a UCS platform, Cisco provides Open Virtualization Archive (OVA) templates that can be loaded onto the hypervisor. Different templates specify different capacities. For example, the 10000 template defines a virtual machine with 4 virtual CPUs, 6 GB of RAM, and 160 GB of hard disk space that has a maximum capacity of serving up to 10,000 endpoints. There are similar templates defined for 1000, 2500, and 7500 endpoints as well.

The formal definitions of the OVA templates for Unified CM and other Unified Communication products are available at

> http://docwiki.cisco.com/wiki/Unified_Communications_Virtualization_Downloads_(including_OVA/OVF_Templates)

Note Choice of placement of virtual machines running Unified CM and other Unified Communications products can have an impact on performance and availability. For a discussion of these and other considerations for Unified Communications on UCS deployments, refer to the documentation at http://www.cisco.com/go/uc-virtualized.

Endpoints

The type and number of endpoints are an important part of the net load that the system must support. There are different types of endpoints, and each type imposes a different load on the Unified CM. Endpoints can be differentiated by:

- Digital (IP) or analog (using an adaptor)
- Software-based or hardware
- The protocol they support (SIP or SCCP)
- Whether they are configured with security
- Dialing modes (en-bloc or overlap)
- Audio only or both audio and video
- Other devices such as gateways (H.323 or MGCP)

Each type of endpoint defined in the system uses system resources – for example, static memory just by being defined and registered, and CPU and dynamic memory based on its call rate. Each endpoint could also place additional load on Unified CM by running applications interacting with services running inside of Unified CM.

There are defined maximum supported quantities of endpoints for a given server, as shown in Table 29-5. Note that these values are guidelines only, and it is possible that the system may not be able to support these maximum amounts because of other applications running in the deployment.

Table 29-5 Maximum Number of Endpoints Per Server Platform or OVA Template

Server Platform Characteristics	Maximum Endpoints per Server or OVA Template	High-Availability Server
Cisco MCS 7845-I3 or OVA equivalent	10,000	Yes
Cisco MCS 7845 (All other supported models) or OVA equivalent	7,500	Yes
Cisco MCS 7835 (All supported models) or OVA equivalent	2,500	Yes
Cisco MCS 7825 (All supported models) or OVA equivalent	1,000	No
Cisco MCS 7816 (All supported models)	500	No
Cisco MCS 7815 (All supported models)	300	No

The designation of High Availability in Table 29-5 indicates whether those servers may be paired for high availability within a cluster consisting of those servers.

Some endpoints may operate in one of two modes. Endpoints such as Cisco Unified Personal Communicator and those based on Common Services Framework (CSF), such as Cisco WebEx Connect, Cisco UC Integration^{TM} for Microsoft Lync, and others, can work either as soft-phones registered directly with Unified CM as phones, or in desk-phone control mode where they act as applications that use CTI to communicate with Unified CM to control a desk phone. Either way, they use Unified CM resources (endpoints or CTI applications) but count against different operating limits.

Along with the endpoints, the number of busy hour users must also be taken into account. The number of users and their collective usage of the endpoints determine the call processing load on the system.

Cisco Collaboration Clients and Applications

Cisco Collaboration Clients include the following software applications that run on the user's desktops or other access devices:

- Cisco Unified Personal Communicator, page 29-24
- Cisco WebEx Connect, page 29-25
- Cisco UC Integration^{TM} for Microsoft Lync, page 29-26
- Cisco Unified Mobile Communicator, page 29-26
- Third-Party XMPP Clients and Applications, page 29-26

In addition, the following client provides an integrated telephony client with virtualized desktop access:

- Cisco Virtualization Experience Clients, page 29-27

Cisco Unified Client Services Framework

The Cisco Unified Client Services Framework provides the underlying services layer for several of these clients, including Cisco Unified Personal Communicator, Cisco WebEx Connect, and Cisco UC Integration for Microsoft Lync. The framework can operate in one of two modes, each of which uses different resources with Unified CM. The Client Services Framework may be configured to operate as a softphone. In this mode, it acts as a SIP registered endpoint and counts toward the total number of endpoints in the system. The Client Services Framework may alternatively be configured to control the user's desk phone, in which case it uses CTI resources. Moreover, the user may switch these Framework-based clients to work in either mode. Therefore, it is necessary to properly account for the system resources necessary for the anticipated usage.

The following additional items must be considered for a Client Services Framework deployment:

- TFTP — When configured in softphone (audio on computer) mode, a Client Services Framework device configuration file is downloaded to the client for Unified CM call control configuration information. In addition, any application dial rules or directory lookup rules are also downloaded through TFTP.
- CTI — When configured in deskphone (using desk phone for audio) mode, the Client Services Framework establishes a CTI connection to Unified CM upon login and registration to allow for control of the IP phone.
- CCMCIP — The Client Services Framework uses the Unified CM IP Phone Services to gather information about the devices associated with the user in order to list the IP phones available for control.
- IMAP — When configured for voicemail, the Client Services Framework updates and retrieves voicemail through an IMAP connection to the mailstore.
- LDAP — Client login and authentication, contact profile information, and incoming caller identification are all handled through an LDAP query, unless stored in the local Client Services Framework cache.

With the exception of IP Phone Services for the integrated Extension Mobility and Unified CM Assistant applications, IP Phone Services must reside on a separate web server. Running phone services other than Extension Mobility and Unified CM Assistant on the Unified CM server is not supported.

Cisco Unified Personal Communicator

When designing and sizing a solution for Cisco Unified Personal Communicator, you must consider the following scalability impacts for all the components:

- Client scalability

  The Cisco IM and Presence Service hardware deployment determines the number of users a cluster can support. The Cisco Unified Personal Communicator deployment must balance all users equally across all servers in the cluster. This can be done automatically by setting the User Assignment Mode Sync Agent service parameter to **balanced**. The maximum number of contacts in the contact list is 200.

- IMAP scalability

  The number of IMAP or IMAP-Idle connections is determined by the platform overlay (Cisco Unity or Cisco Unity Connection) for messaging integration. For specific configuration sizing, refer to the Cisco Unity or Cisco Unity Connection product documentation available at http://www.cisco.com.

- Web conferencing

  Cisco Unified MeetingPlace web licensing determines the number of concurrent web conferencing participants allowed. For specific configuration sizing, refer to the Cisco Unified MeetingPlace product documentation available at http://www.cisco.com.

- Video sizing capability

  Videoconferencing and switching are determined by Cisco Unified Videoconferencing MCU sizing and configuration, by Cisco MeetingPlace Hardware Media Server (HMS) sizing and configuration, or by Cisco Unified MeetingPlace Express VT for concurrent voice, video, and web participants. For specific configuration sizing, refer to the Cisco Unified Videoconferencing or Cisco Unified MeetingPlace Express VT product documentation available at http://www.cisco.com.

Cisco Unified Personal Communicator interfaces with Unified CM. Therefore, the following guidelines for the current functionality of Unified CM apply when Cisco Unified Personal Communicator voice or video calls are initiated:

- CTI scalability

  In Desk Phone mode, calls from Cisco Unified Personal Communicator use the CTI interface on Unified CM. Therefore, observe the CTI limits as defined in the chapter on Call Processing, page 8-1. You must include these CTI devices when sizing Unified CM clusters.

- Call admission control

  Cisco Unified Personal Communicator applies call admission control for voice and video calls by means of Unified CM locations or RSVP.

- Codec selection

  Cisco Unified Personal Communicator voice and video calls utilize codec selection through the Unified CM regions configurations.

All Cisco Unified Personal Communicator configuration and contacts are stored in the Cisco IM and Presence database and have the potential to contain large amounts of data. The current conversation history list is limited to 50 entries for each tab (Chats, Voice Messages, Calls), while the contact list size is limited to 200 contacts. Therefore, bandwidth utilization must be taken into consideration for presence data exchange as well as for conferencing, video, and messaging traffic.

The following bandwidth considerations also apply to Cisco Unified Personal Communicator:

- A Presence User Profile (PUP) takes into consideration the number of logins, presence state changes, and roster changes to determine a user deployment traffic pattern. With a typical PUP, where the number of logins is 0.5 per hour, the number of presence state changes is 0.5 per hour, and the number of roster changes is 0.25 per hour, you can use the following formula as a general guideline for calculating bandwidth utilization (in kilobits per second) between Cisco Unified Personal Communicator and Cisco IM and Presence (see Table 29-6 for examples):

  USERS ∗ [30 + ROSTER∗7 + IM∗3 + CALLS ∗ (33 + 3∗ROSTER)] / 1000

  where:

  USERS = number of users using Unified Personal Communicator.

  ROSTER = average roster size of a Unified Personal Communicator user.

  IM = number of instant messages per hour for a Unified Personal Communicator user.

  CALLS = number of softphone calls per hour.

Table 29-6 Examples of Bandwidth Requirement for Unified Personal Communicator

Enterprise	Number of Users	Roster Size	Number of IMs	Calls per Hour	Bandwidth Utilization
Small	1,000	100	25	4	2,100 kbps (2.1 Mbps)
Large	5,000	200	25	4	20,185 kbps (20.2 Mbps)

- For Cisco Unified MeetingPlace voice, video, and web collaboration sessions, see Cisco Unified MeetingPlace, page 22-21.
- For video calls, see the chapter on IP Video Telephony, page 12-1.
- For Cisco Unity or Unity Connection, see the section on Managing Bandwidth, page 21-28, in the chapter on Cisco Voice Messaging, page 21-1.

Cisco WebEx Connect

A single end-user requires only a 56 kbps dial-up Internet connection to be able to log in to the Cisco WebEx Messenger service (formerly WebEx Connect service) and get the basic capabilities such as presence, instant messaging, and VoIP calling. However, for a small office or branch office, a broadband connection with a minimum of 512 kbps is required in order to use the advanced features such as file transfer, screen capture, PC-to-PC video calling, and team spaces. For higher quality video such as High Definition 720p, the minimum bandwidth connection recommendation is 2 Mbps.

For additional information on network and desktop requirements, refer to the Cisco WebEx administrator's guide available at

http://www.webex.com/webexconnect/orgadmin/help/index.htm

The Cisco Unified Communications integrations use Unified CM CTI Manager for click-to-call applications, as well as deskphone control mode with the Cisco Unified Client Services Framework. Therefore, observe the CTI limits as defined in the section on Applications and CTI, page 29-30. When Cisco UC Integration^{TM} for Connect is operating in a softphone (audio on computer) mode, the Cisco Unified Client Services Framework is a SIP registered endpoint with Cisco Unified CM. When sizing a solution involving Cisco Unified Communications, you must include the CTI devices and the SIP endpoint devices utilizing resources on the Unified CM clusters.

Network Requirements

Cisco Webex Messenger service deployment network requirements are available at:

http://www.webex.com/webexconnect/orgadmin/help/17161.htm

Cisco UC Integration^{TM} for Microsoft Lync

Cisco UC Integration^{TM} for Microsoft Lync uses Unified CM CTI Manager for click-to-dial applications, as well as deskphone control mode with the Cisco Unified Client Services Framework. Therefore, observe the CTI limits as defined in the chapter on Call Processing, page 8-1. When Cisco UC Integration^{TM} for Microsoft Lync is operating in a softphone (audio on computer) mode, the Cisco Unified Client Services Framework is a SIP registered endpoint with Cisco Unified CM. When sizing a solution involving Cisco Unified Communications, you must include the CTI devices and the SIP endpoint devices utilizing resources on the Unified CM clusters.

Cisco Unified Mobile Communicator

The Cisco Unified Mobility Advantage Server supports the following user capacities:

- Cisco MCS 7845-H2/I2 supports up to 1,000 Unified Mobile Communicator clients.
- Cisco MCS 7825-H4/I4 supports up to 500 Unified Mobile Communicator clients.
- Cisco MCS 7825-H2/I2 or 7825-H3/I3 supports up to 250 Unified Mobile Communicator clients.

To support more than 1,000 Unified Mobile Communicator users within a deployment, additional Unified Mobility Advantage Servers may be installed. However, Unified Mobile Communicator clients configured and associated to one Cisco Unified Mobility Advantage Server will not be able to send text messages to clients on another server.

When integrating Unified Mobile Communicator with Cisco Unified CM for enterprise call log integration, the Unified Mobility Advantage Server interacts with Unified CM CTIManager for deskphone line monitoring. For each Unified Mobile Communicator enabled for call log integration, the Cisco Unified Mobility Advantage Server generates one CTI connection to the CTIManager. Therefore, with a deployment of Unified Mobile Communicator with one fully populated Unified Mobility Advantage Server running on an MCS 7845 with call log integration enabled for all users, 1,000 CTI connections will be consumed. For this reason, when you deploy Unified Mobile Communicator with call log integration, you must consider the number of required CTI connections as explained in the section on Applications and CTI, page 29-30.

If additional CTI connections are required for other applications, they can limit the capacity of Unified Mobile Communicator users with call log integration enabled.

Integration of Unified Mobile Communicator with Unified CM for dial-via-office and Unified Mobility functionality requires the configuration of each Unified Mobile Communicator as a Unified CM device and configuration of the mobile number as a mobility identity. Therefore, when implementing these integrations, you must also consider overall Unified CM phone and mobility-enabled user capacities.

Third-Party XMPP Clients and Applications

Third-party Extensible Messaging and Presence Protocol (XMPP) clients may be used with both the WebEx Messenger service platform and the Cisco IM and Presence Service. Voice, video, and other collaboration mechanisms (except for instant messaging and chat) are typically not supported with these clients. Depending on their capabilities, these clients may be counted against the device capacities supported by the above products on their servers.

Cisco Virtualization Experience Clients

All Cisco Virtualization Experience Clients are deployed with a Virtual Desktop Infrastructure (VDI) component, while some of the deployments may also contain a Unified Communications component. Capacity planning and datacenter resource utilization for VDI when using the Cisco Virtualization Experience Clients is covered as part of the Virtualization Experience Infrastructure (VXI) sizing. For details, refer to the VXI documentation available at

http://www.cisco.com/en/US/solutions/ns340/ns414/ns742/ns1100/landing_vxi.html

Capacity planning for the Unified Communications components depends on which Virtualization Experience Client is deployed:

- Cisco VXC 2111 and 2112 integrated form factor zero clients are paired with a Cisco Unified IP Phone 8961, 9951, or 9971. The Cisco client running in the user's virtual desktop uses the deskphone control mode of the Cisco Unified IP Phone; therefore, Computer Telephony Integration (CTI) planning guidelines must be followed for each client deployed.
- Cisco VXC 2211 and 2212 standalone form factor zero clients can be deployed as VDI-only or as a fully integrated voice, video, and virtual desktop with a number of different Cisco Unified IP Phones. When deployed in a Unified Communications environment, the Cisco client running in the user's virtual desktop uses the deskphone control mode of the Cisco Unified IP Phone; therefore, CTI planning guidelines must be followed for each client deployed.
- Cisco VXC 4000 software appliance is a software-only VXC deployment option. The Cisco client running in the user's virtual desktop uses the deskphone control mode of the VXC 4000; therefore, CTI planning guidelines must be followed for each VXC 4000 deployed.
- Cisco VXC 6215 thin client running in VDI-only mode follows VDI capacity planning; however, when the VXC 6215 is deployed as a fully integrated voice, video, and virtual desktop, additional Unified Communications capacity must be accounted for. The Cisco client running in the user's virtual desktop uses the deskphone control mode of the VXC software appliance running locally on the Linux thin client; therefore, CTI planning guidelines must be followed for each client deployed. The VXC software appliance is a SIP line-side registered device on Cisco Unified CM; therefore, for each VXC 6215 thin client running as a fully integrated voice, video, and virtual desktop, a SIP line device and CTI connection is used.

Mobile Unified Communications

Mobility in Unified Communications is multi-faceted. Each of the different aspects of mobile communications consumes different Unified CM resources and must be accounted for both independently and as a part of the whole system. The following sizing considerations apply to mobility, but note that aspects of mobility that do not affect Unified CM are not discussed here.

Cisco Unified Mobility

There are two parameters that are key to Unified CM's capacity to support Mobile Access and Enterprise Feature Access. For these functions to work appropriately, users must be enabled for mobility and remote destinations with shared lines must be defined for the users. Table 29-7 shows the limits for users and remote destinations in a cluster consisting of each class of Unified CM servers.

Table 29-7 Maximum Number of Mobility Users and Remote Destination per Cluster

Cluster Servers	Maximum Number of Users Enabled for Mobility per Cluster	Maximum Number of Remote Destinations per Cluster
MCS-7845 or OVA equivalent	15,000	15,000 (or 3,750 per server)

Table 29-7 Maximum Number of Mobility Users and Remote Destination per Cluster (continued)

Cluster Servers	Maximum Number of Users Enabled for Mobility per Cluster	Maximum Number of Remote Destinations per Cluster
MCS-7835 or OVA equivalent	10,000	10,000 (or 2,500 per server)
MCS-7825 or OVA equivalent	4,000	4,000 (or 1,000 per server)

Note A mobility-enabled user is defined as a user that has a remote destination profile and at least one remote destination or a mobility identity configured.

Each remote destination defined in the system affects Unified CM in several ways:

- The remote destination occupies static memory and configuration space in the database.
- Each occurrence uses a shared line with the users primary device and hence calls to that line use more CPU resources.
- If the remote destination is an external number (such as the user's cell phone or home), then gateway resources will be used to extend the call.

Call Traffic

Next to the number of endpoints, the quantity and quality of call traffic places the second biggest requirement on Unified CM. It is important to differentiate between call types because call origination and termination are considered as distinct events in the half-call model. A single server needs to handle both halves for calls made between two endpoints registered on it. For calls made between two servers in the same cluster, each of the participating servers needs to handle only half of the call. For calls made between endpoints registered on different clusters, a server and the cluster as a whole need to handle only half of each call. For calls made between an endpoint in a cluster and the PSTN, a PSTN gateway needs to handle half of the call, and these calls form the basis for sizing the gateways themselves.

When considering call traffic, other complexities arise from calls between endpoints that work on different protocols, such as between SIP and SCCP-based phones, if calls are transferred, and if conferencing is invoked.

In general, the following factors require consideration:

- Overall Busy Hour Call Attempts (BHCA) per user
- Average Call Holding Time (ACHT) per call
- BHCA from and to the PSTN using MGCP, H.323, and SIP protocols
- BHCA from and to other clusters using H.323 intercluster trunks or SIP protocols
- BHCA from and to other enterprises using Cisco Intercompany Media Engine (IME)
- BHCA within the cluster

Each different type of call takes a different amount of CPU resources to set up. The rate of call placement, or the BHCA, determines the CPU usage. CPU requirements vary directly with the call placement rate. The ACHT determines the dynamic memory requirements to sustain calls for their duration. A higher ACHT means that more dynamic memory must remain allocated, thus increasing the memory requirement.

Call traffic can arise from other sources as well. Each time a call is redirected in a transfer or to voicemail, it requires processing by the CPU. If a directory number is configured on multiple phones, an incoming call to that number needs to be presented to all of those phones, thus increasing CPU usage at call setup time. As another example, if advanced features such as the Intercompany Media Engine (IME) are being used, calls made using this technology, and the percentage of these calls that need to be redirected to the PSTN because of call quality, must also be accounted for.

Dial Plan

The dial plan in Unified CM consists of static configuration elements that determine call routing and associated policies. In general, dial plan elements occupy static memory space in Unified CM servers, and the following dial plan elements impact the amount of memory required:

- Directory numbers
- Shared directory numbers and the average number of endpoints that share the same DN
- Partitions, calling search spaces, and translation patterns
- Route patterns, route lists, and route groups
- Advertised and learned DN patterns
- Hunt pilots and hunt lists
- Circular, sequential, and broadcast line groups and their membership

There are no hard limits enforced by Unified CM for any of the dial plan elements, but there is only a limited amount of shared system memory available.

Of the above dial plan elements, the number of lines shared across multiple endpoints is of particular interest. Each shared line multiplies the CPU cost of a call setup because the call has to be presented to all the endpoints that share that particular directory number.

Another aspect of a large dial plan that comes into play is the space required to hold the elements of the plan in the Informix Database System. There is only a finite amount of disk space available to hold the entire configuration of Unified CM, and extra-large dial plans can overwhelm it. In this case, the only option may be to break up the dial plan and use its parts in multiple clusters.

Applications and CTI

In the context of Unified CM, applications are the "extra" functions beyond simple call processing provided by Unified CM. In general these applications make use of Computer Telephone Integration (CTI), which allows users to initiate, terminate, reroute, or otherwise monitor and treat calls. Features such as Cisco Unified CM Assistant, Attendant Console, Contact Center, and others, depend on CTI to function.

Historically CTI interactions have been relatively expensive operations in Unified CM that severely limited system scalability, but recent optimizations have reduced their impact on scalability. Although the high-end server platforms for Unified CM 9.*x* are able to support CTI for all of their registered devices, the lower-end platforms do not scale that high. Table 29-8 lists the maximum number of CTI resources supported by each type of server platform. These maximum values apply to the following types of CTI resources:

- The maximum number of CTI controlled and/or monitored endpoints that can be registered to a Unified CM subscriber node.
- The maximum number of endpoints that a Unified CM subscriber node running the CTI Manager service can monitor or control.
- The maximum number of TAPI/JTAPI application instances that can connect to a Unified CM subscriber node running the CTI Manager service. The TAPI/JTAPI application instances that can connect to a Unified CM subscriber node running the CTI Manager service are sometimes referred as CTI connections.

Note that the numbers for the high end of each class of server equal the number of devices that the class can support.

In addition to native applications provided by Unified CM, third-party applications may also be deployed that use Unified CM CTI resources. When counting CTI ports and route points, be sure to account for the third-party applications as well.

Table 29-8 CTI Resource Limits in Unified CM

Server Platform	Maximum CTI Resources per Server
MCS 7815	150
MCS 7816-I2/I3/I4	400
MCS 7816-I5	500 with Unified CM 8.6 and later releases; otherwise 400
MCS 7825-I1/I2	800
MCS 7825-I3/I4	900
MCS 7825-I5 and OVA equivalent	1,000 with Unified CM 8.6 and later releases; otherwise 900
MCS 7835-I1/I2	2,000
MCS 7835-I3 and OVA equivalent	2,500 with Unified CM 8.6 and later releases; otherwise 2,000
MCS 7845-I1	2,500
MCS 7845-I2 and OVA equivalent	5,000
MCS 7845-I3 and OVA equivalent	10,000 with Unified CM 8.6 and later releases; otherwise 5,000

In addition to the maximum number of connections and devices, CTI limits are also influenced by:

- The number of lines on each of the controlled devices (up to 5 lines per controlled device with Unified CM 8.6 and later releases; otherwise up to 2 lines per controller device)
- The number of shared occurrences of a line controlled by CTI (up to 5 per line with Unified CM 8.6 and later releases; otherwise up to 2 per line)
- The number of active CTI applications (up to 5 for any device with Unified CM 8.6 and later releases; otherwise up to 2 for any device)
- A maximum of 6 BHCA per controlled device

The CTI resources available on Unified CM are reduced if any of these values is exceeded.

Table 29-9 lists the number of supported CTI devices for Cisco Business Edition.

Table 29-9 Users and CTI Devices in Cisco Business Edition

Model	Maximum Number of Users	Maximum CTI Devices
Business Edition 3000	300	400
Business Edition 5000	500	575
Business Edition 6000	1,000	1,200

Determining CTI Resources Required for a Unified CM Cluster

Step 1 Determine the total CTI device count.

Count the number of CTI devices that will be in use on the cluster.

Step 2 Determine the CTI line factor.

Determine the CTI line factor of all devices in the cluster, according to Table 29-10.

Table 29-10 CTI Line Factor

Number of Lines per CTI Device	CTI Line Factor
1 to 5 lines	1.0
6 lines	1.2
7 lines	1.4
8 lines	1.6
9 lines	1.8
10 lines	2.0

Note If there are multiple line factors for the devices within a cluster; determine the average line factor across all CTI devices in the system.

Step 3 Determine the application factor.

Determine the application factor of all devices in the cluster, according to Table 29-11.

Table 29-11 CTI Application Factor

Number of Applications per CTI Device	CTI Application Factor
1 to 5 applications	1.0
6 applications	1.2
7 applications	1.4
8 applications	1.6
9 applications	1.8
10 applications	2.0

Step 4 Calculate the required number of CTI resources according to the following formula:

Required Number of CTI Resources = (Total CTI Device Count) ∗ (The greater of the CTI Line Factor or the CTI Application Factor)

The following examples illustrate the process.

Example 1: 500 CTI devices deployed with an average of 9 lines per device and an average of 4 applications per device. According to the factor lists in Table 29-10 and Table 29-11, 9 lines per device renders a line factor of 1.8, while 4 applications per device renders an application factor of 1.0. Applying these values in the formula from Step 4 yields:

(500 CTI Devices) ∗ (Greater of {1.8 Line Factor or 1.0 Application Factor})

(500 CTI Devices) ∗ (1.8 Line Factor) = 900 total CTI resources required

Example 2: 2,000 CTI devices deployed with an average of 5 lines per device and an average of 9 applications per device. According to the factor lists in Table 29-10 and Table 29-11, 5 lines per device renders a line factor of 1.0, while 9 applications per device renders an application factor of 1.8. Applying these values in the formula from Step 4 yields:

(2000 CTI Devices) ∗ (Greater of {1.0 Line Factor or 1.8 Application Factor})

(2000 CTI Devices) ∗ (1.8 Application Factor) = 3,600 total CTI resources required

Example 3: 5,000 CTI devices deployed with an average of 2 lines per device and an average of 3 applications per device. According to the factor lists in Table 29-10 and Table 29-11, 2 lines per device renders a line factor of 1, while 3 applications per device renders an application factor of 1. Applying these values in the formula from Step 4 yields:

(5,000 CTI Devices) ∗ (Greater of {1 Line Factor or 1 Application Factor})

(5,000 CTI Devices) ∗ (1 Line or Application Factor) = 5,000 total CTI resources required

Cisco Extension Mobility and Extension Mobility Cross Cluster

Using Extension Mobility (EM) impacts the system performance in the following ways:

- Creation of EM profiles requires both disk database space and static memory.
- The rate at which users may log into their EM accounts affects both CPU and memory usage. Servers have bounds on the maximum number of logins per minute that they can support.

- Extension Mobility Cross Cluster (EMCC) has a higher impact on resources. There is a limit on the number of EMCC users that a server can support. The maximum EMCC login rates supported are lower than those supported for EM. In addition, there is a trade-off between EM and EMCC login rates. If both are occurring at the same time, then the maximum capacity for each will be reduced.
- EM and EMCC login rates per cluster are not simply the login rate of each server multiplied by the number of servers in the cluster because profiles in a shared database have to be accessed. The maximum login rate in a cluster consisting of more than one call processing subscriber should be limited to 1.5 times that of a single server.

Table 29-12 shows the maximum number of EM and EMCC logins per minute for each type of server.

Table 29-12 EM and EMCC Rates Per Server Type

Server Types	Maximum EM Login Rate (per Server)	Maximum EM Login Rate (Dual Servers)	Maximum EMCC Login Rate (Per Server)	Maximum EMCC Login Rate (Dual Servers)	Maximum Concurrent EMCC Devices
MCS-7815, MCS-7816	15	22	5	7	100 (MCS-7815) or 167 (MCS-7816)
MCS-7825 and OVA equivalent	200	300	60	70	333
MCS-7835 (I2/H2, I3/H3) and OVA equivalent	235	352	71	80	833
MCS-7845 and OVA equivalent	250	375	75	90	2,500

Cisco Extension Mobility login and logout functionality can be distributed across a pair of subscriber nodes to increase login/logout cluster capacity. When the EM load is distributed evenly between two MCS 7845-H2/I2 servers, the maximum cluster-wide capacity is 375 sequential logins and/or logouts per minute.

Note The Cisco Extension Mobility service can be activated on more than two nodes for redundancy purposes, but Cisco supports a maximum of two subscriber nodes actively handling logins/logouts at any given time.

Note Enabling EM Security does not diminish performance.

The EMCC login/logout process requires more processing resources than intracluster EM login/logout, therefore the maximum supported login/logout rates are lower for EMCC. In the absence of any intracluster EM logins/logouts, Unified CM 8.*x* supports a maximum rate of 75 EMCC logins/logouts per minute with Cisco MCS 7845-H2/I2 and MCS 7845-I3 servers. Most deployments will have a combination of intracluster and intercluster logins/logouts occurring. For this more common scenario, the mix of EMCC logins/logouts (whether acting as home cluster or visiting cluster) should be modeled for 40 per minute, while the intracluster EM logins should modeled for 185 logins/logouts when using a single EM login server. The intracluster EM login rate can be increased to 280 logins/logouts per minute when using MCS 7845-H2/I2 or MCS 7845-I3 servers in dual EM server configuration. (See Table 29-12.)

EMCC logged-in devices (visiting phones) consume twice as many resources as any other endpoint in a cluster. The maximum supported number of EMCC logged-in devices is 2,500 per cluster, but this also decreases the theoretical maximum number of other devices per cluster from 30,000 to 25,000. Even if the number of other registered devices in the cluster is reduced, the maximum supported number of EMCC logged-in devices is still 2,500.

Cisco Unified CM Assistant

The Cisco Unified CM Assistant application uses CTI resources in Unified CM for line monitoring and phone control. Each line (including intercom lines) on a Unified CM Assistant or Manager phone requires a CTI line from the CTIManager. In addition, each Unified CM Assistant route point requires a CTI line instance from the CTIManager. When you configure Unified CM Assistant, the number of required CTI lines or connections must be considered with regard to the overall cluster limit for CTI lines or connections.

The following limits apply to Unified CM Assistant:

- A maximum of 10 Assistants can be configured per Manager.
- A maximum of 33 Managers can be configured for a single Assistant (if each Manager has one Unified CM Assistant-controlled line).
- A maximum of 3,500 Assistants and 3,500 Managers (7,000 total users) can be configured per cluster using the Cisco MCS 7845 server.
- A maximum of three pairs of primary and backup Unified CM Assistant servers can be deployed per cluster if the **Enable Multiple Active Mode** advanced service parameter is set to **True** and a second and third pool of Unified CM Assistant servers are configured.

In order to achieve the maximum Unified CM Assistant user capacity of 3,500 Managers and 3,500 Assistants (7,000 users total), multiple Unified CM Assistant server pools must be defined. (For more information, see Unified CM Assistant, page 19-20.)

Cisco WebDialer

Cisco WebDialer provides a convenient way for users to initiate a call. Its impact on Unified CM is fairly limited because extra resources are required only at call initiation and are not tied up for the duration of the call. Once the call has been established, its impact on Unified CM is just like any other call.

The WebDialer and Redirector services can run on one or more subscriber nodes within a Unified CM cluster, and they support the following capacities:

- Each WebDialer service can handle up to 2 call requests per second (7,200 calls per hour) per node.
- Each Redirector service can handle up to 8 call requests per second.

The following general formula can be used to determine the number of WebDialer calls per second (cps):

(Number of WebDialer users) ∗ ((Average BHCA) / (3600 seconds/hour))

When performing this calculation, it is important to estimate properly the number of BHCA per user that will be initiated specifically from using the WebDialer service. The following example illustrates the use of these WebDialer design calculations for a sample organization.

Example: Calculating WebDialer Calls per Second

Company XYZ wishes to enable click-to-call applications using the WebDialer service, and their preliminary traffic analysis resulted in the following information:

- 10,000 users will be enabled for click-to-call functionality.
- Each user averages 6 BHCA.

- 50% of all calls are dialed outbound, and 50% are received inbound.
- Projections estimate 30% of all outbound calls will be initiated using the WebDialer service.

Note These values are just examples used to illustrate a WebDialer deployment sizing exercise. User dialing characteristics vary widely from organization to organization.

10,000 users each with 6 BHCA equates to a total of 60,000 BHCA. However, WebDialer deployment sizing calculations must account for placed calls only. Given the initial information for this sizing example, we know that 50% of the total BHCA is for placed or outbound calls. This results in a total of 30,000 placed BHCA for all the users enabled for click-to-call using WebDialer.

Of these placed calls, the percentage that will be initiated using the WebDialer service will vary from organization to organization. For the organization in this example, several click-to-call applications are made available to the users, and it is projected that 30% of all placed calls will be initiated using WebDialer.

(30,000 placed BHCA) ∗ 0.30 = 9,000 placed BHCA using WebDialer

To determine the number of WebDialer servers required to support a load of 9,000 BHCA, we convert this value to the average call attempts per second required to sustain this busy hour:

(9,000 call attempts / hour) ∗ (hour/3,600 seconds) = 2.5 cps

Each WebDialer service can support up to 2 cps, therefore 2 nodes should be configured to run the WebDialer service in this example. This would allow for future growth of WebDialer usage. In order to maintain WebDialer capacity during a server failure, additional backup WebDialer servers should be deployed to provide redundancy.

Attendant Console

The integration of Cisco Unified CM with the Cisco Unified Department, Unified Business, and Unified Enterprise Attendant Consoles centers on their CTI resource usage. These applications monitor the last 2,000 users to whom the attendant sent calls, thus increasing CTI resource usage. In addition, each call uses a number of CTI route points and ports for greetings, queuing, and so forth.

Media Resources

The Unified CM server, by the virtue of the Cisco IP Voice Media Streaming Application, may be used for certain media functions that can be performed in software only and do not require hardware resources. Unified CM can act as a media termination point (MTP), as a conference bridge, or as a source of music-on-hold streams. Although the capabilities of Unified CM are limited in comparison to similar functions provided by Cisco Integrated Service Routers (ISRs), they are generally the key source of music-on-hold streams (both unicast and multicast).

The Cisco IP Voice Media Streaming Application may be deployed in one of two ways:

- Co-resident deployment

  In a co-resident deployment, the streaming application runs on any server (either publisher or subscriber) in the cluster that is also running the Unified CM software.

  Note The term *co-resident* refers to two or more services or applications running on the same server.

- Standalone deployment

  A standalone deployment runs the streaming application on a dedicated server within the Unified CM cluster. That is, the Cisco IP Voice Media Streaming Application service is the only service enabled on the server. The sole function of this dedicated server is to provide media resources to devices within the network.

While the Cisco IP Voice Media Streaming Application provides MTP, annunciation, and conferencing capabilities, you might find it more scalable to place this functionality on external Cisco Integrated Service Routers (ISRs). The music-on-hold functionality of this application is, however, not so easily placed on external sources. Table 29-13 lists the maximum values that may be configured for each of these services.

Table 29-13 Cisco IP Voice Media Streaming Application Capacity Limits

Service	Maximum Number of Streams
Annunciator	400
Conference Bridge	256
Media Termination Point	512

Note To calculate the capacities of each of the media functions on the DSPs supported by each individual ISR, refer to the Cisco ISR product data sheets or to the chapter on Media Resources, page 17-1.

Music on Hold

Table 29-14 lists the server platforms and the maximum number of simultaneous music-on-hold (MoH) sessions each can support. You should ensure that the actual usage does not exceed these limits because, once MoH sessions have reached these limits, additional load could result in poor MoH quality, erratic MoH operation, or even loss of MoH functionality.

Table 29-14 Music on Hold Capacity Limits

Server Platform	Codecs Supported	Maximum Number of MoH Sessions
MCS 7816 MCS 7825 MCS 7878 and OVA equivalent	G.711 (A-law and mu-law) G.729a Wideband audio	Co-resident or standalone server: 250 MoH sessions
MCS 7835 MCS 7845 and OVA equivalent	G.711 (A-law and mu-law) G.729a Wideband audio	Co-resident or standalone server: 500 MoH sessions

You can define a maximum of 51 unique sources of Music on Hold on a Unified CM cluster. Considering that each MoH source may be streamed in up to four encodings, there can be a maximum of 204 multicast streams in the cluster. The limits described in Table 29-14 apply to any combination of unicast, multicast, or simultaneous unicast and multicast sessions.

Impact on Unified CM

Whether deployed in co-resident or standalone mode, the Cisco IP Voice Media Streaming Application consumes CPU and memory resources. This impact must be considered in the overall sizing of Unified CM. In general, usage of media resources can be considered to add to the BHCA that needs to be processed by Unified CM.

LDAP Directory Integration

The Unified CM Database Synchronization feature provides a mechanism for importing a subset of the user configuration data (attributes) from the LDAP store into the Unified CM publisher database. Once synchronization of a user account has occurred, the copy of each user's LDAP account information may then be associated to additional data required to enable specific Unified Communications features for that user. When authentication is also enabled, the user's credentials are used to bind to the LDAP store for password verification. The end user's password is never stored in the Unified CM database when enabled for synchronization and/or authentication.

User account information is cluster-specific. Each Unified CM publisher server maintains a unique list of those users receiving Unified Communications services from that cluster. Synchronization agreements are cluster-specific, and each publisher has its own unique copy of user account information.

The maximum number of users that a Unified CM cluster can handle is limited by the maximum size of the internal configuration database that gets replicated between the cluster members. Starting with Unified CM Release 8.6(1), the maximum number of users that can be configured or synchronized was increased from 60,000 to 80,000. To optimize directory synchronization performance, Cisco recommends considering the following points:

- Directory lookup from phones and web pages may use the Unified CM database or the IP Phone Service SDK. When directory lookup functionality uses the Unified CM database, only users who were configured or synchronized from the LDAP store are shown in the directory. If a subset of users are synchronized, then only that subset of users are seen on directory lookup.
- When the IP Phone Services SDK is used for directory lookup, but authentication of Unified CM users to LDAP is needed, the synchronization can be limited to the subset of users who would log in to the Unified CM cluster.
- If only one cluster exists, if the LDAP store contains fewer than the maximum number of users supported by the Unified CM cluster, and if directory lookup is implemented to the Unified CM database, then it is possible to import the entire LDAP directory.
- If multiple clusters exist and if the number of users in LDAP is less than the maximum number of users supported by the Unified CM cluster, it is possible to import all users into every cluster to ensure directory lookup has all the entries.
- If the number of user accounts in LDAP exceeds the maximum number of users supported by the Unified CM cluster and if the entire user set should be visible to all users, it will be necessary to use the Unified IP Phone Services SDK to off-load the directory lookup from Unified CM.
- If both synchronization and authentication are enabled, user accounts that have either been configured or synchronized into the Unified CM database will be able to log in to that cluster. The decision about which users to synchronize will impact the decision on directory lookup support.

Note Cisco supports the synchronization of user accounts up to the limit mentioned above, but it does not enforce this limit. Synchronizing more user accounts can lead to starvation of disk space, slower database performance, and longer upgrade times.

Cisco Unified CM Megacluster Deployment

A Unified CM cluster is considered to be a megacluster when the number of call processing subscribers exceeds the normal maximum of 4 pairs. A megacluster may have up to 8 pairs of call processing subscribers and no more than 21 servers in all.

A Unified Communications deployment can be simplified in certain cases with a Unified CM megacluster. The following limits increase with such a deployment:

- Maximum number of endpoints supported is now twice the number in a normal cluster (up to 80,000 using MCS-7845-I3 or OVA equivalent).
- Maximum number of CTI devices and connections also doubles.

However, some cluster-wide constants do not increase. Chief among these are:

- Size of the configuration database
- Number of locations and regions

Therefore, care should be taken when deciding whether or not to deploy a megacluster.

Cisco Unified CM Session Management Edition

The Session Management Edition (SME) is Unified CM in a specific deployment mode. Thus, all the sizing discussion for Unified CM applies to SME as well. The big distinction is that the call traffic in a pure SME deployment is solely through trunk interfaces rather than through line interfaces. Therefore, sizing an SME cluster is in general a simpler exercise than for Unified CM as a whole.

An SME cluster follows the same guidance as that for a regular Unified CM cluster. A publisher server provides the master configuration repository. A TFTP server may be co-resident with the publisher server if the number of phones or MGCP gateways in the cluster is relatively small. A redundancy ratio of 1:1 is recommended for call processing subscribers.

To size an SME cluster, assess the functionality that it is expected to perform. In a base configuration, the SME acts as a routing aggregation point for a number of leaf clusters. It also provides centralized PSTN access for all of the leaf clusters connected to it. In more advances configurations, the SME may also host centralized voice messaging, mobility, and conferencing. The performance of the SME is influenced by the type of trunk protocols that the leaf clusters use to connect to it and by the BHCA across these trunks.

The following considerations apply when sizing an SME cluster:

- The various types of trunk interfaces that the cluster services. The following trunk protocols are supported by the SME:
  - SIP
  - H.323
  - MGCP (Q.931)
  - SIP (Q.SIG)
  - H.323 Annex M1
  - MGCP (Q.SIG)
- The number of users that access SME cluster services through each type of trunk interface
- BHCA per user for each trunk interface to leaf clusters for intercluster calls
- BHCA per user for each trunk interface to leaf clusters for off-net (PSTN) calls

- The type of trunk interface used by the SME cluster to connect to the PSTN
- Average holding time for calls
- Number of route and translation patterns

If the SME acts as a service aggregation point, the following relevant sizing parameters come into play as well:

- If using centralized voice messaging, the percentage of calls that are sent to voice mail
- If using mobility, the number of users and the remote destinations per user
- If using conferencing, the conferencing dial-in interval

The performance of the SME is measured as calls per second across each pair of protocols. There are variations across the hardware platforms and software versions.

For SME sizing calculations, use the Cisco Unified CM SME Sizing Tool, available at http://tools.cisco.com/cucst.

Cisco Intercompany Media Engine

The sizing of servers used for running the Cisco Intercompany Media Engine (IME) depends solely on the quantity of directory numbers enrolled for the IME service. Table 29-15 lists the capacity of each supported server.

Table 29-15 IME Server Supported Capacities

Server Platform	Maximum Number of Enrolled DIDs
MCS 7825-I2/H2 and 7825-I4/H4	20,000
MCS 7845-I2/H2 and 7845-I3	40,000

Because all IME call media (audio and video) flow through the IME-enabled Cisco Adaptive Security Appliance (ASA), capacity depends on the type and number of calls flowing through it. The IME-enabled ASA monitors only the audio stream incoming from the internet for voice quality. The video media is not monitored for voice quality, but it does flow through the IME-enabled ASA for RTP-to-SRTP conversion, and the bandwidth of the video directly affects the number of sessions each ASA can handle. Table 29-16 provides capacity limits for the ASA-5550 and ASA-5580. Performance limits of other ASA models have not been validated yet.

Table 29-16 Maximum Number of IME Calls per Type of Call and ASA Model

ASA Model	Voice G.711	Video 300 kbps	Video 800 kbps	Video 1 Mbps
ASA-5500 4 GB	480 Calls	240 Calls	120 Calls	80 Calls
ASA-5580-20 4 GB	900 Calls	600 Calls	300 Calls	200 Calls

Impact of IME on Unified CM

Unified CM does not have a limit on the number of IME calls it can handle, but IME calls should be factored into the overall call capacity provided by the cluster. In addition, some calls through IME might need to be re-routed mid-call through gateways if the call quality is not considered acceptable. The expected number of calls re-routed this way should be considered both for Unified CM processing and for number of calls through the gateways.

Emergency Services

The Cisco Emergency Responder tracks the locations of phones and the access switch ports to which they are connected. The phones may be discovered automatically or entered manually into the Emergency Responder. Table 29-17 shows the server platforms that support the Emergency Responder and their maximum capacities.

Table 29-17 Cisco Emergency Responder Server Platforms and Capacities

Server Platform	Maximum Number of Automatically Tracked Phones	Maximum Number of Manually Configured Phones	Maximum Number of Roaming Phones	Maximum Number of Switches	Maximum Number of Switch Ports	Maximum Number of Emergency Response Locations
MCS-7816	6,000	1,000	600	200	12,000	1,000
MCS-7825 and OVA equivalent	12,000	2,500	1,200	500	30,000	3,000
MCS-7835 and OVA equivalent	20,000	5,000	2,000	1,000	60,000	7,500
MCS 7845 and OVA equivalent	30,000	10,000	3,000	2,000	120,000	10,000

The formal definitions of the OVA templates for Cisco Emergency Responder and other Unified Communication products are available at

> http://docwiki.cisco.com/wiki/Unified_Communications_Virtualization_Downloads_(including_OVA/OVF_Templates)

The capacity of Emergency Responder also affects the Unified CM cluster size. There can be only one Emergency Responder active per cluster. Therefore, choose the server that has sufficient resources to provide emergency coverage to all of the phones in the cluster.

For more details on network hardware and software requirements for Emergency Responder, refer to the *Cisco Emergency Responder Administration Guide*, available at

> http://www.cisco.com/en/US/products/sw/voicesw/ps842/prod_maintenance_guides_list.html

Gateways

The expected traffic in and out of gateways is the key to calculating the number of DS0s required. To calculate this traffic, consider the sources where the traffic can originate and terminate. The endpoints registered to Unified CM are, of course, the major traffic consumers, but there may be others such as interactive voice response (IVR) applications and other parts of a contact center deployment.

In addition to voice call termination, gateways can also perform a variety of other functions that require resources (either CPU and memory or DSP). These functions include media processing such as media termination point (MTP), transcoding, conference bridge, RSVP Agents, and others.

Gateways, especially those based on the Cisco Integrated Service Routers (ISRs), can provide services beyond just terminating PSTN traffic, such as serving as VXML processing engines, acting as border elements, doubling as Cisco Unified Communications Manager Express or Survivable Remote Site Telephony (SRST), or performing WAN edge functions. All of these other activities that the gateway is performing need to be taken into account when calculating the gateway load.

Gateway Groups

When considering the number of gateways, you also need to consider the geographical placement of physical gateway servers. In a deployment model where PSTN access is distributed, you need to size those gateways as a group by themselves and assign the appropriate amount of load to each such group.

A grouping might also be appropriate if certain gateways are expected to be dedicated for certain functions and share common characteristics.

Therefore, to accurately estimate the number of gateways required, the following information is required:

- Groups of gateways that share a common group profile. The common profiles will depend on the complexity of the deployment.
- For each group, the traffic patterns, platform, blocking probability, and so forth, that make up the profile.
- The individual gateway platform that makes up the group. In deciding on a particular gateway model, ensure that the model can support the capabilities and the capacity that is expected of it. Note that more than one gateway might be required in a gateway group, depending on the ability of the selected platform to meet the performance requirements.

PSTN Traffic

The discussion on voice traffic analysis earlier in this chapter (Call Traffic, page 29-28) is particularly pertinent to gateways in deciding the number of PSTN circuit DS0s that are required for time-division multiplexing (TDM) voice termination. Because there are likely to be fewer PSTN circuits than the number of system users, a decision has to be made about the optimum number of such circuits, and consequently the DSP requirements, for the gateways. The blocking factor determines the percentage of calls that may not be serviced at peak traffic levels. A smaller blocking factor requires more circuits.

Traffic is measured in Erlangs, and an Erlang is defined as one call lasting for one hour. This section does not go into any further detail on Erlangs other than to say that there are mathematical tables (Erlang-B and Erlang-C) that are used to calculate how many circuits are required for a given amount of offered traffic.

The amount of traffic received and generated by your business determines the size of the external circuits required. However, many customers typically continue to use the same number of circuits for their IP-based communications system as they previously used for a TDM-based system. While this sizing method might work if no issues are encountered, the process of ongoing system traffic analysis should be part of any routine maintenance practices. Traffic analysis can show that the system is over-provisioned for the current levels of traffic (and, therefore, the customer is paying for circuits that are not needed) or, conversely, that the system is under-provisioned and might be suffering from occasional blocked and/or lost calls, in which case increasing the number of circuits will remedy the situation.

Normal Business Traffic Profile

Most customers have a normal traffic profile, which means that they typically have two busy hours per day, one occurring during the morning from 10:00 to 11:00 and the other in the afternoon from 14:00 to 15:00. These busy-hour patterns can often be attributed to such things as employees starting the work day or returning from a lunch break. The calls tend to have longer hold times, and they tend to arrive and leave in a steady manner. A generally accepted industry average holding time to use for traffic calculations is 3 minutes.

Assuming that the communications system is engineered with the busy-hour traffic in mind, no issues should arise. Engineering a system below these levels will result in blocked and/or lost calls, which can have a detrimental effect on business.

Contact Center Traffic Profile

Contact centers present somewhat different patterns of traffic in that these systems typically handle large volumes of calls for the given number of agents or interactive voice response (IVR) systems available to service them. Contact centers want to get the most out of their resources, therefore their agents, trunks, and IVR systems are kept busy all the while they are in operation, which usually is 24 hours a day. Call queuing is typical (when incoming call traffic exceeds agent capacity, calls wait in queue for the next available agent), and the agents are usually dedicated during their work shifts to taking contact center calls.

Call holding times in contact centers are often of a shorter average duration than normal business calls. Contributing to the shorter average call holding time is the fact that many calls interact only with the IVR system and never need to speak to a human agent (also termed self-service calls). A representative holding time for self-service calls is about 30 seconds, while a call that talks to an agent has an average holding time of 3 minutes (the same as normal business traffic), making the overall average holding time in the contact center shorter than for normal business traffic.

The goal of contact centers to optimize resource use (including IVR ports, PSTN trunks, and human agents), combined with the fact that contact centers are systems dedicated to taking telephone calls, also presents the system with higher call arrival rates than in a typical business environment. These call arrival rates can also peak at different times of day and for different reasons (not the usual busy hour) than normal business traffic. For example, when a television advertisement runs for a particular holiday package with a 1-800 number, the call arrival rate for the system where those calls are received will experience a peak of traffic for about 15 minutes after the ad airs. This call arrival rate can exceed the average call arrival rate of the contact center by an order of magnitude.

Gateway Sizing for Contact Center Traffic

Short call durations as well as bursty call arrival rates impact the PSTN gateway's ability to process the traffic. Under these circumstances the gateway needs more resources to process all calls in a timely manner, as compared to gateways that receive calls of longer duration that are presented more uniformly over time. Because gateways have varying capabilities to deal with these traffic patterns, careful consideration should be given to selecting the appropriate gateway for the environment in which it will operate. Some gateways support more T1/E1 ports than others, and some are more able than others to deal with multiple calls arriving at the same time.

For a traffic pattern with multiple calls arriving in close proximity to each other (that is, high or bursty call arrival rates), a gateway with a suitable rating of calls per second (cps) is the best fit. Under these conditions, using calls with 15-second hold times, the Cisco AS5400XM Universal Gateway can maintain 16 cps with 250 calls active at once, the Cisco 3845 Integrated Services Router can maintain 13 cps with 200 calls active at once, and the Cisco 3945 Integrated Services Router can maintain 28 cps with 420 calls active at once. The performance of the Cisco AS5350XM Universal Gateway is identical to that of the AS5400XM in terms of calls per second.

For traffic patterns with a steady arrival rate, the maximum number of active calls that a gateway can handle is generally the more important consideration. Under these conditions, using calls with 180-second hold times, the Cisco AS5400XM Universal Gateway can maintain 600 simultaneously active calls with a call arrival rate of up to 3.3 cps, the Cisco 3845 Integrated Services Router can maintain 450 simultaneously active calls with a call arrival rate of up to 2.5 cps, and the Cisco 3945 Integrated Services Router can maintain 720 simultaneously active calls with a call arrival rate of up to 4 cps.

These numbers assume that all of the following conditions apply:

- CPU utilization does not exceed 75%
- PSTN gateway calls are made with ISDN PRI trunks using H.323
- The Real Time Control Protocol (RTCP) timer is set to the default value of 5 seconds
- Voice Activity Detection (VAD) is off
- G.711 uses 20 ms packetization
- Cisco IOS Release 15.0.1M is used
- Dedicated voice gateway configurations are used, with Ethernet (or Gigabit Ethernet) egress and no QoS features. (Using QoS-enabled egress interfaces or non-Ethernet egress interfaces, or both, will consume additional CPU resources.)
- No supplementary call features or services are enabled – such as general security (for example, access control lists or firewalls), voice-specific security (TLS, IPSec and/or SRTP), AAA lookups, gatekeeper-assisted call setups, VoiceXML or TCL-enabled call flows, call admission control (RSVP), and SNMP polling/logging. Such extra call features use additional CPU resources.

Voice Activity Detection (VAD)

Voice Activity Detection (VAD) is a digital signal processing feature that suppresses the creation of most of the IP packets during times when the speech path in a particular direction of the call is perceived to be silent. Typically only one party on a call speaks at a time, so that packets need to flow in only one direction, and packets in the reverse (or silent) direction need not be sent except as an occasional keepalive measure. VAD can therefore provide significant savings in the number of IP packets sent for a VoIP call, and thereby save considerable CPU cycles on the gateway platform. While the actual packet savings that VAD can provide varies with the call flow, the application, and the nature of speaker interactions, it tends to use 10% to 30% fewer packets than would be sent for a call made with VAD turned off.

VAD is most often turned off in endpoints and voice gateways deployed in Unified CM networks; VAD is most often turned on in voice gateways in other types of network deployments.

Codec

Both G.711 and G.729A use as their default configuration a 20 ms sampling time, which results in a 50 packets-per-second (pps) VoIP call in each direction. While a G.711 IP packet (200 bytes) is larger than a G.729A packet (60 bytes), this difference has not proven to have any significant effect on voice gateway CPU performance. Both G.711 and G.729 packets qualify as "small" IP packets to the router, therefore the packet rate is the salient codec parameter affecting CPU performance.

Performance Overload

Cisco IOS is designed to have some amount of CPU left over during peak processing, to handle interrupt-level events. The performance figures in this section are measured with the processor running at an average load of approximately 75%. If the load on a given Cisco IOS gateway continually exceeds this threshold, the following results will occur:

- The deployment will not be supported by Cisco Technical Assistance Center (TAC).
- The Cisco IOS Gateway will display anomalous behavior, including Q.921 time-outs, longer post-dial delay, and potentially interface flaps.

Cisco IOS Gateways are designed to handle a short burst of calls, but continual overloading of the recommended call rate (calls per second) is not supported.

Note With any gateway, you might be tempted to assign unused hardware ports to other tasks, such as on a Cisco Communication Media Module (CMM) gateway where traffic calculations have dictated that only a portion of the ports can be used for PSTN traffic. However, the remaining ports must remain unused, otherwise the CPU will be driven beyond supported levels.

Performance Tuning

The CPU utilization of a Cisco IOS Voice Gateway is affected by every process that is enabled in a chassis. Some of the lowest level processes such as IP routing and memory defragmentation will occur even when there is no live traffic on the chassis.

Lowering the CPU utilization can help to increase the performance of a Cisco IOS Voice Gateway by ensuring that there are enough available CPU resources to process the real-time voice packets and the call setup instructions. Table 29-18 describes some of the techniques for decreasing CPU utilization.

Table 29-18 Techniques for Reducing Gateway CPU Utilization

Technique	CPU Savings	Description
Enable Voice Activity Detection (VAD)	Up to 20%	Enabling VAD can result in up to 45% fewer voice packets in typical conversations. The difficultly is that, in scenarios where voice recognition is used or there are long delays, a reduction in voice quality can occur. Voice appears to "pop" in at the beginning and "pop" out at the end of talk spurts.
Disable Real Time Control Protocol (RTCP)	Up to 5%	Disabling RTCP results in less out-of-band information being sent between the originating and terminating gateways. This results in lower quality of statistics displayed on the paired gateway. This can also result in the terminating gateway having a call "hang" for a longer period of time if RTCP packets are being used to determine if a call is no longer active.

Table 29-18 Techniques for Reducing Gateway CPU Utilization (continued)

Technique	CPU Savings	Description
Disable other non-essential functions such as: Authentication, Authorization, and Accounting (AAA); Simple Network Management Protocol (SNMP); and logging	Up to 2%	Any of these processes, when not required, can be disabled and will result in lower CPU utilization by freeing up the CPU to provide faster processing of real-time traffic.
Change the call pattern to increase the length of the call (and reduce the number of calls per second)	Varies	This can be done by a variety of techniques such as including a long(er) introduction prompt played at the beginning of a call or adjusting the call script at the call center.

Additional Information

A full discussion of every gateway, its capabilities, and call processing capacities is not possible in this chapter. For more information on Cisco Voice Gateways, refer to the following documentation:

- Cisco Voice Gateway Solutions:

  http://www.cisco.com/en/US/products/sw/voicesw/index.html#~all-prod
- Gateway protocols supported with Cisco Unified Communications Manager (Unified CM):

  http://www.cisco.com/en/US/docs/voice_ip_comm/cucm/admin/8_0_1/ccmsys/a08gw.html
- Interfaces and signaling types supported by the following Cisco Voice Gateways:
  - Cisco 3900 Series Integrated Services Routers

    http://www.cisco.com/en/US/products/ps10536/products_relevant_interfaces_and_modules.html
  - Cisco 2900 Series Integrated Services Routers

    http://www.cisco.com/en/US/products/ps10537/products_relevant_interfaces_and_modules.html
  - Cisco 3800 Series Integrated Services Routers

    http://www.cisco.com/en/US/products/ps5855/products_relevant_interfaces_and_modules.html
  - Cisco 2800 Series Integrated Services Routers

    http://www.cisco.com/en/US/products/ps5854/products_relevant_interfaces_and_modules.html
- Gateway features supported with MGCP, SIP, and H.323:

  http://www.cisco.com/en/US/prod/collateral/routers/ps259/product_data_sheet0900aecd8057f2e0.pdf
- SIP gateway RFC compliance:

  http://www.cisco.com/en/US/prod/collateral/voicesw/ps6790/gatecont/ps6831/product_data_sheet0900aecd804110a2.html
- Skinny Client Control Protocol (SCCP) feature support with FXS gateways:

  http://www.cisco.com/en/US/prod/collateral/voicesw/ps6790/gatecont/ps2250/ps5516/product_data_sheet09186a00801d87f6.html

- Gateway capacities and minimum releases of Cisco IOS and Unified CM required for conferencing, transcoding, media termination point (MTP), MGCP, SIP, and H.323 gateway features:

  http://www.cisco.com/en/US/prod/collateral/routers/ps259/product_data_sheet0900aecd8057f2e0.pdf

- Various voice traffic calculators, including Erlang calculators:

  http://www.erlang.com/calculator/

Voice Messaging

Voice messaging is an application that needs to be sized not only by itself but also for its effect on other Unified Communications components, mainly Unified CM.

In sizing hardware for the voice messaging system itself (either Cisco Unity or Cisco Unity Connection), the total number of users in the system should be considered. Other items that impact messaging hardware are as follows:

- Number of calls during the busy hour that the application has to handle
- Average length of messages left on the servers
- Number of users who check their messages during the busy hour
- Average length of user sessions
- Any advanced operations such as voice recognition or text-to-speech sessions
- Any media transcoding
- Ports on the voice messaging system are analogous to the DS0s on a gateway and are shared resources that need to be optimized. The same considerations of probabilistic arrival and the need for blocking apply to both types of resources.

Table 29-19 shows the applicability of the various voice messaging solutions to the scalability requirements of the deployment.

Table 29-19 Scaling Voice Messaging Solutions

	Maximum Number of Users Supported on a Single Server (Failover or Clustered Deployment)			Maximum Number of Users in a Digital Networking Solution	
Solutions	**500**	**15,000**	**20,000**	**100,000**	**250,000**
Cisco Unity Express	Y	N	N	Y	Y
Cisco Business Edition	Y	N	N	N	N
Cisco Unity Connection (unified/integrated messaging)	Y	Y	Y	Y	N
Cisco Unity (unified and voice messaging)	Y	Y	N	Y	Y

Table 29-20 shows the maximum limits of various functions of different servers running Cisco Unity Connection.

Table 29-20 Servers and Capacities for Cisco Unity Connection

Server Platform	Maximum Number of Ports	Maximum Voice Recognition Sessions	Maximum Text to Speech Sessions	Maximum Number of Voicemail Users
MCS-7825	48	48	48	2,000
MCS-7835	150	150	150	4,000
MCS-7845	250	250	250	20,000
OVA Template for 5,000 users	100	100	100	5,000
OVA Template for 10,000 users	150	150	150	10,000
OVA Template for 20,000 users	250	250	250	20,000

The formal definitions of the OVA templates for Cisco Unity Connection and other Unified Communication products are available at

http://docwiki.cisco.com/wiki/Unified_Communications_Virtualization_Downloads_(including_OVA/OVF_Templates)

Impact on Unified CM

The impact of a voice messaging system on Unified CM can be gauged by considering the extra processing that Unified CM needs to do. These extra call flows add to the sizing load of Unified CM and is as follows:

- Calls that need to be forwarded to the voice messaging system when the user is not present or if the user deliberately forwards the calls using Do Not Disturb (DND) or other features.
- Calls from users who dial the voice messaging pilot number to access their voice messages go through Unified CM, and these calls must be added to the calls being handled by Unified CM, including both the number and the duration of these calls.

Collaborative Conferencing

Cisco Collaborative Conferencing systems include Cisco Unified CM as a component for call control. When sizing such a system, the function it performs as well as its impact to Unified CM should be considered.

When sizing such conferencing systems, you typically have to consider the following parameters to determine the type and number of servers:

- Number of users who could use the system at any one time
- Number of audio, video, and web users on the system at the peak usage time
- Required dial-in duration
- Video resolution and audio codec requirements

Sizing Guidelines for Audio Conferencing

Cisco recommends the following methods for calculating audio conferencing capacity:

- Calculation based on average monthly usage

  If you know the average voice conferencing usage (average minutes per month), use Table 29-21 to calculate the audio conferencing capacity.

Table 29-21 Audio Conferencing Capacity Based on Average Monthly Usage

Average Monthly Usage (minutes)	Baseline Usage (minutes per port per month)	Estimated Number of Ports
20,000 to 50,000	1,500	15 to 35
50,000 to 500,000	2,000	25 to 250
500,000 to 1,000,000	3,000	165 to 335
1,000,000 to 2,000,000	3,500	285 to 570
2,000,000 to 8,000,000	4,000	500 to 2,000

- Calculation based on number of users

  You should plan on having one port for every 20 users with average usage. If the users are heavy conference users, then provision one port for every 15 users. For example, in a system with 6000 users, you should provision 300 audio ports; however, if those users heavily use conferencing, then plan for 400 audio ports.

- Calculation based on actual peak usage

  Actual voice conferencing usage during peak hours usually can be obtained from existing voice conferencing system logs or service provider bills. Cisco recommends provisioning 30% extra capacity based on the actual peak usage in order to protect against extra conferencing volume.

Factors Affecting System Sizing

In addition to the estimates provided by the methods described above for the system baseline port requirement, the following factors also affect system sizing:

- When migrating from an "operator-scheduled" model to a user-scheduled model, you might need to add another 20% to the baseline.
- The default average meeting size is 4.5 callers per meeting. Use the value that is applicable to your case if it is different than the default.
- Increase the baseline estimate accordingly if the following condition applies:

  (Estimated meetings per day) ∗ (Estimated users) > 80% of baseline

- If the largest single meeting exceeds 20% of the estimated capacity, increase the estimate accordingly.
- If there are continuous meetings with dedicated ports, then you must add those additional ports ((Meetings) ∗ (Dedicated callers)) to the baseline.

The total number of ports will include all the above factors in addition to the baseline. Plan for conferencing system capacity expansion if the total estimated port capacity exceeds 80% of the maximum supported ports.

Sizing Guidelines for Video Conferencing

Cisco recommends the following three methods for calculating video conferencing capacity:

- Calculation based on number of knowledgeable workers

  Cisco recommends provisioning a video user license for every 40 knowledgeable workers.

- Calculation based on number of voice conferencing user licenses

  Cisco recommends provisioning video conferencing capacity in the range of 17% to 25% of existing audio user licenses. The percentage depends on business requirements regarding video conferencing and on the size of the conferencing system.

- Calculation based on existing video Multipoint Control Unit (MCU)

  Cisco recommends deploying a direct replacement for an existing video conferencing system.

Impact on Unified CM

The impact to Unified CM can be analyzed by the extra call traffic that the conferencing system generates. The most impact occurs when conference users dial into their meetings that are typically scheduled at the top of the hour or half-hour. A large amount of call traffic within a few minutes of conference start times increases the load on Unified CM for just those few minutes that must be designed in appropriately. In addition, if conference users include callers from the PSTN or from other clusters, those parameters must also be considered to gauge their impact on the gateways.

Sizing Guidelines for Specific Conferencing Systems

The following sections describe more sizing information specific to particular Cisco Collaborative Conferencing systems.

Cisco WebEx Meetings Server

In Cisco WebEx Meetings Server, usage of various audio codecs and secure conferencing has no impact to the system capacity. For details, refer to the system capacity table in the *Cisco WebEx Meetings Server System Requirements*, available at

http://www.cisco.com/en/US/products/ps12732/products_installation_and_configuration_guides_list.html

Cisco Unified MeetingPlace

In Cisco Unified MeetingPlace, the Hardware Media Server (HMS) and Express Media Server (EMS) are two media server options available to customers. Use the following parameters to determine which option is sufficient:

- Usage of iLBC or other high-complexity audio codecs. Usage of these codecs requires an HMS.
- Media options such as video continuous presence and echo cancellation. Both of these options require an HMS.

Video usage characteristics such as bandwidth and resolution are an important aspect for the sizing of both kinds of media servers (HMS and EMS).

The capacity of a given Cisco Unified MeetingPlace solution depends on the platform on which the Unified MeetingPlace Meeting Directors, Application servers with EMS or HMS, or WebEx Node for MCS or ASR servers are installed, followed by the capacity of the Unified MeetingPlace Media Servers deployed. For example, with the Unified MeetingPlace Application server installed on a Cisco MCS 7845-I3 (or equivalent) server, voice conferencing can scale to 1,200 ports (G.711) with EMS or 2,000 ports (G.711) with HMS in a single system or conferencing node.

Additional Factors Affecting System Sizing

Consider the following recommendations to maintain the maximum capacity with Cisco Unified MeetingPlace:

- If an audio codec other than G.711 is desired, use transcoders based on Cisco Integrated Services Routers (ISR) to achieve maximum capacity.
- Use Line Echo Cancellation (LEC) provided by an external device such as an ISR, rather than the build-in LEC from Unified MeetingPlace.

Express Media Server

The Cisco Unified MeetingPlace Express Media Server (EMS) capacity is directly related to codec and video bandwidth because it is installed co-resident with the Unified MeetingPlace Application server. When the Unified MeetingPlace Application server is installed on a Cisco MCS 7835-H2/I2 server, the overall system capacity decreases for both EMS and HMS deployments. Standards-based video as well as G.729 and G.722 audio codecs all affect the capacity of the EMS system. For the detailed capacity numbers, refer to the latest version of the *Planning Guide for Cisco Unified MeetingPlace*, available at

http://www.cisco.com/en/US/products/sw/ps5664/ps5669/products_implementation_design_guides_list.html

The EMS introduces the concept of System Resource Units (SRUs), where the system capacity (or the Total SRUs value) is based on the type of hardware platform on which the Unified MeetingPlace Application Server resides and the speed and number of processors on that system. The system immediately consumes some of these SRUs from the total for normal operation, and it puts the remaining resources in an SRU pool and makes them available for enhanced audio and video features. Table 29-22 shows the number of total SRUs available for enhanced audio and video per supported platform.

Table 29-22 Total System Resource Units per Supported EMS Platform

Server Platform	Total System Resource Units (SRUs) Available for Enhanced Audio and Video
MCS 7835-I3	400
MCS 7845-I2/H2	500
MCS 7845-I3	1,200
UCS B200 or C210 Series	1,200 (with or without Meeting Director co-resident)
UCS C200 Series	500 (2 nodes with redundancy)

Table 29-23 Number of System Resource Units Consumed for Various Audio Codecs and Video Bandwidths

Session Type	Number of SRUs Used
One G.711 audio port	1
One G.729 or one G.722 audio port	6
One video port at 320 kbps[1]	1

Table 29-23 Number of System Resource Units Consumed for Various Audio Codecs and Video Bandwidths (continued)

Session Type	Number of SRUs Used
One video port at 384 kbps	1
One video port at 768 kbps	2
One video port at 2,000 kbps	6

[1] The lowest rate that is guaranteed for a video license is 320 kbps.

As shown by the data in Table 29-22 and Table 29-23, on an MCS 7845-I3 server handling only G.711 audio calls, the EMS supports 1,200 audio sessions. Alternatively, it supports 600 video sessions at up to 384 kbps with G.711 audio (a video session also consumes SRUs for the audio session).

In Unified CM, the regions setting of the SIP trunk used for call delivery to Unified MeetingPlace can be configured to control the audio codec and video bandwidth of calls sent to the EMS. Understanding the nature and capabilities of the endpoints dialing into Unified MeetingPlace is critical to proper design. For more information on EMS capacity planning, refer to the latest version of the *Planning Guide for Cisco Unified MeetingPlace*, available at

http://www.cisco.com/en/US/products/sw/ps5664/ps5669/products_implementation_design_guides_list.html

Hardware Media Server

The Cisco Unified MeetingPlace Hardware Media Server (HMS) uses some different settings than the EMS. The Global Audio Mode setting in Unified MeetingPlace Application Administration directly affects the voice capacity of Unified MeetingPlace HMS audio blades. The Global Audio Mode can be configured in either of the following ways:

- G.711 and G.729 without Line Echo Cancellation (LEC)

  With this configuration setting, a single audio blade in the HMS can support a maximum of 250 voice ports. It would require 8 audio blades to reach the maximum supported system limit of 2,000 concurrent audio sessions.

- G.711, G.722, iLBC, or G.729 with Line Echo Cancellation (LEC)

  With this configuration, a single audio blade can support a maximum of 166 voice ports. With 8 audio blades, the maximum supported number of concurrent audio sessions using these additional codecs is 1,328.

The Global Video Mode setting in Unified MeetingPlace Application Administration determines the video capacity of Unified MeetingPlace HMS video blades. The Global Video Mode can be configured in either of the following ways:

- Standard Rate (video call speed up to 384 kbps)

  In this mode, a video blade in the HMS can support a maximum of 40 video ports.

- High Rate (video call speed up to 2,048 kbps)

  In this mode, a video blade can support a maximum of 20 video ports.

For a complete list of the video formats supported by Unified MeetingPlace, refer to the latest version of the *Compatibility Matrix for Cisco Unified MeetingPlace*, available at

http://www.cisco.com/en/US/products/sw/ps5664/ps5669/products_device_support_tables_list.html

Unified MeetingPlace Hardware Media Servers can be either the Cisco Unified MeetingPlace 3515 or the Cisco Unified MeetingPlace 3545 chassis. The Unified MeetingPlace 3515 is a fixed platform that comes with one audio blade and one video blade installed. The Unified MeetingPlace 3545 is a modular platform consisting of a chassis that supports four audio blades or video blades in various combinations.

Cascading of Audio and Video Blades

If multiple audio blades and video blades are installed in the Unified MeetingPlace 3545, the media server uses virtual cascading to overflow voice and video streams from one audio or video blade to another. The audio blade has built-in cascading ports that do not decrease the audio session capacity. With a single video blade deployed in the Unified MeetingPlace system, all video ports are available for video conferencing. With multiple video blades deployed, the media server will automatically reserve video ports for cascading purposes. For Standard Rate video, 8 video ports are reserved for cascading, leaving 40 video ports available. For High Rate video, 4 video ports are reserved for cascading, leaving 20 video ports available.

Example 29-1 Cisco Unified MeetingPlace Audio Conference

A Cisco Unified MeetingPlace 3545 media server is deployed with two audio blades and two video blades. A meeting is scheduled with 350 audio ports, and the Global Audio Mode is configured for G.711 with LEC. In this case:

- The media server allocates 251 ports from the first audio blade, out of which 250 ports are used for audio participants and one port is used for voice cascading or connecting to the second audio blade.
- The media server allocates 101 ports from the second audio blade, out of which 100 ports are used for audio participants and one port is used for voice cascading.

Example 29-2 Cisco Unified MeetingPlace Video Conference

A Cisco Unified MeetingPlace 3545 media server is deployed with two audio blades and two video blades. For this example, assume a meeting is scheduled with 65 video ports and the Global Video Mode is configured for Standard Rate video. In this case:

- The media server allocates 41 ports from the first video blade, out of which 40 ports are used for video participants and one port is used for video cascading or connecting to the second video blade.
- The media server allocates 26 ports from the second video blade, out of which 25 ports are used for video participants and one port is used for video cascading.

Unified MeetingPlace Web Server

The Unified MeetingPlace Web Server is required only for Unified MeetingPlace scheduling deployments to schedule and attend meetings from the Web user interface, for Lotus Notes integrations, or for accessing the recording storage. There is no capacity planning consideration for these servers. Cisco MCS 7835 servers are sufficient for the largest Unified MeetingPlace deployment, but MCS 7845 servers may be used as well.

WebEx Node for MCS

Web conferencing optionally using Cisco WebEx Node for MCS can accommodate up to 500 web sessions, depending on the type of hardware on which the WebEx Node for MCS resides. A maximum of four WebEx Nodes for MCS can be deployed per solution, or an unlimited number of nodes can be deployed with WebEx Node for ASR, allowing for scalability up to 2,000 web sessions on-premises with redundancy. WebEx Nodes can be distributed anywhere in a customer network, but Cisco recommends deploying them closest to the larger groups of web users. Only internal users will have web sessions with the WebEx Node for MCS or ASR; external users will always connect to the cloud. For detailed capacities of the Cisco WebEx Node for MCS or ASR, refer to the latest version of the *Planning Guide for Cisco Unified MeetingPlace*, available at

> http://www.cisco.com/en/US/products/sw/ps5664/ps5669/products_implementation_design_guides_list.html

Cisco IM and Presence

As with all other applications, sizing for Cisco IM and Presence is accomplished in the following manner:

- Decomposing the system into its most elemental services
- Measuring the unit cost of each of these services
- Analyzing the given system description as an aggregation of the identified services and arriving at a net system cost
- Determining the number of required servers based on system cost and deployment options

For IM and Presence, the following system variables in the system under analysis are relevant and must be considered for accurate sizing:

- Number and type of users
  - Clients employed by users to obtain presence services
  - Operating mode for users (instant messaging only or full Unified Communications facilities)
- Presence-related activities performed by typical users
  - Contact list size and composition (intra-cluster, inter-cluster, and federated)
  - Number of instant messages (directly between two users) per user during the busy hour
  - Chat support with number of chat rooms, users per chat room, and instant messages per user per chat room
  - State changes per user (both call related and user initiated)
- Deployment model
  - Whether intercluster presence is supported
  - Whether federation is supported
  - Whether high availability is desired
- Server preferences
  - The class of server or voice messaging platform desired
- System options
  - Whether compliance recording is required

Once the system requirements are quantified, the number of required servers can be determined from the data in Table 29-24.

Table 29-24 Maximum Number of Users Supported per IM and Presence Cluster

Server Platform	Maximum Users Supported in Full Unified Communications Mode	Maximum Users Supported in Instant Messaging Only Mode
MCS-7816	3,000	7,500
MCS-7825 and OVA equivalent	6,000	6,000
MCS-7835 and OVA equivalent	15,000	37,500
MCS-7845 and OVA equivalent	45,000	75,000

For additional information, refer to the latest version of *Hardware and Software Compatibility Information for IM and Presence Service on Cisco Unified Communications Manager*, available at

http://www.cisco.com/en/US/products/sw/voicesw/ps556/products_device_support_tables_list.html

The formal definitions of the OVA templates for Cisco IM and Presence and other Unified Communication products are available at

http://docwiki.cisco.com/wiki/Unified_Communications_Virtualization_Downloads_(including_OVA/OVF_Templates)

Impact on Unified CM

The Cisco IM and Presence Service influences the performance of Unified CM in the following ways:

- User synchronization through an AXL/SOAP interface
- Presence information through a SIP trunk
- CTI traffic to enable phone control

In general, the impact of user synchronization (except for a one-time hit) and that of presence information through the SIP trunk are negligible. The affect of CTI control of phones, however, must be counted against CTI limits.

Cisco Unified Communications Management Suite

The Cisco Unified Communications Management Suite consists of four applications. Sizing for these applications is relatively simple and depends directly on the number of endpoints or network devices that they are expected to manage. These applications can work either in a standalone mode hosted on separate hardware servers or in a co-resident environment on a single server.

The server characteristics to host the Unified Communications Management Suite applications are generally stated in terms of hardware specifications: CPU characteristics (processor speed and number of cores), memory, and disk space for each level of desired capacity.

The co-resident servers, for example, can host from one to all four of the Unified Communications Management Suite applications, and they specified in two configurations – one for managing up to 2,000 endpoints and a larger configuration for managing up to 10,000 endpoints. The specifications for such co-resident servers are as follows:

- Large co-resident configuration:
  - Processor: 3 GHz, 8 Core
  - Memory: 16 GB
  - Disk space: 320 GB
- Small co-resident configuration:
  - Processor: 3 GHz, 4 Core
  - Memory: 8 GB
  - Disk Space: 100 GB

These hardware characteristics can be mapped to the equivalent Cisco MCS or UCS servers.

Cisco Prime Unified Provisioning Manager

The Cisco Prime Unified Provisioning Manager (Unified PM) can support up to 60,000 phones and can be implemented either on a single machine or on two machines. A two-machine deployment is recommended when the number of phones exceeds 30,000.

Hardware resources required for various levels of performance are described in the *Cisco Unified Provisioning Manager Data Sheet*, available at

http://www.cisco.com/en/US/products/ps7125/products_data_sheets_list.html

Cisco Prime Unified Operations Manager

The Cisco Prime Unified Operations Manager (Unified OM) can manage phones and other network devices such as routers and switches. The Unified Operations Manager operates in a single machine configuration. The Unified OM supports up to 45,000 phones and 2,000 other IP devices.

Hardware resources required for various levels of performance are described in the *Cisco Unified Operations Manager Data Sheet*, available at

http://www.cisco.com/en/US/products/ps6535/products_data_sheets_list.html

Cisco Prime Unified Service Monitor

The Cisco Prime Unified Service Monitor (Unified SM) consists of not only the server to run the Unified SM software but also on Cisco 1040 Sensor and Network Analysis Modules (NAMs) to measure voice quality.

Table 29-25 Performance for 1040 Sensor and Different NAM Types

	Cisco Network Analysis Module Type				
	1040 Sensor	NME-NAM	NAM-2	NAM 2204 Appliance	NAM 2220 Appliance
Maximum number of concurrent RTP streams supported	100	100	400	1,500	4,000

Hardware resources required for various levels of performance are described in the *Cisco Unified Service Monitor Data Sheet*, available at

http://www.cisco.com/en/US/products/ps6536/products_data_sheets_list.html

Unified SM supports the following voice quality monitoring capacities:

- Up to 50 Cisco 1040 Sensors
- Up to 45,000 IP phones
- Up to 5,000 sensor-based RTP streams per minute (with Cisco 1040 Sensors or NAM modules)
- Up to 1,600 Cisco Voice Transmission Quality (CVTQ) calls per minute
- Up to 1,500 RTP streams and 666 CVTQ calls per minute

Cisco Unified Service Statistics Manager

The Cisco Unified Service Statistics Manager (Unified SSM) operates in a single server mode and can scale to manage up to 45,000 phones.

Hardware resources required for various levels of performance are described in the *Cisco Unified Service Statistics Manager Data Sheet*, available at

http://www.cisco.com/en/US/products/ps7285/products_data_sheets_list.html

Conclusion

In summary, it can be challenging to determine the hardware composition of a large Unified Communications system consisting of a number of separate applications working together. However, an understanding of the functional requirements of the software and the performance capabilities of the hardware platforms certified for running the software, can be very helpful in making an accurate estimation of the servers required. Tools developed by Cisco for this purpose are available as described in this chapter. For further assistance, contact your Cisco Partner or Cisco Systems Engineer, and they can use the Cisco Unified Communications Sizing Tool (http://tools.cisco.com/cucst) to validate all designs.

GLOSSARY

Revised: June 28, 2012; OL-27282-05

A

AA	Automated attendant
AAD	Alerts and Activities Display
AAR	Automated Alternate Routing
AC	Cisco Attendant Console
ACD	Automatic call distribution
ACE	Cisco Application Control Engine
ACF	Admission Confirm
ACL	Access control list
ACS	Access Control Server
AD	Microsoft Active Directory
ADAM	Active Directory Application Mode
ADPCM	Adaptive Differential Pulse Code Modulation
ADUC	Active Directory Users and Computers
AES	Advanced Encryption Standards
AFT	ALI Formatting Tool
AGM	Cisco Access Gateway Module
ALG	Application Layer Gateway
ALI	Automatic Location Identification
AMI	Alternate mark inversion
AMIS	Audio Messaging Interchange Specification
AMWI	Audible message waiting indication
ANI	Automatic Number Identification

AP	Access point
APDU	Application protocol data unit
API	Application Program Interface
ARJ	Admission Reject
ARP	Address Resolution Protocol
ARQ	Admission Request
ASA	Cisco Adaptive Security Appliance
ASP	Active server page
ASR	Automatic speech recognition
ATA	Cisco Analog Telephone Adapter
ATM	Asynchronous Transfer Mode
AXL	Administrative XML Layer

B

BAT	Cisco Bulk Administration Tool
BBWC	Battery-backed write cache
BES	Blackberry Enterprise Server
BFCP	Binary Flow Control Protocol
BGP	Border Gateway Protocol
BHCA	Busy hour call attempts
BHCC	Busy hour call completions
BIB	Built-in bridge
BLF	Busy lamp field
BOSH	Bidirectional-streams Over Synchronous HTTP
BPDU	Bridge protocol data unit
bps	Bits per second
BRI	Basic Rate Interface
BTN	Bill-to number

C

CA	Certificate Authority
CAC	Call admission control
CAM	Content-addressable memory
CAMA	Centralized Automatic Message Accounting
CAPF	Certificate Authority Proxy Function
CAPWAP	Control and Provisioning of Wireless Access Points
CAR	Cisco CDR Analysis and Reporting
CAS	Channel Associated Signaling
CBWFQ	Class-Based Weighted Fair Queuing
CCA	Clear channel assessment
CCD	Call Control Discovery
CCS	Common channel signaling
CDP	Cisco Discovery Protocol
CDR	Call detail record
CGI	Common Gateway Interface
CIF	Common Intermediate Format
CIR	Committed information rate
CKM	Cisco Centralized Key Management
CLEC	Competitive local exchange carrier
CLID	Calling line identifier
CM	Cisco Unified Communications Manager (Unified CM)
CMC	Client Matter Code
CME	Cisco Unified Communications Manager Express (Unified CME)
CMI	Cisco Messaging Interface
CMM	Cisco Communication Media Module
CNG	Comfort noise generation
CO	Central office

Co-located	Two or more devices in the same physical location, with no WAN or MAN connection between them
COM	Component Object Model
COP	Cisco Options Package
COR	Class of restriction
Co-resident	Two or more services or applications running on the same server
CoS	Class of service
CPCA	Cisco Unity Personal Assistant
CPI	Cisco Product Identification tool
CPN	Calling party number
CRS	Cisco Customer Response Solution
cRTP	Compressed Real-Time Transport Protocol
CSF	Client Services Framework
CSTA	Computer-Supported Telecommunications Applications
CSUF	Cross-Stack UplinkFast
CSV	Comma-separated values
CTI	Computer telephony integration
CTL	Certificate Trust List
CUBE	Cisco Unified Border Element, formerly the Cisco Multiservice IP-to-IP Gateway (IP-IP Gateway)
CUE	Cisco Unity Express
CUMI	Cisco Unity Connection Messaging Interface
CUPI	Cisco Unity Connection Provisioning Interface
CUSP	Cisco Unified SIP Proxy
CVTQ	Cisco Voice Transmission Quality

D

DC	Domain controller
DDNS	Dynamic Domain Name Server
DDR	Delayed Delivery Record

DFS	Dynamic Frequency Selection
DHCP	Dynamic Host Configuration Protocol
DID	Direct inward dial
DIT	Directory Information Tree
DMVPN	Dynamic Multipoint Virtual Private Network
DMZ	Demilitarized zone
DN	Directory number
DNIS	Dialed number identification service
DNS	Domain Name System
DoS	Denial of service
DPA	Digital PBX Adapter
DSCP	Differentiated Services Code Point
DSE	Digital set emulation
DSP	Digital signal processor
DTIM	Delivery Traffic Indicator Message
DTLS	Datagram Transport Layer Security protocol
DTMF	Dual tone multifrequency
DTPC	Dynamic Transmit Power Control
DUC	Domino Unified Communications Services

E

E&M	Receive and transmit, or ear and mouth
EAP	Extensible Authentication Protocol
EAPOL	Extensible Authentication Protocol over LAN
EC	Echo cancellation
ECM	Error correction mode
ECS	Empty Capabilities Set
EI	Enhanced Image

EIGRP	Enhanced Interior Gateway Routing Protocol
E-L CAC	Enhanced Locations call admission control
ELIN	Emergency location identification number
ELM	Enterprise License Manager
EM	Extension Mobility
EMCC	Extension Mobility Cross Cluster
ER	Cisco Emergency Responder
ERL	Emergency response location
ESF	Extended Super Frame
E-SRST	Enhanced Survivable Remote Site Telephony

F

FAC	Forced Authorization Code
FCC	Federal Communications Commission
FCoE	Fibre Channel over Ethernet
FECC	Far End Camera Control
FIFO	First-in, first-out
FQDN	Fully qualified domain name
FR	Frame Relay
FWSM	Firewall Services Module
FXO	Foreign Exchange Office
FXS	Foreign Exchange Station

G

GARP	Gratuitous Address Resolution Protocol
GC	Global catalog
GKTMP	Gatekeeper Transaction Message Protocol
GLBP	Gateway Load Balancing Protocol

GMS	Greeting management system
GPO	Group Policy Object
GPRS	General Packet Radio Service
GSM	Global System for Mobile Communication
GSS	Global Site Selector
GUI	Graphical user interface
GUP	Gatekeeper Update Protocol

H

H.225D	H.225 daemon
HDLC	High-Level Data Link Control
HMS	Hardware Media Server
HP	Hewlett-Packard
HSRP	Hot Standby Router Protocol
HTTP	Hyper-Text Transfer Protocol
HTTPS	HTTP Secure
HVD	Hosted virtual desktop
Hz	Hertz

I

IANA	Internet Assigned Numbers Authority
IAPP	Inter-Access Point Protocol
ICA	Independent Computing Architecture
ICCS	Intra-Cluster Communication Signaling
ICMP	Internet Control Message Protocol
ICS	IBM Cabling System
ICT	Intercluster trunk
IE	Information Element

IETF	Internet Engineering Task Force
IGMP	Internet Group Management Protocol
IIS	Microsoft Internet Information Server
IM	Instant messaging
IMAP	Internet Message Access Protocol
IntServ	Integrated Services
IntServ/DiffServ	Integrated Services/Differentiated Services
IOPS	Input/output operations per second
IP	Internet Protocol
IPCC	Cisco IP Contact Center
IPMA	Cisco IP Manager Assistant
IPPM	Cisco IP Phone Messenger
IPSec	IP Security
ISO	International Standards Organization
ISR	Integrated Services Router
ITEM	CiscoWorks IP Telephony Environment Monitor
ITU	International Telecommunication Union
IVR	Interactive voice response

J

JTAPI	Java Telephony Application Programming Interface

K

kbps	Kilobits per second
KPML	Key Press Markup Language

L

LAN	Local area network

LBM	Locations Bandwidth Manager
LBR	Low bit-rate
LCD	Liquid crystal display
LCF	Location Confirm
LCS	Live Communications Server
LDAP	Lightweight Directory Access Protocol
LDAPS	LDAP over SSL
LDIF	LDAP Data Interchange Format
LDN	Listed directory number
LEAP	Lightweight Extensible Authentication Protocol
LEC	Local Exchange Carrier
LFI	Link fragmentation and interleaving
LLDP	Link Layer Discovery Protocol
LLDP-MED	Link Layer Discovery Protocol for Media Endpoint Devices
LLQ	Low-latency queuing
LRG	Local route group
LRJ	Location Reject
LRQ	Location Request
LSC	Locally significant certificate
LUN	Logical unit number
LWAP	Light Weight Access Point
LWAPP	Light Weight Access Point Protocol

M

MAC	Media Access Control
MAN	Metropolitan area network
Mbps	Megabits per second
MCM	Multimedia Conference Manager

MCS	Media Convergence Server
MCU	Multipoint Control Unit
MDN	Mobile Data Network
MDS	Mobile Data Services
MFT	Multiflex trunk
MGCP	Media Gateway Control Protocol
MIB	Management Information Base
MIC	Manufacturing installed certificate
MIME	Multipurpose Internet Mail Extension
MIPS	Millions of instructions per second
MISTP	Multiple Instance Spanning Tree Protocol
MITM	Man-in-the-middle
MLA	Cisco Multi-Level Administration
MLP	Multilink Point-to-Point Protocol
MLPP	Multilevel Precedence and Preemption
MLPPP	Multilink Point-to-Point Protocol
MLTS	Multi-line telephone system
MMoIP	Multimedia Mail over IP
MMP	Mobile Multiplexing Protocol
MOC	Microsoft Office Communicator
MoH	Music on hold
MOS	Mean Opinion Score
MPLS	Multiprotocol Label Switching
MRG	Media resource group
MRGL	Media resource group list
ms	Millisecond
MSP	Managed service provider
MTP	Media termination point

mW	Milli-Watt
MWI	Message Waiting Indicator

N

NAT	Network Address Translation
NDR	Non-delivery receipt
NENA	National Emergency Number Association
NFAS	Non-Facility Associated Signaling
NIC	Network interface card
NOC	Network operations center
NPA	Numbering Plan Area
NSE	Named Service Event
NSF	Network Specific Facilities
NTE	Named Telephony Event
NTP	Network Time Protocol

O

ORA	Open Recording Architecture
OSPF	Open Shortest Path First
OU	Organizational unit
OVA	Open Virtualization Archive
OWA	Outlook Web Access

P

PAC	Protected Access Credential
PBX	Private branch exchange
PC	Personal computer
PCI	Peripheral Component Interconnect

PCM	Pulse code modulation
PCoIP	PC over IP
PCTR	Personal call transfer rule
PD	Powered device
PHB	Per-hop behavior
PIN	Personal identification number
PINX	Private integrated services network exchange
PIX	Private Internet Exchange
PKI	Public Key Infrastructure
PLAR	Private Line Automatic Ringdown
PoE	Power over Ethernet
POTS	Plain old telephone service
PPP	Point-to-Point Protocol
pps	Packets per second
PQ	Priority Queue
PRACK	Provisional Reliable Acknowledgement
PRI	Primary Rate Interface
PSAP	Public safety answering point
PSE	Power source equipment
PSK	Pre-Shared Key
PSTN	Public switched telephone network
PVC	Permanent virtual circuit

Q

QBE	Quick Buffer Encoding
QBSS	QoS Basic Service Set
QoS	Quality of Service
QSIG	Q signaling

R

RADIUS	Remote Authentication Dial-In User Service
RAS	Registration Admission Status
RCP	Remote Copy Protocol
RDNIS	Redirected Dialed Number Information Service
REST	Representational State Transfer
RF	Radio frequency
RFC	Request for Comments
RIM	Research In Motion
RIP	Routing Information Protocol
RIS	Real-Time Information Server
RMTP	Reliable Multicast Transport Protocol
RoST	RSVP over SIP Trunks
RSNA	Reservationless Single Number Access
RSP	Route/Switch Processor
RSSI	Relative Signal Strength Indicator
RSTP	Rapid Spanning Tree Protocol
RSVP	Resource Reservation Protocol
RTCP	Real-Time Transport Control Protocol
RTMP	Real-Time Messaging Protocol
RTMT	Cisco Real-Time Monitoring Tool
RTP	Real-Time Transport Protocol
RTSP	Real Time Streaming Protocol
RTT	Round-trip time

S

S1, S2, S3, and S4	Severity levels for service requests
SaaS	Software-as-a-Service
SAF	Service Advertisement Framework
SAN	Storage area networking
SBC	Session Border Controller
SCCP	Skinny Client Control Protocol
SCSI	Small Computer System Interface
SDI	System Diagnostic Interface
SDK	Software Development Kit
SDL	Signaling Distribution Layer
SDP	Session Description Protocol
SE	Cisco Systems Engineer
SF	Super Frame
SFTP	Secure File Transfer Protocol
SI	Standard Image
SIMPLE	SIP for Instant Messaging and Presence Leveraging Extensions
SIP	Session Initiation Protocol
SIS	Symbian installation system
SIW	Service Inter-Working
SLB	Server load balancing
SLDAP	Secure LDAP
SMA	Segmented Meeting Access
SMDI	Simplified Message Desk Interface
SMS	Short Message Service
SMTP	Simple Mail Transfer Protocol
SNMP	Simple Network Management Protocol
SOAP	Simple Object Access Protocol

SPA	Shared Port Adapter
SQL	Structured Query Language
SRND	Solution Reference Network Design
SRST	Survivable Remote Site Telephony
SRSV	Survivable Remote Site Voicemail
SRTP	Secure Real-Time Transport Protocol
SRV	Server
SS7	Signaling System 7
SSID	Service set identifier
SSL	Secure Sockets Layer
SSO	Single Sign-On
STP	Spanning Tree Protocol
SUP1	Cisco Supervisor Engine 1
SUP2	Cisco Supervisor Engine 2
SUP2+	Cisco Supervisor Engine 2+
SUP3	Cisco Supervisor Engine 3

T

TAC	Cisco Technical Assistance Center
TAPI	Telephony Application Programming Interface
TCD	Telephony Call Dispatcher
TCER	Total Character Error Rate
TCL	Tool Command Language
TCP	Transmission Control Protocol
TCS	Terminal Capabilities Set
TDD	Telephone Device for the Deaf
TDM	Time-division multiplexing
TEHO	Tail-end hop-off

TFTP	Trivial File Transfer Protocol
TKIP	Temporal Key Integrity Protocol
TLS	Transport Layer Security
ToD	Time of day
ToS	Type of service
TPC	Transmit Power Control
TRaP	Telephone record and playback
TRP	Trusted Relay Point
TSP	Telephony Service Provider
TTL	Time to live
TTS	Text-to-speech
TTY	Terminal teletype
TUI	Telephony user interface

U

UAC	User agent client
UAS	User agent server
UCCN	Unified Client Change Notifier
UCS	Cisco Unified Computing System
UDC	Universal data connector
UDLD	UniDirectional Link Detection
UDP	User Datagram Protocol
UDPTL	Unnumbered Datagram Protocol Transport Layer
UMTS	Universal Mobile Telecommunications System
UN	Unsolicited SIP Notify
UNC	Universal Naming Convention
UP	User Priority
UPS	Uninterrupted power supply

URI	Uniform resource identifier
USB	Universal Serial Bus
UTIM	Cisco Unity Telephony Integration Manager
UTP	Unshielded twisted pair
UUIE	User-to-User Information Element

V

V3PN	Cisco Voice and Video Enabled Virtual Private Network
VAD	Voice activity detection
VAF	Voice-Adaptive Fragmentation
VATS	Voice-Adaptive Traffic Shaping
VCS	Cisco TelePresence Video Communication Server
VDI	Virtual Desktop Infrastructure
VIC	Voice interface card
VLAN	Virtual local area network
VMO	Cisco ViewMail for Outlook
VoIP	Voice over IP
VoPSTN	Voice over the PSTN
VoWLAN	Voice over Wireless LAN (WLAN)
VPIM	Voice Profile for Internet Mail protocol
VPN	Virtual private network
VRRP	Virtual Router Redundancy Protocol
VUI	Voice User Interface
VWIC	Voice/WAN interface card
VXC	Cisco Virtualization Experience Client
VXI	Cisco Virtualization Experience Infrastructure

W

WAN	Wide area network
WebDAV	Web-Based Distributed Authoring and Versioning
WEP	Wired Equivalent Privacy
WFQ	Weighted fair queuing
WINS	Windows Internet Naming Service
WLAN	Wireless local area network
WLC	Wireless LAN controller
WLSM	Cisco Wireless LAN Services Module
WMM	Wi-Fi Multimedia
WMM TSPEC	Wi-Fi Multimedia Traffic Specification
WPA	Wi-Fi Protected Access

X

XCP	Jabber Extensible Communications Platform
XML	Extensible Markup Language
XMPP	Extensible Messaging and Presence Protocol

INDEX