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.

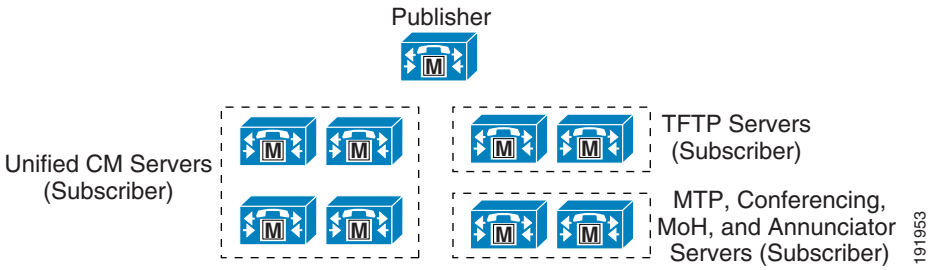
Note While Cisco recommends using the same server model for all servers in a cluster, mixing server models within a cluster is supported provided that all of the individual hardware versions are supported and that all servers are running the same version of Unified CM. However, differences in capacity 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

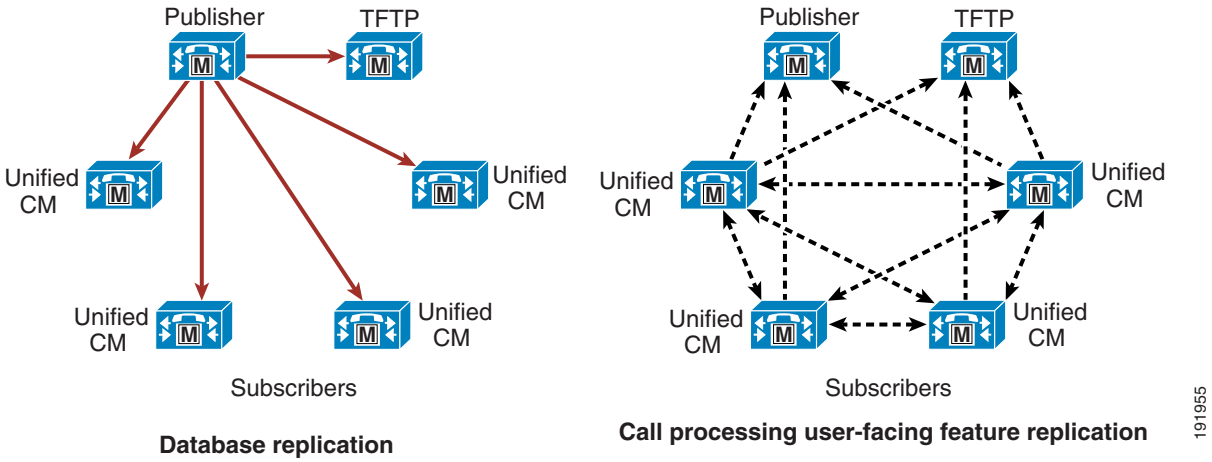
There are two primary kinds of intracluster communications, or communications within a Unified CM cluster (see Figure 8-2 and Figure 8-3.) The first is a mechanism for distributing the database that contains all the device configuration information (see “Database replication” in Figure 8-2). The configuration 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

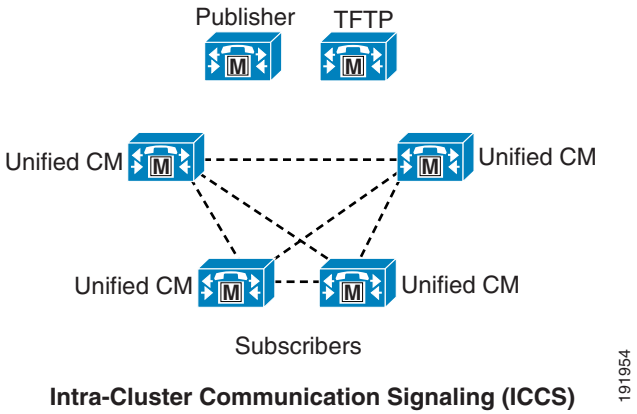
- Extension Mobility login/logout
- Hunt Group login/logout
- Device Mobility
- Certificate Authority Proxy Function (CAPF) status for end users and applications users
- Credential hacking and authentication

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.

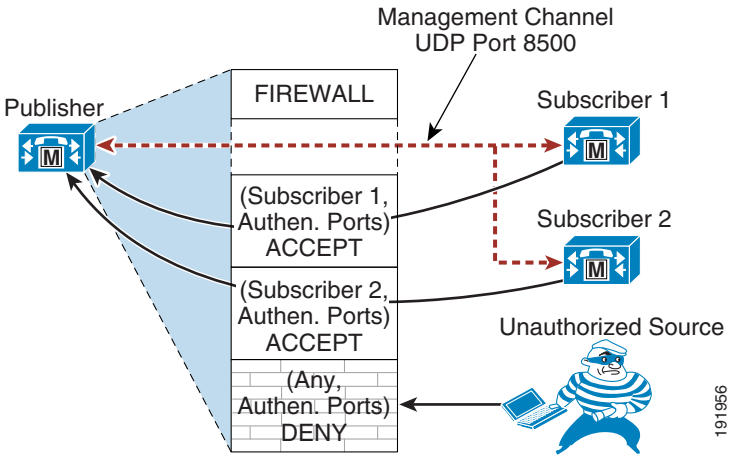
Figure 8-3 Intra-Cluster Communication Signaling (ICCS)



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

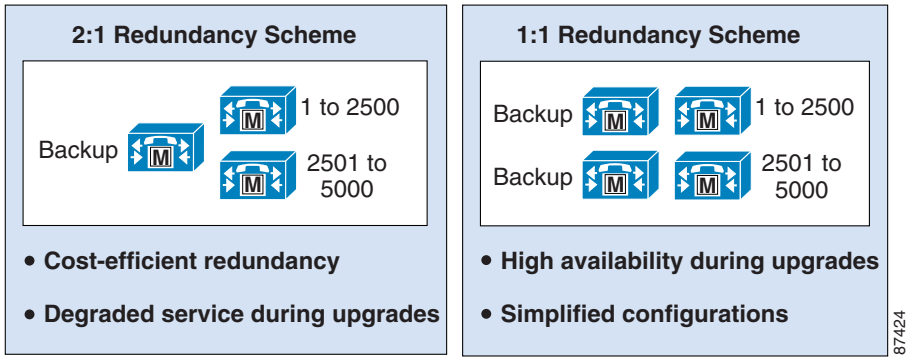


Figure 8-5 illustrates the two basic redundancy schemes available. In each case the backup server must be capable of handling the capacity of at least a single primary call processing server failure. In the 2:1 redundancy scheme, the backup might have to be capable of handling the failure of a single 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

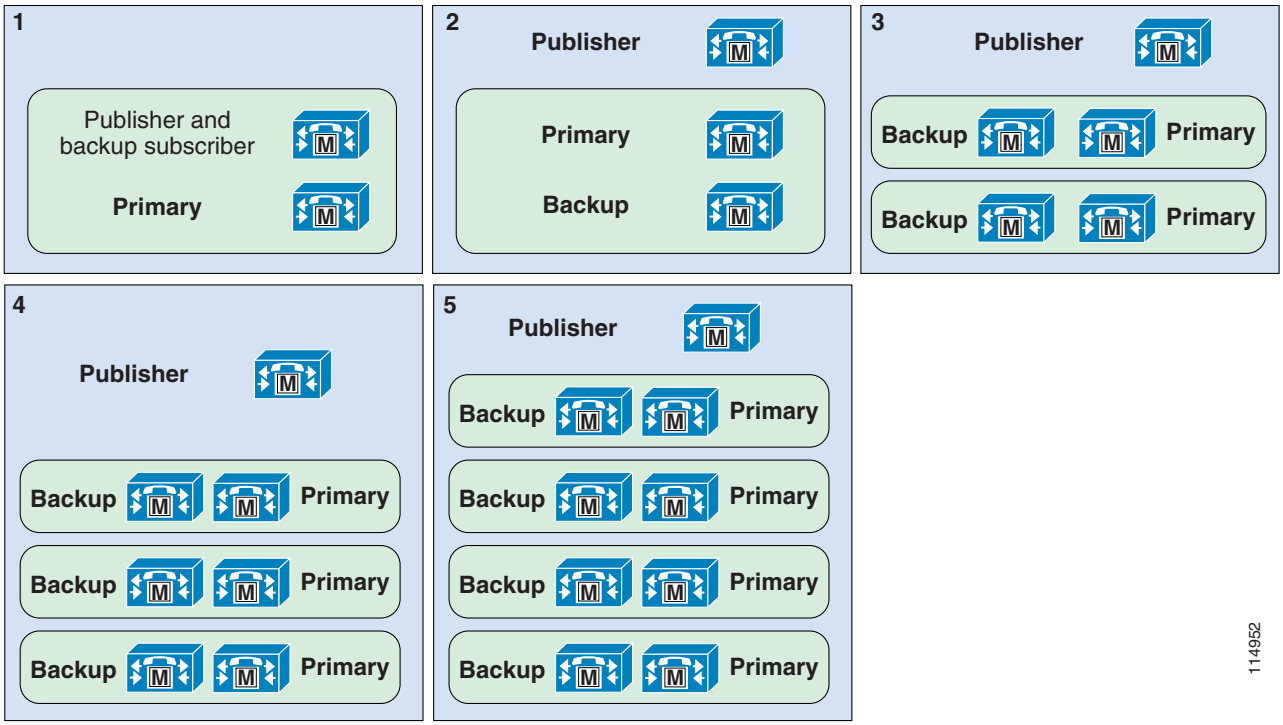
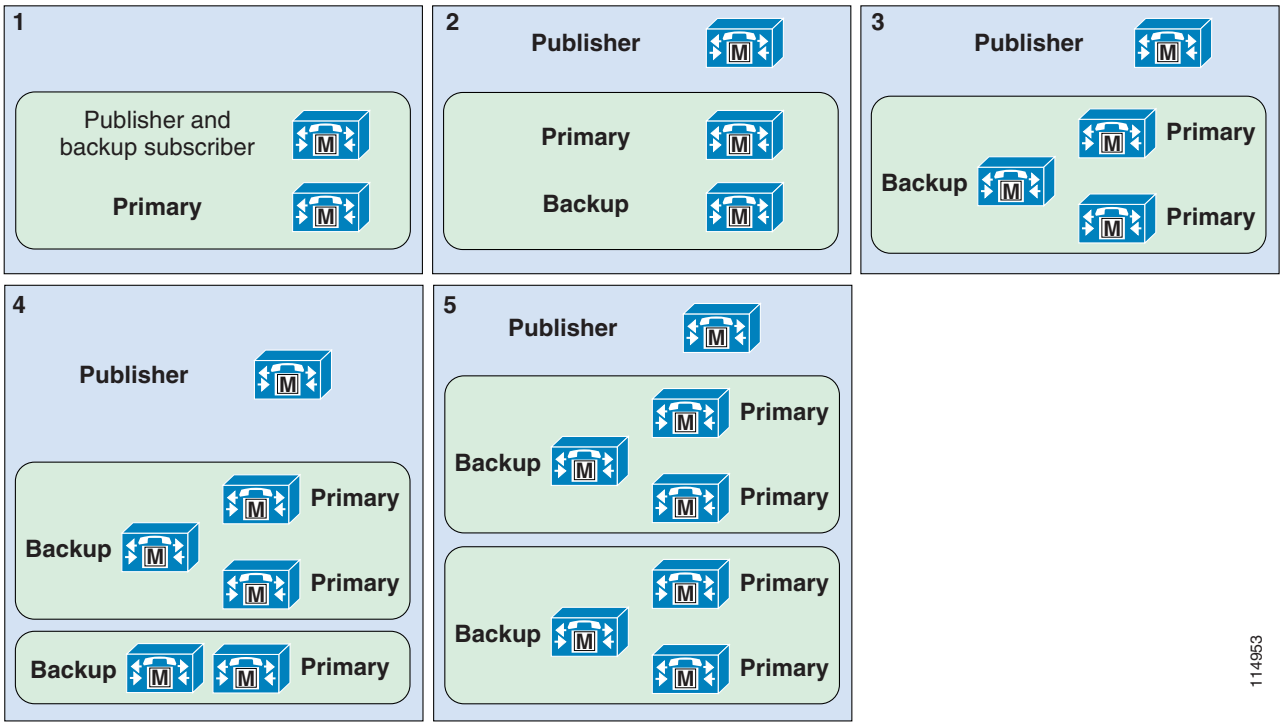


Figure 8-7 2:1 Redundancy Configuration Options



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.

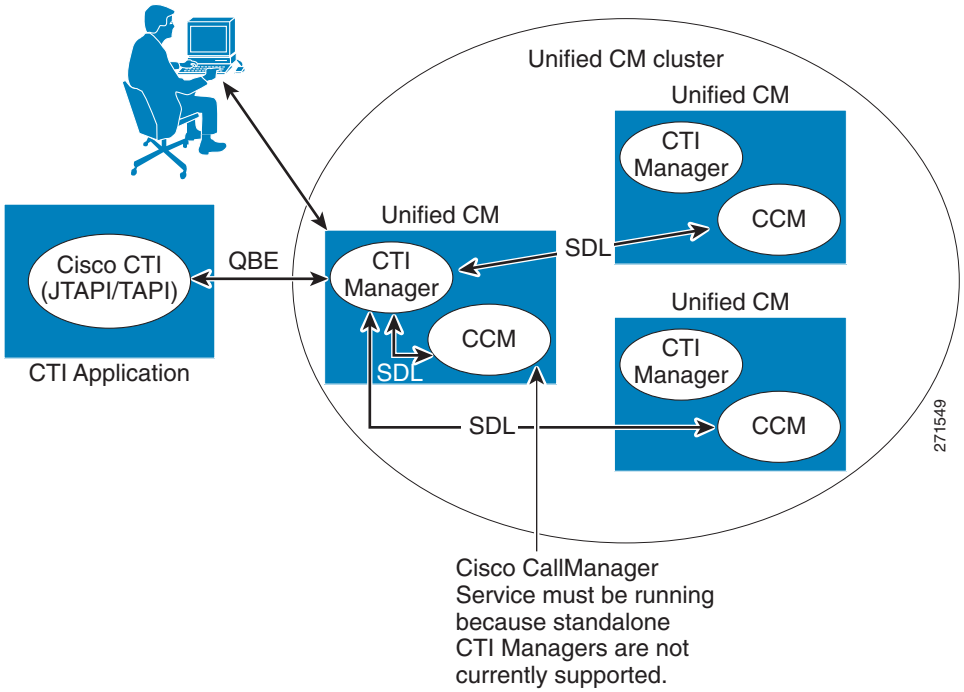
CTI Architecture

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.

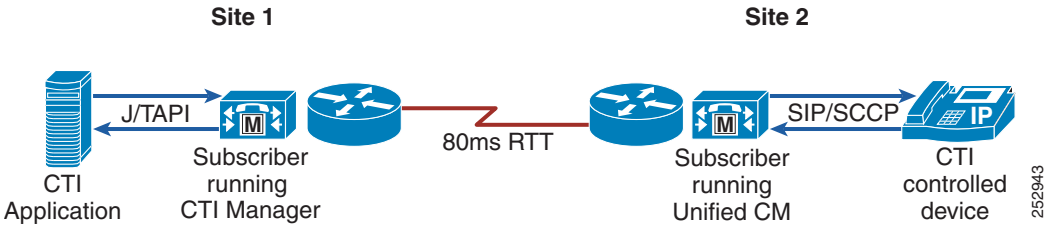
CTI Applications and Clustering Over the WAN

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).

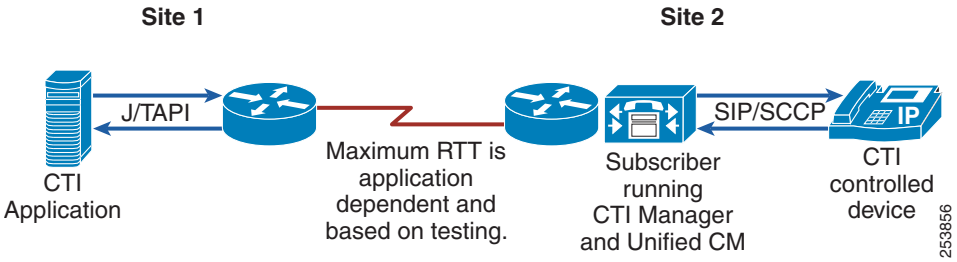
Figure 8-10 CTI Over the WAN



- 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

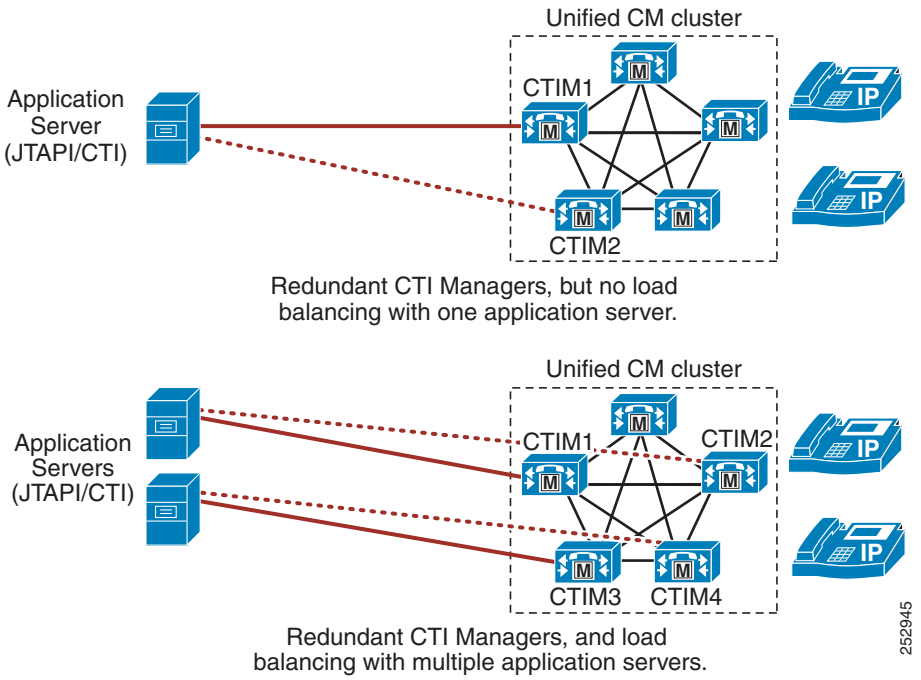


Figure 8-13 shows an example of this type of configuration for Cisco Unified Contact Center Enterprise (Unified CCE). This type of configuration has the following characteristics:

- 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

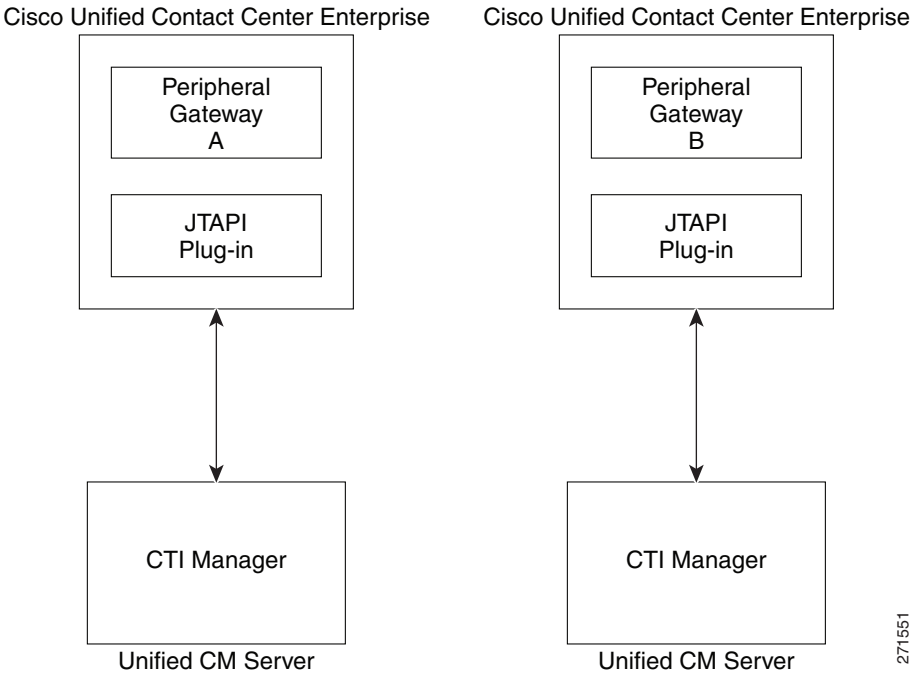
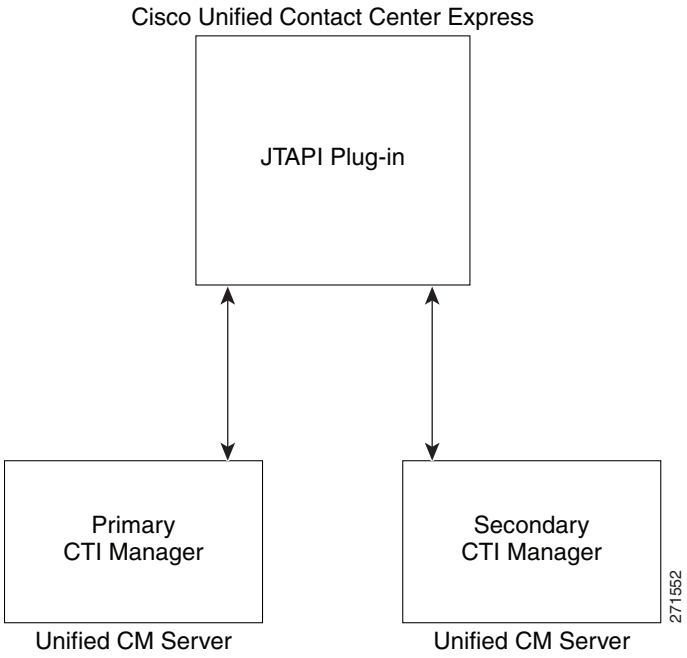


Figure 8-14 shows an example of this type of configuration for Cisco Unified Contact Center Express (Unified CCX). This type of configuration has the following characteristics:

- 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:

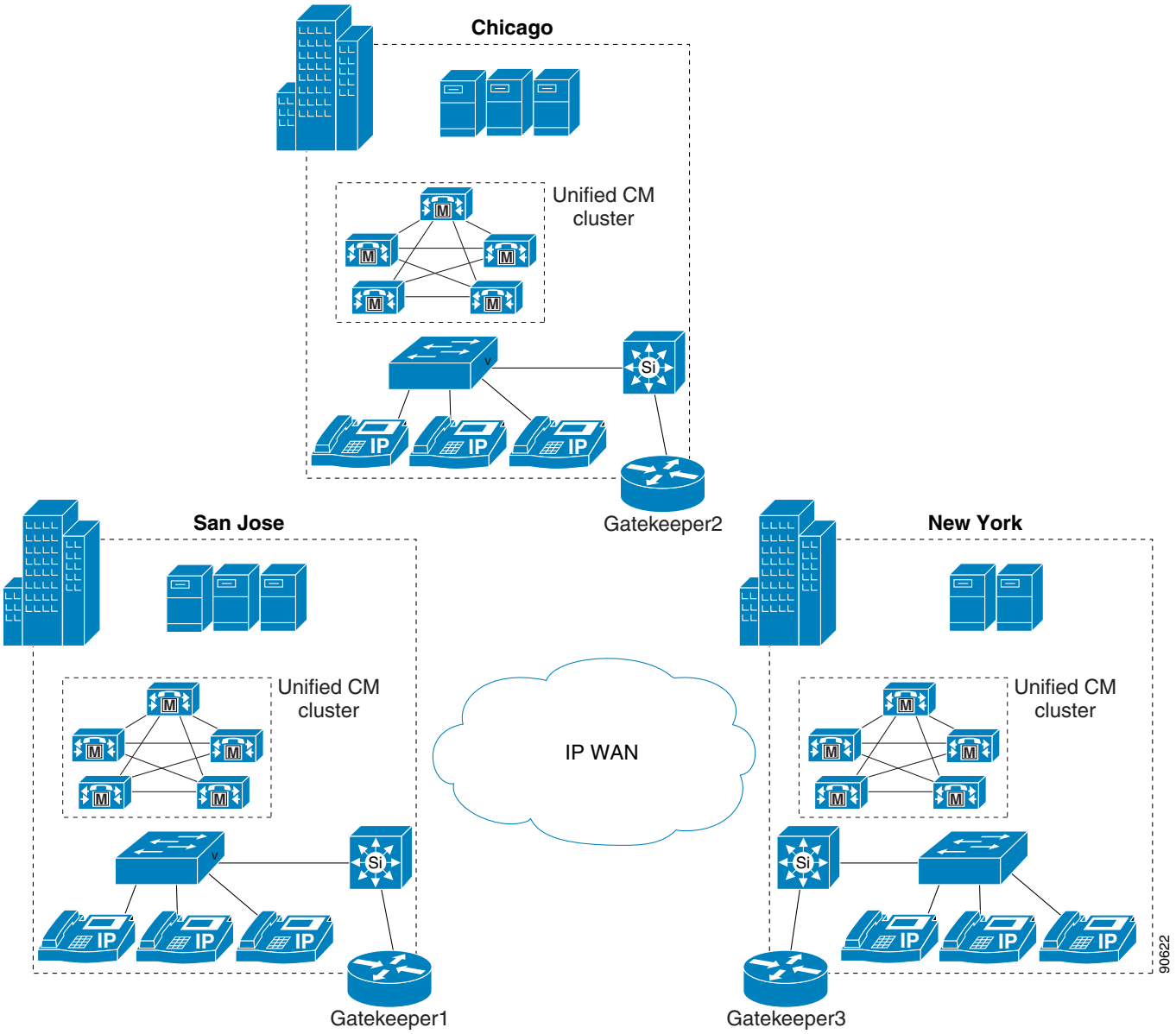
- CPU utilization
- Memory utilization
- Number of active calls
- Number of registered endpoints

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.

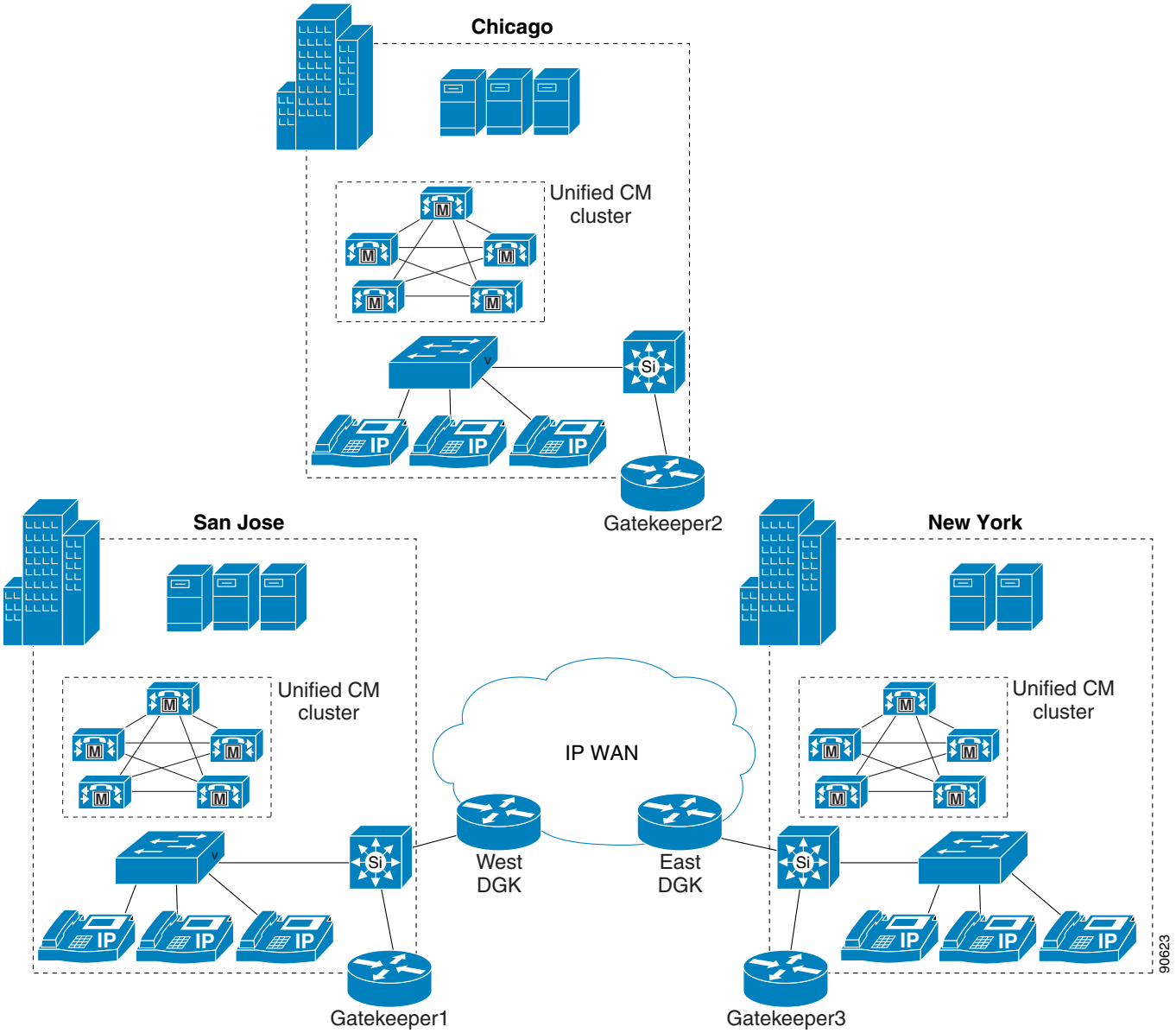
- LRQ Blast

  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.

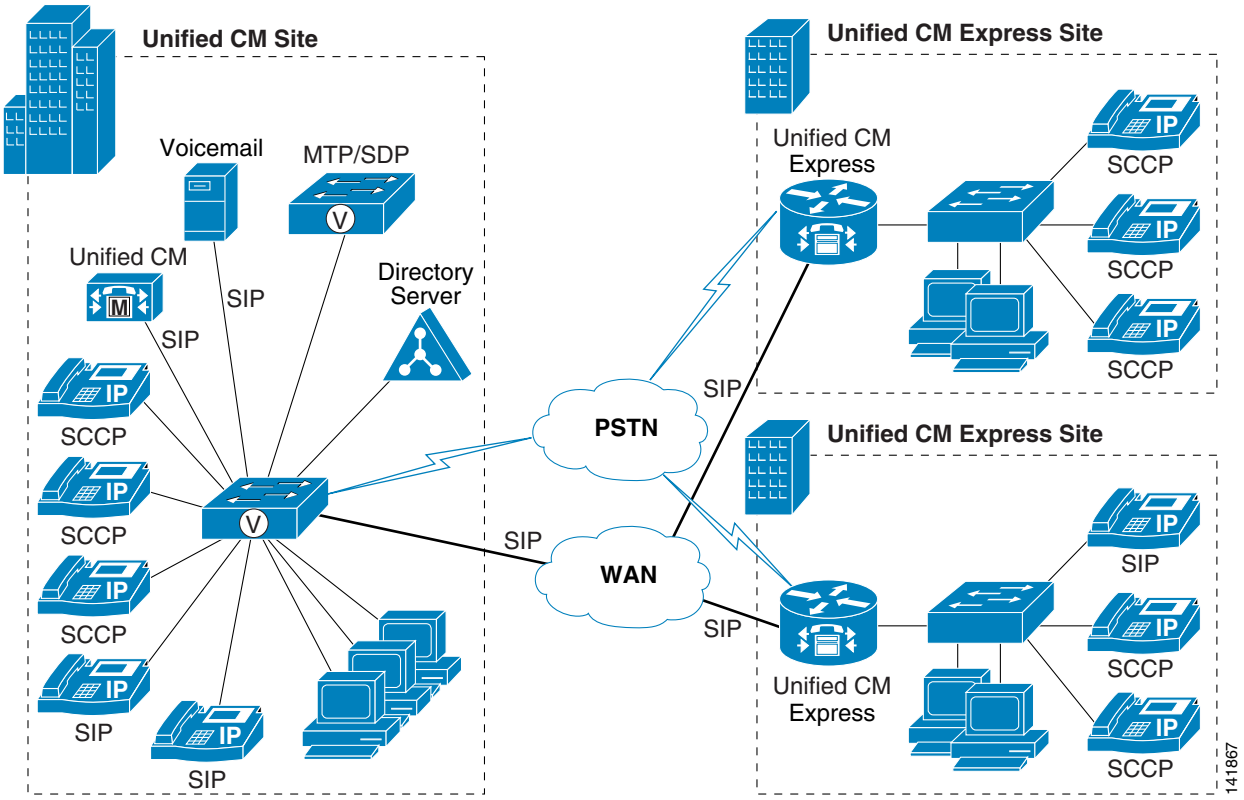
For information on required and supported DSP resources and the maximum number of conference participants allowed for Ad Hoc or Meet Me conferences, refer to the Unified CME product documentation available at

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.

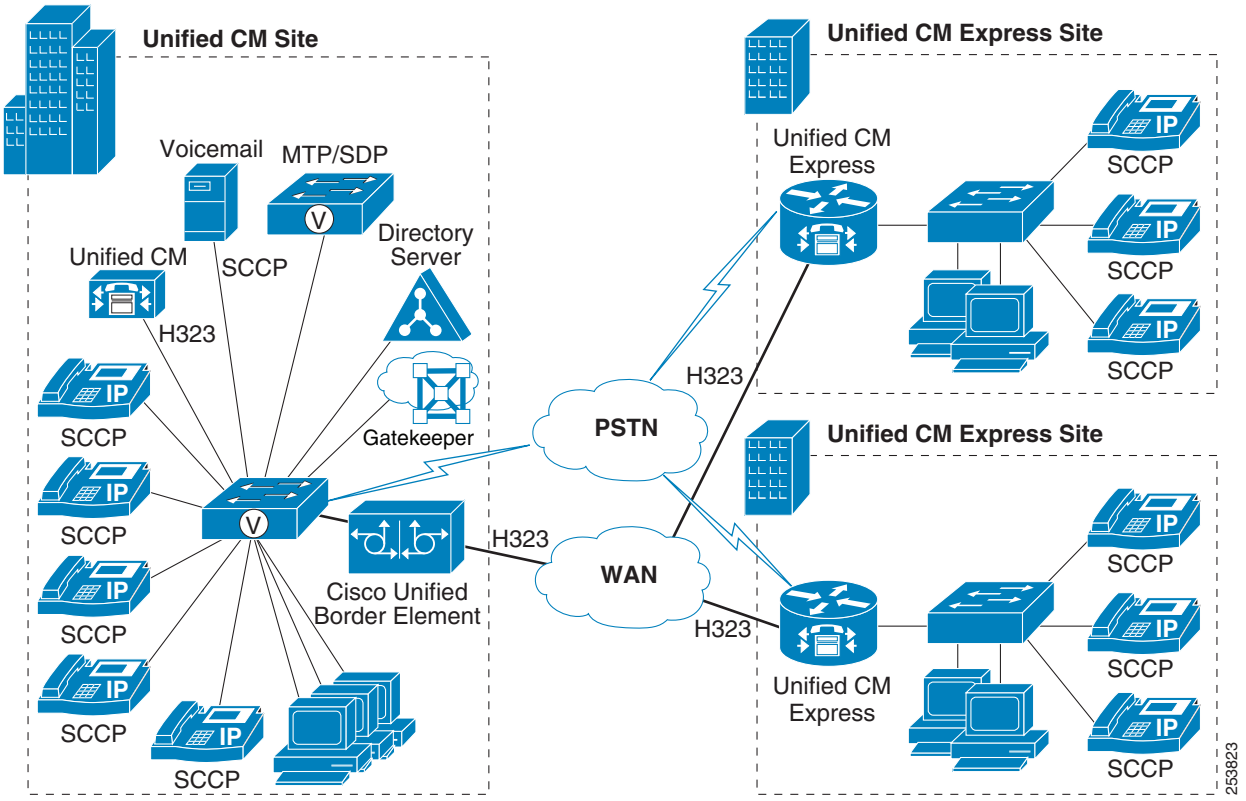
The second option is to deploy a via-zone gatekeeper. Unified CM, Unified CME, and the Cisco Unified Border Element all register with the via-zone gatekeeper as VoIP gateway devices. The via-zone gatekeeper performs dial plan resolution and bandwidth restrictions between Unified CM and Unified CME. The 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