CHAPTER 5

Unified Communications Deployment Models

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

This chapter describes the deployment models for Cisco Unified Communications Systems.

Earlier versions of this chapter based the deployment models discussion on the call processing deployment models for Cisco Unified Communications Manager (Unified CM) exclusively. The current version of this chapter, by contrast, introduces a site-based approach to the design guidance for the constituent technologies of the Cisco Unified Communications System. The intent is to offer design guidance for the entire Cisco Unified Communications System, which includes much more than just the call processing service.

For design guidance with earlier releases of Cisco Unified Communications, refer to the Cisco Unified Communications Solution Reference Network Design (SRND) documentation available at

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

What's New in This Chapter

Table 5-1 lists the topics that are new in this chapter or that have changed significantly from previous releases of this document.

Table 5-1 New or Changed Information Since the Previous Release of This Document

New or Revised Topic	Described in	Revision Date
Minor correction to Cisco IOS extended ping example	Delay Testing, page 5-36	April 30, 2013
Cisco Unified Survivable Remote Site Telephony (SRST) Manager	Cisco Unified Survivable Remote Site Telephony Manager, page 5-18	August 31, 2012
Minor updates for Cisco Unified Communications System Release 9.0	Various sections throughout this chapter	June 28, 2012

Deployment Model Architecture

In general terms, the deployment model architecture follows that of the enterprise it is deployed to serve. Deployment models describe the reference architecture required to satisfy the Unified Communications needs of well-defined, typical topologies of enterprises. For example, a centralized call processing deployment model caters to enterprises whose operational footprint is based on multiple sites linked to one or few centralized headquarters offices.

In some cases, the deployment model of a technology will depart from that of the enterprise, due to technological constraints. For example, if an enterprise has a single campus whose scale exceeds that of a single service instance (such as a call processing service provided by Cisco Unified Communications Manager), then a single campus might require more than a single instance of a call processing cluster or a single messaging product.

Another option for customers who exceed the sizing limits of a standard cluster is to consider deploying a megacluster, which can provide increased scalability. For more information about megaclusters, see Megacluster, page 8-26.

Note Unless otherwise specified, all information contained within this SRND that relates to call processing deployments (including capacity, high availability, and general design considerations) applies only to a standard cluster.

High Availability for Deployment Models

Unified Communications services offer many capabilities aimed at achieving high availability. They may be implemented in various ways, such as:

- Failover redundancy

  For services that are considered essential, redundant elements should be deployed so that no single point of failure is present in the design. The redundancy between the two (or more) elements is automated. For example, the clustering technology used in Cisco Unified Communications Manager (Unified CM) allows for up to three servers to provide backup for each other. This type of redundancy may cross technological boundaries. For example, a phone may have as its first three preferred call control agents, three separate Unified CM servers belonging to the same call processing cluster. As a fourth choice, the phone can also be configured to rely on a Cisco IOS router for call processing services.

- Redundant links

  In some instances, it is advantageous to deploy redundant IP links, such as IP WAN links, to guard against the failure of a single WAN link.

- Geographical diversity

  Some products support the distribution of redundant service nodes across WAN links so that, if an entire site is off-line (such as would be the case during an extended power outage exceeding the capabilities of provisioned UPS and generator backup systems), another site in a different location can ensure business continuance.

Capacity Planning for Deployment Models

The capacities of various deployment models are typically integrally linked to the capacities of the products upon which they are based. Where appropriate in this chapter, capacities are called out. For some of the products supporting services covered in more detail in other sections of this document, the capacities of those products are discussed in their respective sections.

Site-Based Design

Across all technologies that make up the Cisco Unified Communications System, the following common set of criteria emerges as the main drivers of design:

Size

In this context, size generally refers to the number of users, which translates into a quantity of IP telephones, voice mail boxes, presence watchers, and so forth. Size also can be considered in terms of processing capacity for sites where few (or no) users are present, such as data centers.

Network Connectivity

The site's connectivity into the rest of the system has three main components driving the design:

- Bandwidth enabled for Quality of Service (QoS)
- Latency
- Reliability

These components are often considered adequate in the Local Area Network (LAN): QoS is achievable with all LAN equipment, bandwidth is typically in the Gigabit range, latency is minimal (in the order of a few milliseconds), and excellent reliability is the norm.

The Metropolitan Area Network (MAN) often approaches the LAN in all three dimensions: bandwidth is still typically in the multiple Megabit range, latency is typically in the low tens of milliseconds, and excellent reliability is common. Packet treatment policies are generally available from MAN providers, so that end-to-end QoS is achievable.

The Wide Area Network (WAN) generally requires extra attention to these components: the bandwidth is at a cost premium, the latencies may depend not only on effective serialization speeds but also on actual transmission delays related to physical distance, and the reliability can be impacted by a multitude of factors. The QoS performance can also require extra operational costs and configuration effort.

Bandwidth has great influence on the types of Unified Communications services available at a site, and on the way these services are provided. For example, if a site serving 20 users is connected with 1.5 Mbps of bandwidth to the rest of the system, the site's voice, presence, instant messaging, email, and video services can readily be hosted at a remote datacenter site. If that same site is hosting 1000 users, some of the services would best be hosted locally to avoid saturating the comparatively limited bandwidth with signaling and media flows. Another alternative is to consider increasing the bandwidth to allow services to be delivered across the WAN from a remote datacenter site.

The influence of latency on design varies, based on the type of Unified Communications service considered for remote deployment. If a voice service is hosted across a WAN where the one-way latency is 200 ms, for example, users might experience issues such as delay-to-dialtone or increased media cut-through delays. For other services such as presence, there might be no problem with a 200 ms latency.

Reliability of the site's connectivity into the rest of the network is a fundamental consideration in determining the appropriate deployment model for any technology. When reliability is high, most Unified Communications components allow for the deployment of services hosted from a remote site; when reliability is inconsistent, some Unified Communications components might not perform reliably when hosted remotely; if the reliability is poor, co-location of the Unified Communications services at the site might be required.

High Availability Requirements

The high availability of services is always a design goal. Pragmatic design decisions are required when balancing the need for reliability and the cost of achieving it. The following elements all affect a design's ability to deliver high availability:

- Bandwidth reliability, directly affecting the deployment model for any Unified Communications service
- Power availability

  Power loss is a very disruptive event in any system, not only because it prevents the consumption of services while the power is out, but also because of the ripple effects caused by power restoration. A site with highly available power (for example, a site whose power grid connection is stable, backed-up by uninterruptible power supplies (UPSs) and by generator power) can typically be chosen to host any Unified Communications service. If a site has inconsistent power availability, it would not be judicious to use it as a hosting site.
- Environmental factors such as heat, humidity, vibration, and so forth
- Availability of qualified personnel

  Some Unified Communications services are delivered though the use of equipment such as servers that require periodical maintenance. Some Unified Communications functions such as the hosting of Unified Communications call agent servers are best deployed at sites staffed with qualified personnel.

Site-Based Design Guidance

Throughout this document, design guidance is organized along the lines of the various Unified Communications services and technologies. For instance, the call processing chapter contains not only the actual description of the call processing services, but also design guidance pertaining to deploying IP phones and Cisco Unified Communications servers based on a site's size, network connectivity, and high availability requirements. Likewise, the call admission control chapter focuses on the technical explanation of that technology while also incorporating site-based design considerations.

Generally speaking, most aspects of any given Unified Communications service or technology are applicable to all deployments, no matter the site's size or network connectivity. When applicable, site-based design considerations are called out. Services can be centralized, distributed, inter-networked, and geographically diversified.

Centralized Services

For applications where enterprise branch sites are geographically dispersed and interconnected over a Wide Area Network, the Cisco Unified Communications services can be deployed at a central location while serving endpoints over the WAN connections. For example, the call processing service can be deployed in a centralized manner, requiring only IP connectivity with the remote sites to deliver

telephony services. Likewise, voice messaging services, such as those provided by the Cisco Unity Connection platform, can also be provisioned centrally to deliver services to endpoints remotely connected across an IP WAN.

Centrally provisioned Unified Communications services can be impacted by WAN connectivity interruptions; for each service, the available local survivability options should be planned. As an example, the call processing service as offered by Cisco Unified CM can be configured with local survivability functionality such as SRST or Cisco Unified Communications Manager Express (Unified CME). Likewise, a centralized voice messaging service such as that of Cisco Unity Connection can be provisioned to allow remote sites operating under SRST or Unified CME to access voice messaging services at the central site, through the PSTN.

The centralization of services need not be uniform across all Unified Communications services. For example, a system can be deployed where multiple sites rely on a centralized call processing service, but can also be provisioned with a de-centralized (distributed) voice messaging service such as Cisco Unity Express. Likewise, a Unified Communications system could be deployed where call processing is provisioned locally at each site through Cisco Unified Communications Manager Express, with a centralized voice messaging service such as Cisco Unity Connection.

In many cases, the main criteria driving the design for each service are the availability and quality of the IP network between sites. The centralization of Unified Communications services offers advantages of economy of scale in both capital and operational expenses associated with the hosting and operation of equipment in situations where the IP connectivity between sites offers the following characteristics:

- Enough bandwidth for the anticipated traffic load, including peak hour access loads such as those generated by access to voicemail, access to centralized PSTN connectivity, and inter-site on-net communications including voice and video
- High availability, where the WAN service provider adheres to a Service Level Agreement to maintain and restore connectivity promptly
- Low latency, where local events at the remote site will not suffer if the round-trip time to the main central site imparts some delays to the system's response times

Also, when a given service is deployed centrally to serve endpoints at multiple sites, there are often advantages of feature transparency afforded by the use of the same processing resources for users at multiple sites. For example, when two sites are served by the same centralized Cisco Unified Communications Manager cluster, the users can share line appearances between the two sites. This benefit would not be available if each site were served by different (distributed) call processing systems.

These advantages of feature transparency and economies of scale should be evaluated against the relative cost of establishing and operating a WAN network configured to accommodate the demands of Unified Communications traffic.

Distributed Services

Unified Communications services can also be deployed independently over multiple sites, in a distributed fashion. For example, two sites (or more) can be provisioned with independent call processing Cisco Unified CME nodes, with no reliance on the WAN for availability of service to their co-located endpoints. Likewise, sites can be provisioned with independent voice messaging systems such as Cisco Unity Express.

The main advantage of distributing Unified Communications services lies in the independence of the deployment approach from the relative availability and cost of WAN connectivity. For example, if a company operates a site in a remote location where WAN connectivity is not available, is very expensive,

or is not reliable, then provisioning an independent call processing node such as Cisco Unified Communications Manager Express within the remote site will avoid any call processing interruptions if the WAN goes down.

Inter-Networking of Services

If two sites are provisioned with independent services, they can still be interconnected to achieve some degree of inter-site feature transparency. For example, a distributed call processing service provisioned through Cisco Unified Communications Manager Express can be inter-networked through H.323 or SIP trunks to permit IP calls between the sites. Likewise, separate instances of Cisco Unity Connection or Cisco Unity Express can partake in the same messaging network to achieve the routing of messages and the exchange of subscriber and directory information within a unified messaging network.

Geographical Diversity of Unified Communications Services

Some services can be provisioned in multiple redundant nodes across the IP WAN, allowing for continued service through site disruptions such as loss of power, network outages, or even compromises in the physical integrity of a site by events such as fire or earthquake.

To achieve such geographical diversity, the individual service must support redundant nodes as well as the deployment of these nodes across the latency and bandwidth constraints of the IP WAN. For example, the call processing service of Unified CM does support the deployment of a single cluster's call processing nodes across an IP WAN as long as the total end-to-end round-trip time between the nodes does not exceed 80 ms and an appropriate quantity of QoS-enabled bandwidth is provisioned. By contrast, Unified CME does not offer redundancy, and thus cannot be deployed in a geographically diverse configuration.

Table 5-2 summarizes the ability of each Cisco Unified Communications service to be deployed in the manners outlined above.

Table 5-2 Available Deployment Options for Cisco Unified Communications Services

Service	Centralized	Distributed	Inter-Networked	Geographical Diversity
Cisco Unified CM	Yes	Yes	Yes	Yes
Cisco Unified CME	No	Yes	Yes	No
Cisco Business Edition 6000	Yes	Yes	Yes	Yes
Cisco Business Edition 5000	Yes	Yes	Yes	No
Cisco Business Edition 3000	Yes	No	No	No
Cisco Unity Express	No	Yes	Yes, with Cisco Unified Messaging Gateway	No
Cisco Unity Connection	Yes	Yes (One Cisco Unity Connection per site)	Yes, with Cisco Unified Messaging Gateway	Yes
Cisco Emergency Responder	Yes	Yes (One Emergency Responder group per site)	Yes, through Emergency Responder clustering	Yes

Table 5-2 Available Deployment Options for Cisco Unified Communications Services (continued)

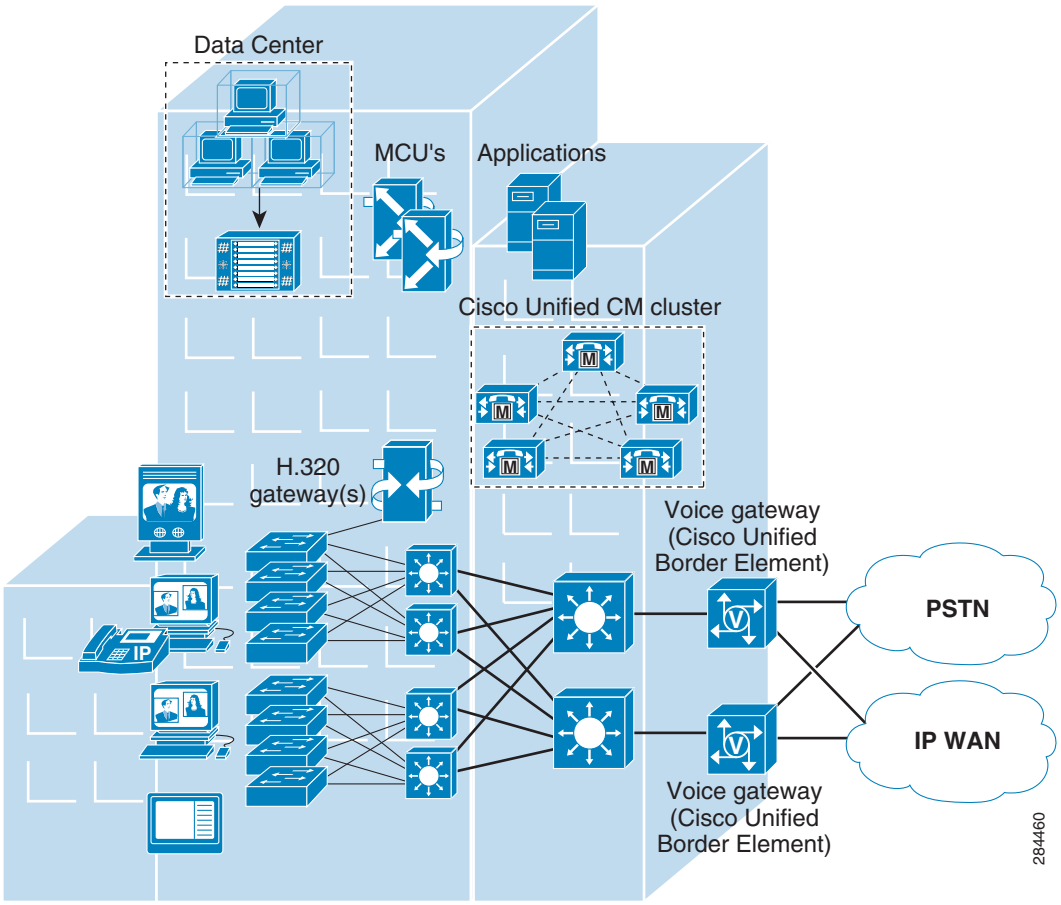
Service	Centralized	Distributed	Inter-Networked	Geographical Diversity
Cisco IM and Presence	Yes	Yes (one Cisco IM and Presence Service per site)	Yes, through inter-domain federation	Yes
Cisco Unified Mobility	Yes	Yes, as Unified CM Single Number Reach	No	Yes

Because call processing is a fundamental service, the basic call processing deployment models are introduced in this chapter. For a detailed technical discussion on Cisco Unified Communications Manager call processing, refer to the chapter on Call Processing, page 8-1.

Campus

In this call processing deployment model, the Unified Communications services and the endpoints are co-located in the campus, and the QoS-enabled network between the service nodes, the endpoints, and applications is considered highly available, offering virtually unlimited bandwidth with less than 15 ms of latency end-to-end. Likewise, the quality and availability of power are very high, and services are hosted in an appropriate data center environment. Communications between the endpoints traverses a LAN or a MAN, and communications outside the enterprise goes over an external network such as the PSTN. An enterprise would typically deploy the campus model over a single building or over a group of buildings connected by a LAN or MAN.

Figure 5-1 Example of a Campus Deployment



The campus model typically has the following design characteristics:

- Single Cisco Unified CM cluster. Some campus call processing deployments may require more than one Unified CM cluster, for instance, if scale calls for more endpoints than can be serviced by a single cluster or if a cluster needs to be dedicated to an application such as a call center.
- Alternatively for smaller deployments, Cisco Business Edition 3000, 5000, or 6000 may be deployed in the campus.
- Maximum of 40,000 configured and registered Skinny Client Control Protocol (SCCP) or Session Initiation Protocol (SIP) IP phones, Cisco Cius, video endpoints, mobile clients, and Cisco Virtualization Experience Clients (VXC) per Unified CM cluster.
- Maximum of 2,100 gateways and trunks (that is, the total number of H.323 gateways, H.323 trunks, digital MGCP devices, and SIP trunks) per Unified CM cluster.
- Trunks and/or gateways (IP or PSTN) for all calls to destinations outside the campus.
- Co-located digital signal processor (DSP) resources for conferencing, transcoding, and media termination point (MTP).
- Other Unified Communications services, such as messaging (voicemail), presence, and mobility are typically co-located.

- Interfaces to legacy voice services such as PBXs and voicemail systems are connected within the campus, with no operational costs associated with bandwidth or connectivity.
- Multipoint Control Unit (MCU) resources are required for multipoint video conferencing. Depending on conferencing requirements, these resources may be either SCCP or H.323, or both.
- H.323 and H.320 video gateways are needed to communicate with H.320 videoconferencing devices on the public ISDN network.
- High-bandwidth audio is available (for example, G.711 or G.722) between devices within the site.
- High-bandwidth video (for example, 384 kbps to 1.5 Mbps) is available between devices within the site.

Best Practices for the Campus Model

Follow these guidelines and best practices when implementing the single-site model:

- Ensure that the infrastructure is highly available, enabled for QoS, and configured to offer resiliency, fast convergence, and inline power.
- Know the calling patterns for your enterprise. Use the campus model if most of the calls from your enterprise are within the same site or to PSTN users outside your enterprise.
- Use G.711 codecs for all endpoints. This practice eliminates the consumption of digital signal processor (DSP) resources for transcoding, and those resources can be allocated to other functions such as conferencing and media termination points (MTPs).
- Implement the recommended network infrastructure for high availability, connectivity options for phones (in-line power), Quality of Service (QoS) mechanisms, and security. (See Network Infrastructure, page 3-1.)
- Follow the provisioning recommendations listed in the chapter on Call Processing, page 8-1.

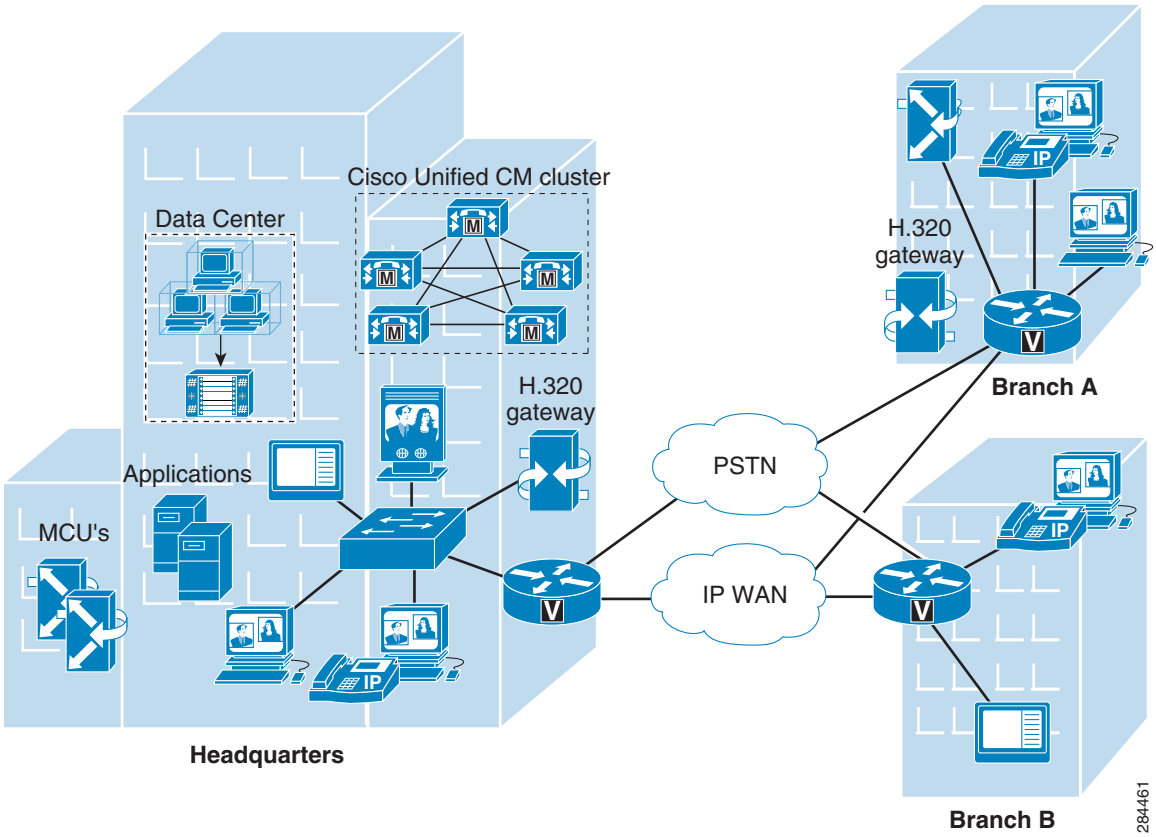
Multisite with Centralized Call Processing

In this call processing deployment model, endpoints are remotely located from the call processing service, across a QoS-enabled Wide Area Network. Due to the limited quantity of bandwidth available across the WAN, a call admission control mechanism is required to manage the number of calls admitted on any given WAN link, to keep the load within the limits of the available bandwidth. On-net communication between the endpoints traverses either a LAN/MAN (when endpoints are located in the same site) or a WAN (when endpoints are located in different sites). Communication outside the enterprise goes over an external network such as the PSTN, through a gateway or Cisco Unified Border Element (CUBE) session border controller (SBC) that can be co-located with the endpoint or at a different location (for example, when using a centralized gateway at the main site or when doing Tail End Hop Off (TEHO) across the enterprise network).

The IP WAN also carries call control signaling between the central site and the remote sites. Figure 5-2 illustrates a typical centralized call processing deployment, with a Unified CM cluster as the call processing agent at the central site and a QoS-enabled IP WAN to connect all the sites. In this deployment model, other Unified Communications services such as voice messaging, presence and mobility are often hosted at the central site as well to reduce the overall costs of administration and maintenance. In situations where the availability of the WAN is unreliable or when WAN bandwidth costs are high, it is possible to consider decentralizing some Unified Communications services such as voice messaging (voicemail) so that the service's availability is not impacted by WAN outages.

Note In each solution for the centralized call processing model presented in this document, the various sites connect to an IP WAN with QoS enabled.

Figure 5-2 Multisite Deployment with Centralized Call Processing



The multisite model with centralized call processing has the following design characteristics:

- Single Unified CM cluster. Some centralized call processing deployments may require more than one Unified CM cluster, for instance, if scale calls for more endpoints than can be serviced by a single cluster or if a cluster needs to be dedicated to an application such as a call center.
- For smaller deployments, Cisco Business Edition 3000 may be deployed in centralized call processing configurations for up to 9 remote sites.
- Cisco Business Edition 5000 may be deployed in centralized call processing configurations for up to 19 remote sites.
- Cisco Business Edition 6000 may be deployed in centralized call processing configurations for up to 49 remote sites.
- Maximum of 40,000 configured and registered Skinny Client Control Protocol (SCCP) or Session Initiation Protocol (SIP) IP phones, Cisco Cius, video endpoints, mobile clients, and Cisco Virtualization Experience Clients (VXC) per Unified CM cluster.
- Maximum of 2,000 locations or branch sites per Unified CM cluster.

- Maximum of 2,100 gateways and trunks (that is, the total number of H.323 gateways, H.323 trunks, digital MGCP devices, and SIP trunks) per Unified CM cluster.
- PSTN connectivity for all off-net calls.
- Digital signal processor (DSP) resources for conferencing, transcoding, and media termination point (MTP) are distributed locally to each site to reduce WAN bandwidth consumption on calls requiring DSPs.
- Capability to integrate with legacy private branch exchange (PBX) and voicemail systems. Interfaces to legacy voice services such as PBXs and voicemail systems can connected within the central site, with no operational costs associated with bandwidth or connectivity. Connectivity to legacy systems located at remote sites may require the operational expenses associated with the provisioning of extra WAN bandwidth.
- MCU resources are required for multipoint video conferencing. Depending on conferencing requirements, these resources may be either SCCP or H.323, or both, and may all be located at the central site or may be distributed to the remote sites if local conferencing resources are required.
- H.323/H.320 video gateways are needed to communicate with H.320 videoconferencing devices on the public ISDN network. These gateways may all be located at the central site or may be distributed to the remote sites if local ISDN access is required.
- The system allows for the automated selection of high-bandwidth audio (for example, G.711 or G.722) between devices within the site, while selecting low-bandwidth audio (for example, G.729) between devices in different sites.
- The system allows for the automated selection of high-bandwidth video (for example, 384 kbps to 1.5 Mbps) between devices in the same site, and low-bandwidth video (for example, 128 kbps) between devices at different sites.
- A minimum of 768 kbps or greater WAN link speed should be used when video is to be placed on the WAN.
- Call admission control is achieved through Enhanced Locations CAC or RSVP.
- For voice and video calls, automated alternate routing (AAR) provides the automated rerouting of calls through the PSTN when call admission control denies a call due to lack of bandwidth. AAR relies on a gateway being available to route the call from the calling phone toward the PSTN, and another gateway to accept the call from the PSTN at the remote site, to be connected to the called phone.
- Call Forward Unregistered (CFUR) functionality provides the automated rerouting of calls through the PSTN when an endpoint is considered unregistered due to a remote WAN link failure. CFUR relies on a gateway being available to route the call from the calling phone toward the PSTN, and another gateway to accept the call from the PSTN at the remote site, to be connected to the called phone.
- Survivable Remote Site Telephony (SRST) for video. SCCP video endpoints located at remote sites become audio-only devices if the WAN connection fails.
- Cisco Unified Communications Manager Express (Unified CME) may be used for remote site survivability instead of an SRST router.
- Cisco Unified Communications Manager Express (Unified CME) can be integrated with the Cisco Unity Connection server in the branch office or remote site. The Cisco Unity Connection server is registered to the Unified CM at the central site in normal mode and can fall back to Unified CME in SRST mode when Unified CM is not reachable, or during a WAN outage, to provide the users at the branch offices with access to their voicemail with MWI.

- As with other call processing types that support multisite centralized call processing, Cisco Business Edition 3000 allows PSTN routing through both central and remote site gateways. Providing a local gateway at remote sites for local PSTN breakout is a necessary requirement for countries providing emergency services for users located at remote sites. The local gateway at the remote site provides call routing to the local PSAP of the remote site location. Local PSTN breakout at remote sites might also be needed or required for countries having strict regulations requiring separation of IP telephony networks from the PSTN. Where regulations allow, local PSTN breakout through the remote site gateway can be used to enable toll bypass or tail-end hop off (TEHO). Business Edition 3000 provides country-based dial plan configuration to enable routing to configured PSTN gateways as well as policy mechanisms to control PSTN access restrictions (as applicable based on local country regulations). Business Edition 3000 supports local PSTN breakout only through the MGCP-controlled Cisco 2901 Integrated Services Router (ISR). Local breakout at a remote site can also be provided through analog trunks using a Cisco SPA8800 IP Telephony Gateway or through SIP trunks using Cisco Unified Border Element on a Cisco SPA8800 or SPA8900 IP Telephony Gateway (sometimes referred to as “CUBE Lite”).
- Business Edition 3000 does not support SRST or remote site survivability.

Connectivity options for the IP WAN include:

- Leased lines
- Frame Relay
- Asynchronous Transfer Mode (ATM)
- ATM and Frame Relay Service Inter-Working (SIW)
- Multiprotocol Label Switching (MPLS) Virtual Private Network (VPN)
- Voice and Video Enabled IP Security Protocol (IPSec) VPN (V3PN)

Routers that reside at the WAN edges require quality of service (QoS) mechanisms, such as priority queuing and traffic shaping, to protect the voice traffic from the data traffic across the WAN, where bandwidth is typically scarce. In addition, a call admission control scheme is needed to avoid oversubscribing the WAN links with voice traffic and deteriorating the quality of established calls. For centralized call processing deployments, *locations* (static or RSVP-enabled) configured within Unified CM provide call admission control. (Refer to the chapter on Call Admission Control, page 11-1, for more information on locations.)

A variety of Cisco gateways can provide the remote sites with TDM and/or IP-based PSTN access. When the IP WAN is down, or if all the available bandwidth on the IP WAN has been consumed, calls from users at remote sites can be rerouted through the PSTN. The Cisco Unified Survivable Remote Site Telephony (SRST) feature, available for both SCCP and SIP phones, provides call processing at the branch offices for Cisco Unified IP Phones if they lose their connection to the remote primary, secondary, or tertiary Unified CM or if the WAN connection is down. Cisco Unified SRST functionality is available on Cisco IOS gateways running the SRST feature or on Cisco Unified CME running in SRST mode. Unified CME running in SRST mode provides more features for the phones than SRST on a Cisco IOS gateway.

Best Practices for the Centralized Call Processing Model

Follow these guidelines and best practices when implementing multisite centralized call processing deployments:

- Minimize delay between Unified CM and remote locations to reduce voice cut-through delays (also known as clipping).
- Configure Enhanced Locations CAC or RSVP in Unified CM to provide call admission control into and out of remote branches. See the chapter on Call Admission Control, page 11-1, for details on how to apply this mechanism to the various WAN topologies.
- The number of IP phones and line appearances supported in Survivable Remote Site Telephony (SRST) mode at each remote site depends on the branch router platform, the amount of memory installed, and the Cisco IOS release. SRST on a Cisco IOS gateway supports up to 1,500 phones, while Unified CME running in SRST mode supports 450 phones. (For the latest SRST or Unified CME platform and code specifications, refer to the SRST and Unified CME documentation available at http://www.cisco.com.) Generally speaking, however, the choice of whether to adopt a centralized call processing or distributed call processing approach for a given site depends on a number of factors such as:
  - IP WAN bandwidth or delay limitations
  - Criticality of the voice network
  - Feature set needs
  - Scalability
  - Ease of management
  - Cost

  If a distributed call processing model is deemed more suitable for the customer's business needs, the choices include installing a Unified CM cluster at each site or running Unified CME at the remote sites.
- At the remote sites, use the following features to ensure call processing survivability in the event of a WAN failure:
  - For SCCP phones, use SRST on a Cisco IOS gateway or Unified CME running in SRST mode.
  - For SIP phones, use SIP SRST.
  - For MGCP phones, use MGCP Gateway Fallback.

  SRST or Unified CME in SRST mode, SIP SRST, and MGCP Gateway Fallback can reside with each other on the same Cisco IOS gateway.

Remote Site Survivability

When deploying Cisco Unified Communications across a WAN with the centralized call processing model, you should take additional steps to ensure that data and voice services at the remote sites are highly available. Table 5-3 summarizes the different strategies for providing high availability at the

remote sites. The choice of one of these strategies may depend on several factors, such as specific business or application requirements, the priorities associated with highly available data and voice services, and cost considerations.

Table 5-3 Strategies for High Availability at the Remote Sites

Strategy	High Availability for Data Services?	High Availability for Voice Services?
Redundant IP WAN links in branch router	Yes	Yes
Redundant branch router platforms + Redundant IP WAN links	Yes	Yes
Data-only ISDN backup + SRST or Unified CME	Yes	Yes
Data and voice ISDN backup	Yes	Yes (see rules below)
Cisco Unified Survivable Remote Site Telephony (SRST) or Unified CME in SRST mode	No	Yes

The first two solutions listed in Table 5-3 provide high availability at the network infrastructure layer by adding redundancy to the IP WAN access points, thus maintaining IP connectivity between the remote IP phones and the centralized Unified CM at all times. These solutions apply to both data and voice services, and are entirely transparent to the call processing layer. The options range from adding a redundant IP WAN link at the branch router to adding a second branch router platform with a redundant IP WAN link.

The third and forth solutions in Table 5-3 use an ISDN backup link to provide survivability during WAN failures. The two deployment options for ISDN backup are:

- Data-only ISDN backup

  With this option, ISDN is used for data survivability only, while SRST or Unified CME in SRST mode is used for voice survivability. Note that you should configure an access control list on the branch router to prevent traffic from telephony signaling protocols such as Skinny Client Control Protocol (SCCP), H.323, Media Gateway Control Protocol (MGCP), or Session Initiation Protocol (SIP) from entering the ISDN interface, so that signaling from the IP phones does not reach the Unified CM at the central site. This is to ensure that the telephony endpoints located at the branch detect the WAN's failure and rely on local SRST resources.

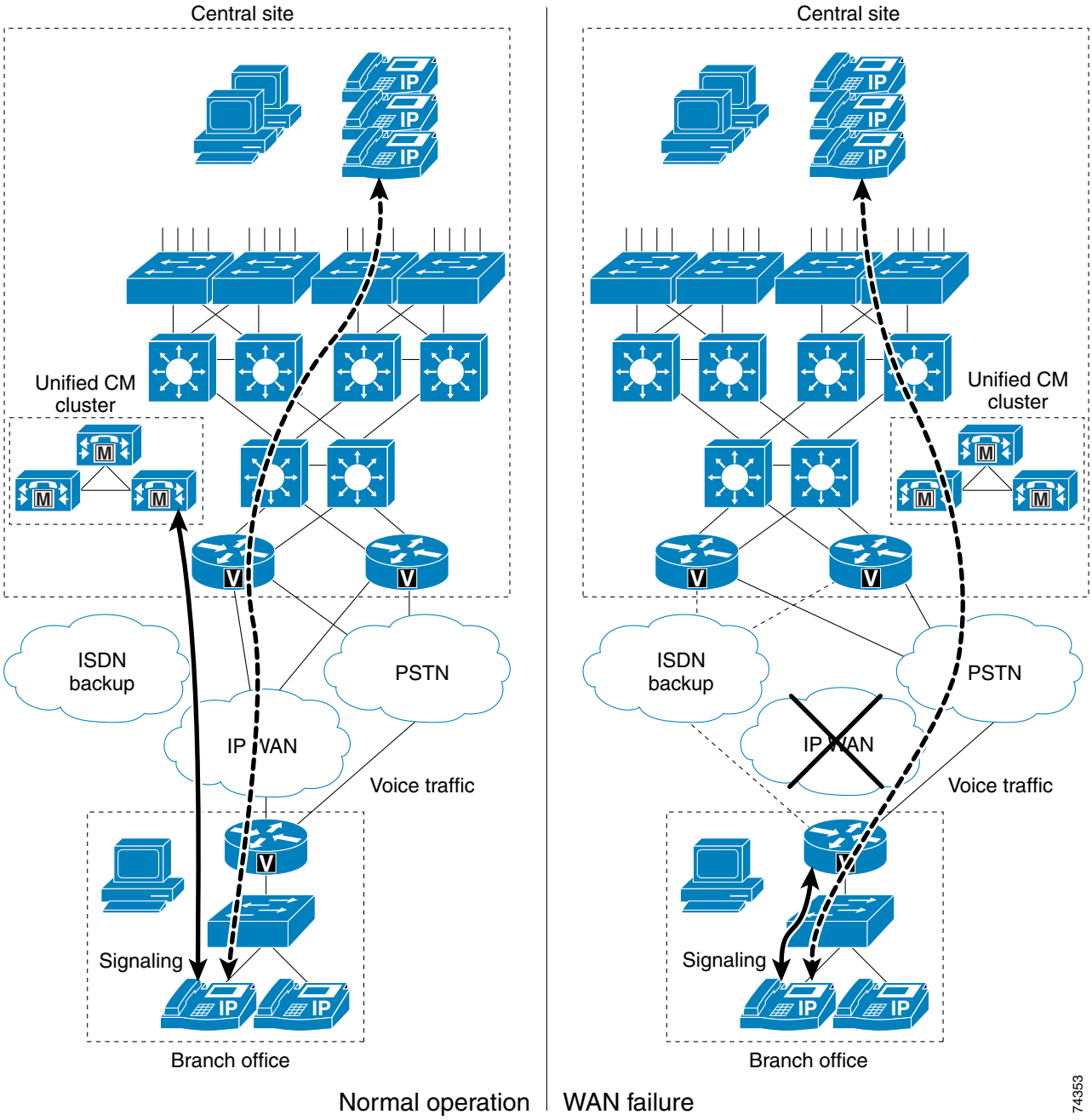
- Data and voice ISDN backup

  With this option, ISDN is used for both data and voice survivability. In this case, SRST or Unified CME in SRST mode is not used because the IP phones maintain IP connectivity to the Unified CM cluster at all times. However, Cisco recommends that you use ISDN to transport data and voice traffic only if all of the following conditions are true:

  - The bandwidth allocated to voice traffic on the ISDN link is the same as the bandwidth allocated to voice traffic on the IP WAN link.
  - The ISDN link bandwidth is fixed.
  - All the required QoS features have been deployed on the router's ISDN interfaces. Refer to the chapter on Network Infrastructure, page 3-1, for more details on QoS.

The fifth solution listed in Table 5-3, Survivable Remote Site Telephony (SRST) or Unified CME in SRST mode, provides high availability for voice services only, by providing a subset of the call processing capabilities within the remote office router and enhancing the IP phones with the ability to "re-home" to the call processing functions in the local router if a WAN failure is detected. Figure 5-3 illustrates a typical call scenario with SRST or Unified CME in SRST mode.

Figure 5-3 Survivable Remote Site Telephony (SRST) or Unified CME in SRST Mode



Under normal operations shown in the left part of Figure 5-3, the branch office connects to the central site via an IP WAN, which carries data traffic, voice traffic, and call signaling. The IP phones at the branch office exchange call signaling information with the Unified CM cluster at the central site and place their calls across the IP WAN. The branch router or gateway forwards both types of traffic (call signaling and voice) transparently and has no knowledge of the IP phones.

If the WAN link to the branch office fails, or if some other event causes loss of connectivity to the Unified CM cluster, the branch IP phones re-register with the branch router in SRST mode. The branch router, SRST, or Unified CME running in SRST mode, queries the IP phones for their configuration and

uses this information to build its own configuration automatically. The branch IP phones can then make and receive calls either within the branch's network or through the PSTN. The phone displays the message “Unified CM fallback mode,” and some advanced Unified CM features are unavailable and are grayed out on the phone display.

When WAN connectivity to the central site is reestablished, the branch IP phones automatically re-register with the Unified CM cluster and resume normal operation. The branch SRST router deletes its information about the IP phones and reverts to its standard routing or gateway configuration. Unified CME running in SRST mode at the branch can choose to save the learned phone and line configuration to the running configuration on the Unified CME router by using the auto-provision option. If **auto-provision none** is configured, none of the auto-provisioned phone or line configuration information is written to the running configuration of the Unified CME router. Hence, no configuration change is required on Unified CME if the IP phone is replaced and the MAC address changes.

Note When WAN connectivity to the central site is reestablished, or when Unified CM is reachable again, phones in SRST mode with active calls will not immediately re-register to Unified CM until those active calls are terminated.

Note The remote site survivability features explained above are not supported with Business Edition 3000.

Unified CME in SRST Mode

When Unified CME is used in SRST mode, it provides more call processing features for the IP phones than are available with the SRST feature on a router. In addition to the SRST features such as call preservation, auto-provisioning, and failover, Unified CME in SRST mode also provides most of the Unified CME telephony features for the SCCP phones, including:

- Paging
- Conferencing
- Hunt groups
- Basic automatic call distribution (B-ACD)
- Call park, call pickup, call pickup groups
- Overlay-DN, softkey templates
- Cisco IP Communicator
- Cisco Unified Video Advantage
- Integration with Cisco Unity with MWI support at remote sites, with distributed Microsoft Exchange or IBM Lotus Domino server

Unified CME in SRST mode provides call processing support for SCCP phones in case of a WAN failure. However, Unified CME in SRST mode does not provide fallback support for MGCP phones or endpoints. To enable SIP and MGCP phones to fall back if they lose their connection to the SIP proxy server or Unified CM, or if the WAN connection fails, you can additionally configure both the SIP SRST feature and the MGCP Gateway Fallback feature on the same Unified CME server running as the SRST fallback server.

Best Practices for Unified CME in SRST Mode

- Use the Unified CME IP address as the IP address for SRST reference in the Unified CM configuration.
- The Connection Monitor Duration is a timer that specifies how long phones monitor the WAN link before initiating a fallback from SRST to Unified CM. The default setting of 120 seconds should be used in most cases. However, to prevent phones in SRST mode from falling back and re-homing to Unified CM with flapping links, you can set the Connection Monitor Duration parameter on Unified CM to a longer period so that phones do not keep registering back and forth between the SRST router and Unified CM. Do not set the value to an extensively longer period because this will prevent the phones from falling back from SRST to Unified CM for a long amount of time.
- Phones in SRST fallback mode will not re-home to Unified CM when they are in active state.
- Phones in SRST fallback mode revert to non-secure mode from secure conferencing.
- Configure **auto-provision none** to prevent writing any learned ephone-dn or ephone configuration to the running configuration of the Unified CME router. This eliminates the need to change the configuration if the IP phone is replaced or the MAC address changes.

For more information on using Unified CME in SRST mode, refer to the *Cisco Unified Communications Manager Express System Administrator Guide*, available at

http://www.cisco.com/en/US/products/sw/voicesw/ps4625/products_installation_and_configuration_guides_list.html

For more information on SIP SRST, refer to the *Cisco Unified SIP SRST System Administrator Guide*, available at

http://www.cisco.com/en/US/products/sw/voicesw/ps2169/products_installation_and_configuration_guides_list.html

For more information on MGCP Gateway fallback, refer to the information on MGCP gateway fallback in the *Cisco CallManager and Cisco IOS Interoperability Guide*, available at

http://www.cisco.com/en/US/docs/ios/12_3/vvf_c/interop/ccm_c.html

Best Practices for SRST Router

Use a Cisco Unified SRST router, rather than Unified CME in SRST mode, for the following deployment scenarios:

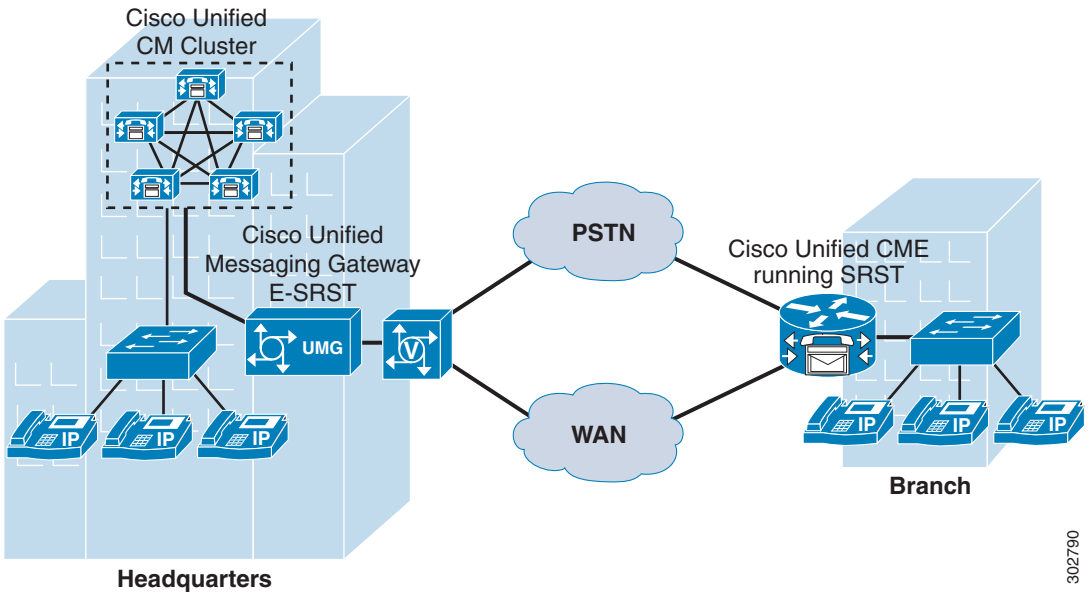
- For supporting a maximum of 1,500 phones on a single SRST router. (Unified CME in SRST mode supports a maximum of 450 phones.)
- For up to 3,000 phones, use two SRST routers. Dial plans must be properly configured to route the calls back and forth between the SRST routers.
- For simple, one-time configuration of basic SRST functions.
- For SRTP media encryption, which is available only in Cisco Unified SRST (Secure SRST).
- For support of the Cisco VG248 Voice Gateway.

For routing calls to and from phones that are unreachable or not registered to the SRST router, use the **alias** command.

Cisco Unified Survivable Remote Site Telephony Manager

Cisco Unified Survivable Remote Site Telephony (SRST) Manager simplifies the deployment of Cisco Unified CME running SRST as well as traditional SRST in the branch. (See Figure 5-4.) It is Linux-based software running inside a virtual machine on Cisco supported virtualized platforms (for example, Cisco UCS). Cisco Unified SRST Manager supports only the centralized call processing deployment model, where the Cisco Unified CM cluster runs in the central location. Cisco Unified SRST Manager can be deployed in the central location along with the Cisco Unified CM cluster or in the remote branch location. Figure 5-4 illustrates the deployment of Cisco Unified SRST Manager in the central location. During normal operation, Cisco Unified SRST Manager regularly retrieves configurations (for example, calling search space, partition, hunt group, call park, call pickup, and so forth, if configured) from Cisco Unified CM and uploads them to provision the branch router with similar functionality for use in SRST mode. Thus, Cisco Unified SRST Manager reduces manual configuration required in the branch SRST router and enables users to have a similar calling experience in both SRST and normal modes.

Figure 5-4 Cisco Unified Survivable Remote Site Telephony Manager Deployed in the Central Location



Cisco Unified SRST Manager consumes bandwidth from the WAN link when uploading the Unified CM configurations to provision the branch router. The Cisco Unified SRST Manager software does not perform packet marking, therefore the Cisco Unified SRST Manager traffic will travel as best-effort on the network. Cisco recommends maintaining this best-effort marking, which is IP Precedence 0 (DSCP 0 or PHB BE), to ensure that it does not interfere with real-time high priority voice traffic. To ensure that Cisco Unified SRST Manager traffic does not cause congestion and to reduce the chances of packet drop, Cisco recommends scheduling the configuration upload to take place during non-peak hours (for example, in the evening hours or during the weekend). The configuration upload schedule can be set from the Cisco Unified SRST Manager web interface.

Consider the following guidelines when you deploy Cisco Unified SRST Manager:

- Cisco Unified SRST Manager is not supported with the Cisco Unified Communications 500 Series platform or the Cisco Business Edition 3000 and 5000 platforms.
- The branch voice gateway must be co-resident with (reside on) the SRST router.
- There is no high availability support with Cisco Unified SRST Manager. If Cisco Unified SRST Manager is unavailable, configuration upload is not possible.
- Cisco Unified SRST Manager is not supported in deployments where NAT is used between the headquarters and branch locations.

Voice Over the PSTN as a Variant of Centralized Call Processing

Centralized call processing deployments can be adapted so that inter-site voice media is sent over the PSTN instead of the WAN. With this configuration, the signaling (call control) of all telephony endpoints is still controlled by the central Unified CM cluster, therefore this Voice over the PSTN (VoPSTN) model variation still requires a QoS-enabled WAN with appropriate bandwidth configured for the signaling traffic.

You can implement VoPSTN in one of the following ways:

- Using the automated alternate routing (AAR) feature. (For more information on AAR, see the section on Automated Alternate Routing, page 9-117.)
- Using a combination of dial plan constructs in both Unified CM and the PSTN gateways.

VoPSTN can be an attractive option in deployments where IP WAN bandwidth is either scarce or expensive with respect to PSTN charges, or where IP WAN bandwidth upgrades are planned for a later date but the Cisco Unified Communications system is already being deployed.

Note VoPSTN deployments offer basic voice functionality that is a reduced subset of the Unified CM feature set.

In particular, regardless of the implementation choice, the system designer should address the following issues, among others:

- Centralized voicemail requires:
  - A telephony network provider that supports redirected dialed number identification service (RDNIS) end-to-end for all locations that are part of the deployment. RDNIS is required so that calls redirected to voicemail carry the redirecting DN, to ensure proper voicemail box selection.
  - If the voicemail system is accessed through an MGCP gateway, the voicemail pilot number must be a fully qualified E.164 number.
- The Extension Mobility feature is limited to IP phones contained within a single branch site.
- All on-net (intra-cluster) calls will be delivered to the destination phone with the same call treatment as an off-net (PSTN) call. This includes the quantity of digits delivered in the call directories such as Missed Calls and Received Calls.
- Each inter-branch call generates two independent call detail records (CDRs): one for the call leg from the calling phone to the PSTN, and the other for the call leg from the PSTN to the called phone.
- There is no way to distinguish the ring type for on-net and off-net calls.
- All destination phones require a fully qualified Direct Inward Dial (DID) PSTN number that can be called directly. Non-DID DNs cannot be reached directly from a different branch site.

- With VoPSTN, music on hold (MoH) is limited to cases where the holding party is co-located with the MoH resource. If MoH servers are deployed at the central site, then only calls placed on hold by devices at the central site will receive the hold music.
- Transfers to a destination outside the branch site will result in the hairpinning of the call through the branch's gateway. Traffic engineering of the branch's gateway resources must be adjusted accordingly.
- Call forwarding of any call coming into the branch's gateway to a destination outside the branch site will result in hairpinning of the call through the gateway, thus using two trunk ports. This behavior applies to:
  - Calls forwarded to a voicemail system located outside the branch
  - Calls forwarded to an on-net abbreviated dialing destination located in a different branch

  The gateway port utilization resulting from these call forwarding flows should be taken into account when sizing the trunks connecting the branch to the PSTN.
- Conferencing resources must be co-located with the phone initiating the conference.
- VoPSTN does not support applications that require streaming of IP audio from the central site (that is, not traversing a gateway). These applications include, but are not limited to:
  - Centralized music on hold (MoH) servers
  - Interactive Voice Response (IVR)
  - CTI-based applications
- Use of the Attendant Console outside of the central site can require a considerable amount of bandwidth if the remote sites must access large user account directories without caching them.
- Because all inter-branch media (including transfers) are sent through the PSTN, the gateway trunk group must be sized to accommodate all inter-branch traffic, transfers, and centralized voicemail access.
- Cisco recommends that you do not deploy shared lines across branches, such that the devices sharing the line are in different branches.

In addition to these general considerations, the following sections present recommendations and issues specific to each of the following implementation methods:

- VoPSTN Using AAR, page 5-20
- VoPSTN Using Dial Plan, page 5-22

VoPSTN Using AAR

This method consists of configuring the Unified CM dial plan as in a traditional centralized call processing deployment, with the automated alternate routing (AAR) feature also properly configured. AAR provides transparent re-routing over the PSTN of inter-site calls when the locations mechanism for call admission control determines that there is not enough available WAN bandwidth to accept an additional call.

To use the PSTN as the primary (and only) voice path, you can configure the call admission control bandwidth of each location (branch site) to be 1 kbps, thus preventing *all* calls from traversing the WAN. With this configuration, all inter-site calls trigger the AAR functionality, which automatically re-routes the calls over the PSTN.

The AAR implementation method for VoPSTN offers the following benefits:

- An easy migration path to a complete Cisco Unified Communications deployment. When bandwidth becomes available to support voice media over the WAN, the dial plan can be maintained intact, and the only change needed is to update the location bandwidth value for each site.
- Support for some supplementary features, such as callback on busy.

In addition to the general considerations listed for VoPSTN, the following design guidelines apply to the AAR implementation method:

- AAR functionality must be configured properly.
- As a general rule, supported call initiation devices include IP phones, gateways, and line-side gateway-driven analog phones.
- Inter-branch calls can use AAR only if the destination devices are IP phones or Cisco Unity ports.
- Inter-branch calls to other endpoints must use a fully qualified E.164 number.
- All on-net, inter-branch calls will display the message, "Network congestion, rerouting."
- If destination phones become unregistered (for example, due to WAN connectivity interruption), AAR functionality will not be invoked and abbreviated dialing will be possible only if Call Forward Unregistered (CFUR) is configured. If the destination phone has registered with an SRST router, then it can also be reached by directly dialing its PSTN DID number.
- If originating phones become unregistered (for example, due to WAN connectivity interruption), they will go into SRST (or Unified CME as SRST) mode. To preserve abbreviated dialing functionality under these conditions, configure the SRST (or Unified CME as SRST) router with an appropriate set of translation rules to match the abbreviated dialing form of the destination and translate it into the form required by the PSTN to route calls to the destination.
- Shared lines within the same branch should be configured in a partition included only in that branch's calling search spaces. Inter-site access to the shared line requires one of the following:
  - The originating site dials the DID number of the shared line.
  - If inter-site abbreviated dialing to the shared line is desired, use a translation pattern that expands the user-dialed abbreviated string to the DID number of the shared line.

Note In this case, direct dialing of the shared line's DN from another branch would trigger multiple AAR-based PSTN calls.

VoPSTN Using Dial Plan

This method relies on a specific dial plan configuration within Unified CM and the PSTN gateways to route all inter-site calls over the PSTN. The dial plan must place IP phone DN's at each site into a different partition, and their calling search space must provide access only to the site's internal partition and a set of route patterns that point to the local PSTN gateway.

Abbreviated inter-site dialing can still be provided via a set of translations at each branch site, one for each of the other branch sites. These translations are best accomplished with H.323 gateways and translation rules within Cisco IOS.

The dial plan method for implementing VoPSTN offers the following benefits:

- Easier configuration because AAR is not needed.
- Abbreviated dialing automatically works even under WAN failure conditions on either the originating or destination side, because the Cisco IOS translation rules within the H.323 gateway are effective in SRST mode.

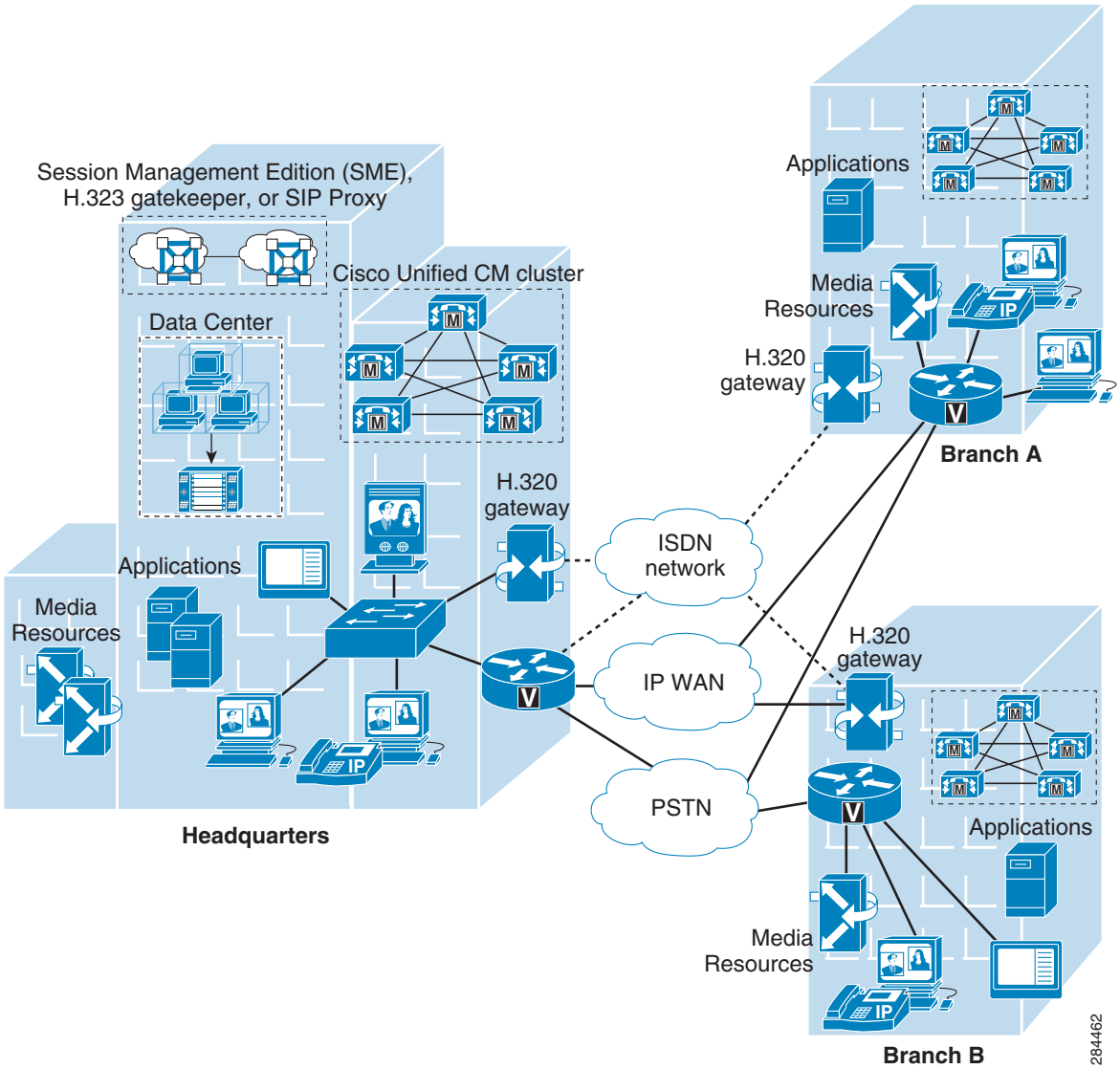
In addition to the general considerations listed for VoPSTN, the following design guidelines apply to the dial plan implementation method:

- There is no support for supplementary features such as callback on busy.
- Some CTI-based applications do not support overlapping extensions (that is, two or more phones configured with the same DN, although in different partitions).
- There is no easy migration to a complete Cisco Unified Communications deployment because the dial plan needs to be redesigned.

Multisite with Distributed Call Processing

The model for a multisite deployment with distributed call processing consists of multiple independent sites, each with its own call processing agent cluster connected to an IP WAN that carries voice traffic between the distributed sites. Figure 5-5 illustrates a typical distributed call processing deployment.

Figure 5-5 Multisite Deployment with Distributed Call Processing



Each site in the distributed call processing model can be one of the following:

- A single site with its own call processing agent, which can be either:
  - Cisco Unified Communications Manager (Unified CM)
  - Cisco Business Edition 5000 and Business Edition 6000
  - Cisco Unified Communications Manager Express (Unified CME)

  - Other IP PBX
- A centralized call processing site and all of its associated remote sites
- A legacy PBX with Voice over IP (VoIP) gateway

The multisite model with distributed call processing has the following design characteristics:

- Maximum of 40,000 configured and registered Skinny Client Control Protocol (SCCP) or Session Initiation Protocol (SIP) IP phones, Cisco Cius, video endpoints, mobile clients, and Cisco Virtualization Experience Clients (VXC) per Unified CM cluster.
- Maximum of 2,100 gateways and trunks (that is, the total number of H.323 gateways, H.323 trunks, digital MGCP devices, and SIP trunks) per Unified CM cluster.
- TDM or IP-based PSTN for all external calls.
- Digital signal processor (DSP) resources for conferencing, transcoding, and media termination point (MTP) are distributed locally to each site to reduce WAN bandwidth consumption on calls requiring DSPs.
- Voicemail, unified messaging, and Cisco IM and Presence components.
- Capability to integrate with legacy private branch exchange (PBX) and voicemail systems.
- Cisco Unified Communications Manager Session Management Edition (SME) clusters, H.323 gatekeepers, or Session Initiation Protocol (SIP) proxy servers can be used to provide intercluster call routing and dial plan aggregation in multisite distributed call processing deployments.
- MCU resources are required in each cluster for multipoint video conferencing. Depending on conferencing requirements, these resources may be either SCCP or H.323, or both, and may all be located at the regional sites or may be distributed to the remote sites of each cluster if local conferencing resources are required.
- H.323/H.320 video gateways are needed to communicate with H.320 videoconferencing devices on the public ISDN network. These gateways may all be located at the regional sites or may be distributed to the remote sites of each cluster if local ISDN access is required.
- High-bandwidth audio (for example, G.711 or G.722) between devices in the same site, but low-bandwidth audio (for example, G.729) between devices in different sites.
- High-bandwidth video (for example, 384 kbps to 1.5 Mbps) between devices in the same site, but low-bandwidth video (for example, 128 kbps) between devices at different sites.
- Minimum of 768 kbps or greater WAN link speeds. Video is *not* recommended on WAN connections that operate at speeds lower than 768 kbps.
- Call admission control is achieved through Enhanced Locations CAC or RSVP.

An IP WAN interconnects all the distributed call processing sites. Typically, the PSTN serves as a backup connection between the sites in case the IP WAN connection fails or does not have any more available bandwidth. A site connected only through the PSTN is a standalone site and is not covered by the distributed call processing model. (See Campus, page 5-7.)

Connectivity options for the IP WAN include:

- Leased lines
- Frame Relay
- Asynchronous Transfer Mode (ATM)
- ATM and Frame Relay Service Inter-Working (SIW)
- Multiprotocol Label Switching (MPLS) Virtual Private Network (VPN)
- Voice and Video Enabled IP Security Protocol (IPSec) VPN (V3PN)

Best Practices for the Distributed Call Processing Model

A multisite deployment with distributed call processing has many of the same requirements as a single site or a multisite deployment with centralized call processing. Follow the best practices from these other models in addition to the ones listed here for the distributed call processing model. (See Campus, page 5-7, and Multisite with Centralized Call Processing, page 5-9.)

Cisco Unified Communications Manager Session Management Edition clusters, H.323 gatekeepers, or Session Initiation Protocol (SIP) proxy servers can be used to provide intercluster call routing and dial plan aggregation in multisite distributed call processing deployments. The following best practices apply to the use of these dial plan aggregation devices:

Unified CM Session Management Edition Clusters

Cisco Unified Communications Manager Session Management Edition is commonly used for intercluster call routing and dial plan aggregation in distributed call processing deployments. Intercluster call routing can be number based using standard numeric route patterns and/or URI based using the Intercluster Look-up Service (ILS). Unified CM Session Management Edition supports multiple protocols (SIP, H.323, MGCP, and SCCP), has sophisticated trunk and digit manipulation features, supports Enhanced Locations CAC and RSVP, and uses the same code and user interface as Unified CM. Unified CM Session Management Edition cluster deployments typically consist of many trunks and no (or very few) Unified Communications endpoints. Unified CM Session Management Edition clusters can use all of the high availability features (such as clustering over the WAN, CallManager Groups, and Run on all Unified CM Nodes) that are available to Unified CM clusters.

For detailed information on Unified CM Session Management Edition cluster deployments, refer to the *Cisco Unified Communizations Manager Session Management Edition Deployment Guide*, available at

http://www.cisco.com/en/US/products/ps10661/products_implementation_design_guides_list.html

Gatekeeper Deployments

- Cisco IOS gatekeepers can be used to provide call admission control into and out of each site.
- To provide high availability of the gatekeeper, use Hot Standby Router Protocol (HSRP) gatekeeper pairs, gatekeeper clustering, and alternate gatekeeper support. In addition, use multiple gatekeepers to provide redundancy within the network. (See Gatekeeper Design Considerations, page 8-37.)
- Size the platforms appropriately to ensure that performance and capacity requirements can be met.
- Use only one type of codec on the WAN because the H.323 specification does not allow for Layer 2, IP, User Data Protocol (UDP), or Real-time Transport Protocol (RTP) header overhead in the bandwidth request. (Header overhead is allowed only in the payload or encoded voice part of the packet.) Using one type of codec on the WAN simplifies capacity planning by eliminating the need to over-provision the IP WAN to allow for the worst-case scenario.

For more information on the various functions performed by gatekeepers, refer to the following sections:

- For gatekeeper call admission control, see Call Admission Control, page 11-1.
- For gatekeeper scalability and redundancy, see Call Processing, page 8-1.
- For gatekeeper dial plan resolution, see Dial Plan, page 9-1.

SIP Proxy Depoloyments

SIP proxies such as the Cisco Unified SIP Proxy provide call routing and SIP signaling normalization.

The following best practices apply to the use of SIP proxies:

- Provide adequate redundancy for the SIP proxies.
- Ensure that the SIP proxies have the capacity for the call rate and number of calls required in the network.
- Planning for call admission control is outside the scope of this document.

Call Processing Agents for the Distributed Call Processing Model

Your choice of call processing agent will vary, based on many factors. The main factors, for the purpose of design, are the size of the site and the functionality required.

For a distributed call processing deployment, each site has its own call processing agent. The design of each site varies with the call processing agent, the functionality required, and the fault tolerance required. For example, in a site with 500 phones, a Unified CM cluster containing two servers can provide one-to-one redundancy, with the backup server being used as a publisher and Trivial File Transfer Protocol (TFTP) server.

The requirement for IP-based applications also greatly affects the choice of call processing agent because only Unified CM provides the required support for many Cisco IP applications.

Table 5-4 lists recommended call processing agents.

Table 5-4 Recommended Call Processing Agents

Call Processing Agent	Recommended Size	Comments
Cisco Unified Communications Manager Express (Unified CME)	Up to 450 phones	• For small remote sites • Capacity depends on Cisco IOS platform
Cisco Business Edition 5000	Up to 575 phones	• For small sites • Supports centralized or distributed call processing
Cisco Business Edition 6000	Up to 1,200 phones	• For small to medium sites • Supports centralized or distributed call processing
Cisco Unified Communications Manager (Unified CM)	50 to 40,000 phones	• Small to large sites, depending on the size of the Unified CM cluster • Supports centralized or distributed call processing
Legacy PBX with VoIP gateway	Depends on PBX	• Number of IP WAN calls and functionality depend on the PBX-to-VoIP gateway protocol and the gateway platform

Unified CM Session Management Edition

Unified Communications deployments using Cisco Unified Communications Manager Session Management Edition are a variation of the multisite distributed call processing deployment model and are typically employed to interconnect large numbers of unified communications systems through a single front-end system, in this case the Unified CM Session Management Edition. This section discusses the relevant design considerations for deploying Unified CM Session Management Edition.

Cisco Unified CM Session Management Edition is essentially a Unified CM cluster with trunk interfaces only and no IP endpoints. It enables aggregation of multiple unified communications systems, referred to as leaf systems.

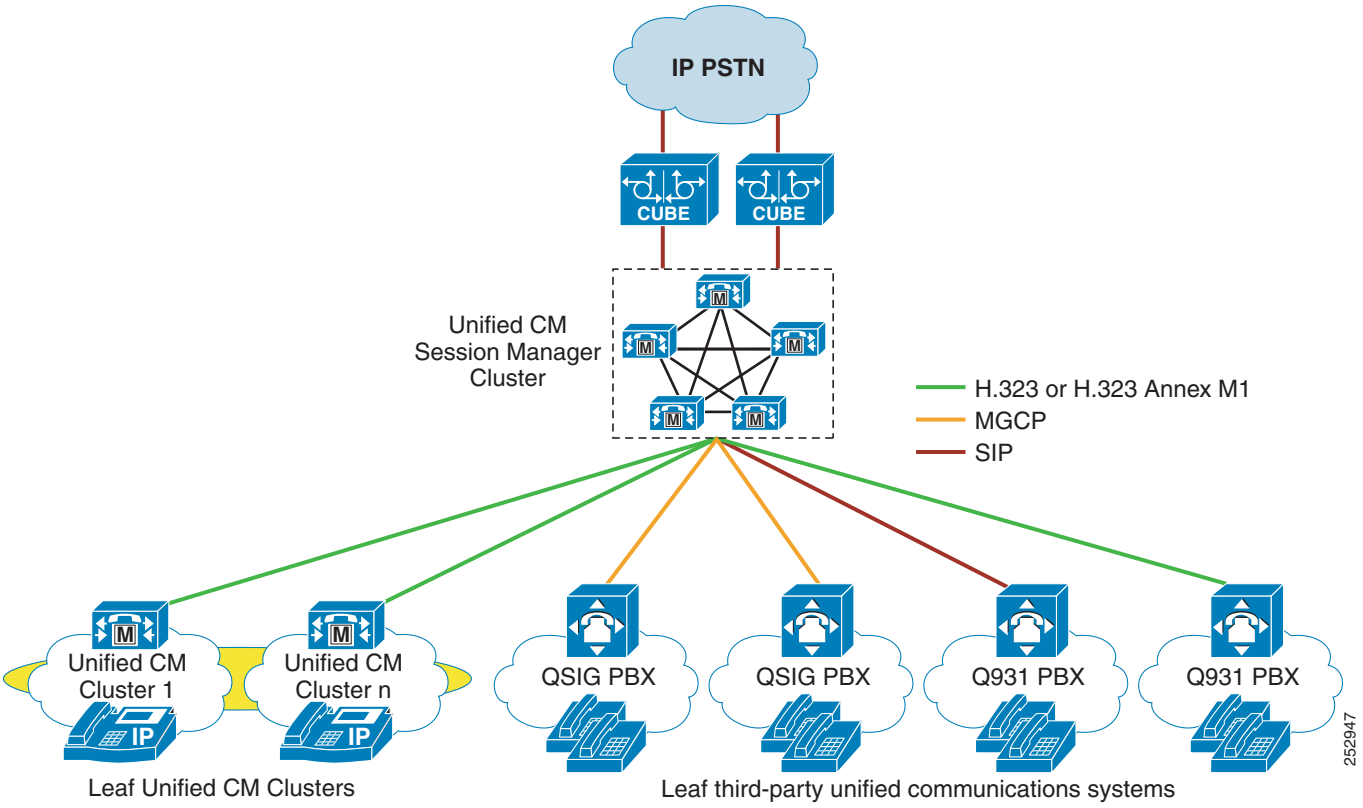
Session Management Edition deployments can be used to migrate a deployment of multiple PBXs and associated phones to a Unified CM cluster with IP phones and relatively few trunks. The Session Management Edition cluster may start with a large number of trunks interconnecting third-party PBXs; and migrate over time to a Unified CM cluster deployment with thousands of IP phones.

With Cisco Unified CM 8.0 and later releases, Unified CM Session Management Edition supports the following features:

- H.323 Annex M1 intercluster trunks
- SIP intercluster trunks
- SIP trunks
- H.323 trunks
- MGCP trunks
- Voice calls
- Video calls
- Encrypted calls
- Fax calls

Unified CM Session Management Edition may also be used to connect to third-party unified communications systems such as PSTN connections, PBXs, and centralized unified communications applications. (See Figure 5-6.) However, as with any standard Unified CM cluster, third-party connections to Unified CM Session Management Edition should be system tested for interoperability prior to use in a production environment.

Figure 5-6 Multisite Deployment with Unified CM Session Management Edition



When to Deploy Unified CM Session Management Edition

Cisco recommends deploying Unified CM Session Management Edition if you want to do any of the following:

- Create and manage a centralized dial plan

  Rather than configuring each unified communications system with a separate dial plan and trunks to connect to all the other unified communications systems, Unified CM Session Management Edition allows you to configure the leaf unified communications systems with a simplified dial plan and trunk(s) pointing to the Session Management cluster. Unified CM Session Management Edition holds the centralized dial plan and corresponding reachability information about all the other unified communications systems.

- Provide centralized PSTN access

  Unified CM Session Management Edition can be used to aggregate PSTN access to one (or more) centralized PSTN trunks. Centralized PSTN access is commonly combined with the reduction, or elimination, of branch-based PSTN circuits.

- Centralize applications

  The deployment of a Unified CM Session Management Edition enables commonly used applications such as conferencing or videoconferencing to connect directly to the Session Management cluster, thus reducing the overhead of managing multiple trunks to leaf systems.

- Aggregate PBXs for migration to a Unified Communications system

  Unified CM Session Management Edition can provide an aggregation point for multiple PBXs as part of the migration from legacy PBXs to a Cisco Unified Communications System.

Differences Between Unified CM Session Management Edition and Standard Unified CM Clusters

The Unified CM Session Management Edition software is exactly the same as Unified CM. However, the software has been enhanced significantly to satisfy the requirements and the constraints of this new deployment model. Unified CM Session Management Edition is designed to support a large number of trunk-to-trunk connections, and as such it is subject to the following design considerations:

- Capacity

  It is important to correctly size the Unified CM Session Management cluster based on the expected BHCA traffic load between leaf Unified Communications systems (for example, Unified CM clusters and PBXs), to and from any centralized PSTN connections, and to any centralized applications. Determine the average BHCA and Call Holding Time for users of your Unified Communications system and share this information with your Cisco account Systems Engineer (SE) or Cisco Partner to size your Unified CM Session Management Edition cluster correctly.

- Trunks

  Where possible, avoid the use of static MTPs on Unified CM trunks (do not enable **MTP required** on the SIP or H.323 trunks of leaf Unified CM or Unified CM Session Management Edition clusters). Trunks that do not use “MTP required” offer more codec choices; support voice, video, and encryption; and do not anchor trunk calls to MTP resources. Dynamically inserted MTPs can be used on trunks (for example, for DTMF translation from in-band to out-of-band). If SIP Early Offer is required by a third-party unified communications system, use either the “Early Offer support for voice and video calls (insert MTP if needed)” on Unified CM SIP trunks or the Delayed Offer to Early Offer feature with Cisco Unified Border Element.

- Unified CM versions

  Both the Unified CM Session Management Edition and Unified CM leaf clusters should be deployed with Cisco Unified CM 7.1(2) or later release. Cisco Unified CM 8.5 or later release is recommended because those versions include features that improve and simplify call routing through Unified CM and Session Management Edition clusters. Earlier versions of Unified CM can be deployed but might experience problems that can be resolved only by upgrading your cluster to Unified CM 7.1(2) or later release.

- Interoperability

  Even though most vendors do conform to standards, differences can and do exist between protocol implementations from various vendors. As with any standard Unified CM cluster, Cisco strongly recommends that you conduct end-to-end system interoperability testing with any unverified third-party unified communications system before deploying the system in a production environment. The interoperability testing should verify call flows and features from Cisco and third-party leaf systems through the Unified CM Session Management cluster. To learn which third-party unified communications systems have been tested by the Cisco Interoperability team, refer to the information available on the Cisco Interoperability Portal at

  http://www.cisco.com/en/US/solutions/ns340/ns414/ns728/interOp_ucSessionMgr.html

- Load balancing for inbound and outbound calls

  Configure trunks on the Unified CM Session Management Edition and leaf unified communications systems so that inbound and outbound calls are evenly distributed across the Unified CM servers within the Session Management cluster. For more information on load balancing for trunk calls, refer to the chapter on Cisco Unified CM Trunks, page 14-1.

- Design guidance and assistance

  For detailed information on Unified CM Session Management Edition designs and deployments, refer to the *Cisco Unified Communications Manager Session Management Edition Deployment Guide*, available at

  http://www.cisco.com/en/US/products/ps10661/products_implementation_design_guides_list.html

  Unified CM Session Management Edition designs should be reviewed by your Cisco SE in conjunction with the Cisco Unified CM Session Management Team.

Hybrid Session Management Edition and SAF CCD Deployments

Session Management Edition deployments provide internal dial plan aggregation. Cisco Service Advertisement Framework (SAF) Call Control Discovery (CCD) deployments distribute both the internal dial plan and the corresponding external "To PSTN" dial plan to participating SAF CCD Unified Communications systems. Combining Session Management Edition and SAF CCD enables Session Management Edition to act as the central Session Manager for all leaf Unified Communications systems, while also using SAF CCD to distribute both the internal and external "To PSTN" dial plans to all SAF CCD participating Unified CM leaf clusters.

A Session Management Edition and SAF hybrid deployment uses a specific configuration of SAF CCD to allow all calls between leaf clusters to be routed only through the Session Management Edition cluster. The SAF configuration consists of two parts:

- Advertising SAF CCD routes to leaf clusters from/through Session Management Edition
- Advertising SAF CCD routes from leaf clusters to Session Management Edition

Note This discussion assumes that you have already configured your Cisco IOS SAF Forwarders and basic SAF CCD configuration on Unified CM (that is, Advertising Service, Requesting Service, SAF enabled Trunks, and so forth). This design uses a single SAF Autonomous System (AS).

Advertising SAF CCD Routes to Leaf Clusters from/through Session Management Edition

On the Session Management Edition cluster, create the DN patterns, DN Groups, and corresponding "to DID" rules for the internal number ranges and external "To PSTN" numbers hosted by each SAF-enabled leaf cluster. Publish these DN patterns to the SAF AS by associating them with one or more SAF-enabled trunks and advertising services. These DN patterns and corresponding routes to Session Management Edition are learned by all SAF-enabled leaf clusters. While Session Management Edition is reachable through the IP WAN, all intercluster calls are routed through Session Management Edition. When Session Management Edition is unreachable, intercluster calls are routed through the leaf cluster's local PSTN gateway after the called number has been modified using the learned DN pattern's "to DID" rule.

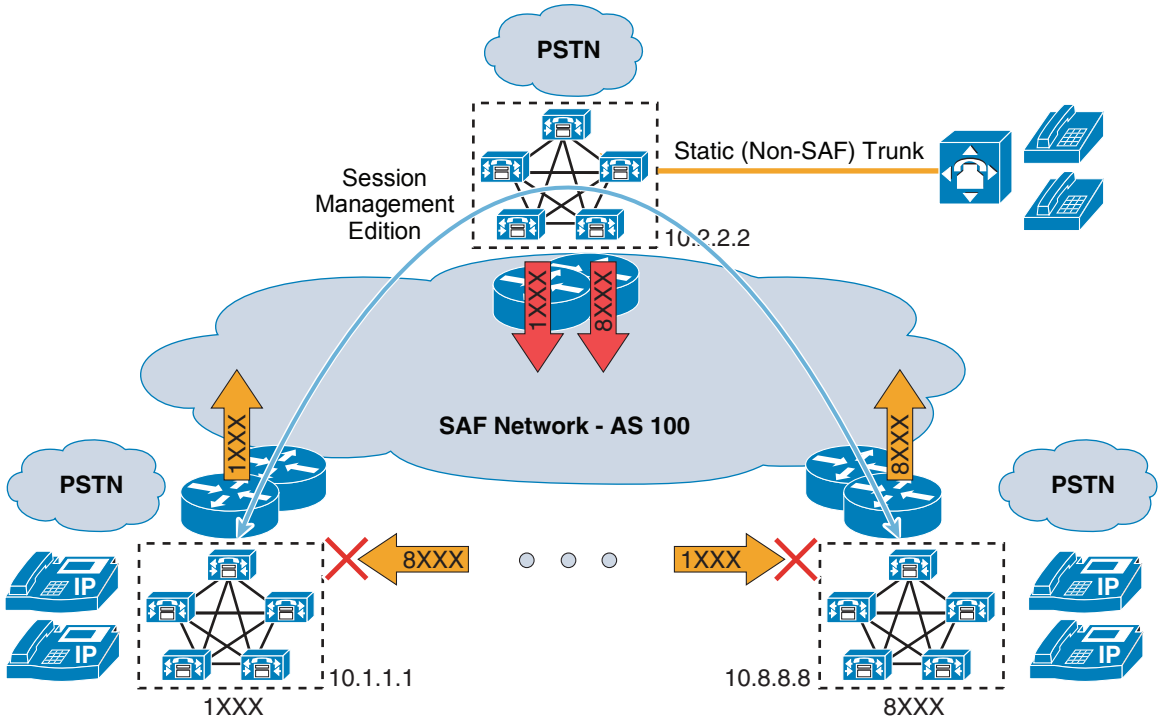
Advertising SAF CCD Routes from Leaf Clusters to Session Management Edition

The purpose of advertising each leaf cluster's hosted DN ranges to the SAF AS is to allow the Session Management Edition cluster to learn about these DN ranges and leaf cluster reachability. These number ranges are also learned by all other leaf clusters. (See Figure 5-7.) To prevent direct leaf-to-leaf routes from being used, in each leaf cluster, block learned routes from all other leaf clusters. Routes can be blocked based on whether they match either the IP address the SAF nodes in each of the leaf clusters or (preferably) the Remote Call Control Entity Name for each leaf cluster. (This is the Unified CM Cluster ID in the Unified CM Enterprise Parameters menu.)

Figure 5-7 Advertising SAF CCD Routes in a Session Management Edition Deployment

Session Management Edition SAF CCD Routing Table

DN Pattern	"to DID"rule	IP address	Protocol
1XXX	0:+1212444	10.1.1.1	SIP
8XXX	0:+1408902	10.8.8.8	SIP



Leaf 1 SAF CCD Routing Table

DN Pattern	"to DID"rule	IP address	Protocol
1XXX	0:+1212444	10.2.2.2	SIP
8XXX	0:+1408902	10.2.2.2	SIP
~~8XXX~~	~~0:+1408902~~	~~10.8.8.8~~	~~SIP~~

Leaf 8 SAF CCD Routing Table

DN Pattern	"to DID"rule	IP address	Protocol
1XXX	0:+1212444	10.2.2.2	SIP
~~1XXX~~	~~0:+1212444~~	~~10.1.1.1~~	~~SIP~~
8XXX	0:+1408902	10.2.2.2	SIP

254275

Operational Considerations for Session Management Edition and SAF CCD Deployments

The following operational considerations apply to deployments of Cisco Unified CM Session Management Edition with Service Advertisement Framework (SAF) Call Control Discovery (CCD).

Leaf Clusters Learning Their Own DN Ranges from Session Management Edition

As can be seen in the SAF CCD routing tables in Figure 5-7, leaf clusters learn about the reachability of their own DN ranges from Session Management Edition. These DN ranges can be blocked in the same way that intercluster DN ranges and routes are blocked. If these Session Management Edition SAF CCD routes are not blocked, they are selected only for intra-cluster calls if the calling search space of the calling device has the SAF CCD learned routes partition ordered above the internal DN's partition. In most cases, the internal DN partition will be ordered above the SAF CCD partition, so that intra-cluster calls are not routed through Session Management Edition.

Routing Calls to the PSTN When IP Routes from Session Management Edition to Leaf Clusters Are Not Available

Two configuration options are available when re-routing calls to the PSTN:

- Re-route calls to the PSTN through a PSTN gateway associated with Session Management Edition

  If the Session Management Edition cluster has PSTN access and you wish to re-route calls that are unreachable through an IP path from Session Management Edition to the destination leaf cluster, make sure each leaf cluster advertises a "to DID" rule for each advertised DN range or group to Session Management Edition. This "to DID" rule is used by Session Management Edition to modify the called number and to route the call through the inbound trunk's Automated Alternate Routing (AAR) calling search space (CSS).

- Re-route calls to the PSTN from the originating leaf cluster

  If the Session Management Edition cluster does not have PSTN access and you wish to re-route calls that are unreachable from Session Management Edition to the destination leaf cluster through the PSTN at the originating leaf cluster, make sure each leaf cluster does not advertises a "to DID" rule for each advertised DN range or group to Session Management Edition. In this case, if a signaling path cannot be established from Session Management Edition to the destination leaf cluster, Session Management Edition signals the call failure to the originating leaf cluster, which in turn uses its "to DID" rule (learned from Session Management Edition) to modify the called number and route the call through the calling device's Automated Alternate Routing (AAR) calling search space (CSS).

Calls to Non-SAF Unified Communications Systems over Static Session Management Edition Trunks

Session Management Edition can use SAF CCD to advertise the DN ranges of non-SAF Unified Communications systems to all SAF-enabled leaf clusters. Calls from leaf clusters to non-SAF Unified Communications systems through the Session Management Edition cluster use SAF trunks to reach Session Management Edition. Session Management Edition then uses a configured route pattern and corresponding static (standard) trunk to reach the non-SAF Unified Communications system.

PSTN Fallback for Calls to Non-SAF Unified Communications Systems

There are two options for PSTN fallback if the non-SAF Unified Communications system is not reachable through a static trunk from Session Management Edition:

- Re-route calls to the PSTN from the originating leaf cluster.

  With this option, a single trunk is configured from Session Management Edition to the destination Unified Communications system. If a signaling path cannot be established from Session Management Edition to the destination Unified Communications system, Session Management

Edition signals the call failure to the originating leaf cluster, which in turn uses its "to DID" rule (learned from Session Management Edition) to modify the called number and route the call through the calling device's Automated Alternate Routing (AAR) calling search space (CSS).

- Re-route calls to the PSTN from Session Management Edition.

  With this option, create two trunks as part of a route list and route group. The first-choice trunk is configured from Session Management Edition to the destination Unified Communications system, while the second-choice trunk is configured from Session Management Edition to its local PSTN gateway. If a signaling path cannot be established from Session Management Edition to the destination Unified Communications system, Session Management Edition chooses the second trunk to the PSTN. The route group that contains the PSTN trunk can be used to modify the internal called number to its PSTN equivalent.

Clustering Over the IP WAN

You may deploy a single Unified CM cluster across multiple sites that are connected by an IP WAN with QoS features enabled. This section provides a brief overview of clustering over the WAN. For further information, refer to the chapter on Call Processing, page 8-1.

Clustering over the WAN can support two types of deployments:

- Local Failover Deployment Model, page 5-37

  Local failover requires that you place the Unified CM subscriber and backup servers at the same site, with no WAN between them. This type of deployment is ideal for two to four sites with Unified CM.

- Remote Failover Deployment Model, page 5-43

  Remote failover allows you to deploy primary and backup call processing servers split across the WAN. Using this type of deployment, you may have multiple sites with Unified CM subscribers being backed up by Unified CM subscribers at another site.

Note Remote failover deployments might require higher bandwidth because a large amount of intra-cluster traffic flows between the subscriber servers.

You can also use a combination of the two deployment models to satisfy specific site requirements. For example, two main sites may each have primary and backup subscribers, with another two sites containing only a primary server each and utilizing either shared backups or dedicated backups at the two main sites.

Some of the key advantages of clustering over the WAN are:

- Single point of administration for users for all sites within the cluster
- Feature transparency
- Shared line appearances
- Extension mobility within the cluster
- Unified dial plan

These features make this solution ideal as a disaster recovery plan for business continuance sites or as a single solution for multiple small or medium sites.

WAN Considerations

For clustering over the WAN to be successful, you must carefully plan, design, and implement various characteristics of the WAN itself. The Intra-Cluster Communication Signaling (ICCS) between Unified CM servers consists of many traffic types. The ICCS traffic types are classified as either priority or best-effort. Priority ICCS traffic is marked with IP Precedence 3 (DSCP 24 or PHB CS3). Best-effort ICCS traffic is marked with IP Precedence 0 (DSCP 0 or PHB BE). The various types of ICCS traffic are described in Intra-Cluster Communications, page 5-34, which also provides further guidelines for provisioning. The following design guidelines apply to the indicated WAN characteristics:

- Delay

  The maximum one-way delay between any two Unified CM servers should not exceed 40 ms, or 80 ms round-trip time. Measuring the delay is covered in Delay Testing, page 5-36. Propagation delay between two sites introduces 6 microseconds per kilometer without any other network delays being considered. This equates to a theoretical maximum distance of approximately 6,000 km for 40 ms delay or approximately 3,720 miles. These distances are provided only as relative guidelines and in reality will be shorter due to other delay incurred within the network.

- Jitter

  Jitter is the varying delay that packets incur through the network due to processing, queue, buffer, congestion, or path variation delay. Jitter for the IP Precedence 3 ICCS traffic must be minimized using Quality of Service (QoS) features.

- Packet loss and errors

  The network should be engineered to provide sufficient prioritized bandwidth for all ICCS traffic, especially the priority ICCS traffic. Standard QoS mechanisms must be implemented to avoid congestion and packet loss. If packets are lost due to line errors or other “real world” conditions, the ICCS packet will be retransmitted because it uses the TCP protocol for reliable transmission. The retransmission might result in a call being delayed during setup, disconnect (teardown), or other supplementary services during the call. Some packet loss conditions could result in a lost call, but this scenario should be no more likely than errors occurring on a T1 or E1, which affect calls via a trunk to the PSTN/ISDN.

- Bandwidth

  Provision the correct amount of bandwidth between each server for the expected call volume, type of devices, and number of devices. This bandwidth is in addition to any other bandwidth for other applications sharing the network, including voice and video traffic between the sites. The bandwidth provisioned must have QoS enabled to provide the prioritization and scheduling for the different classes of traffic. The general rule of thumb for bandwidth is to over-provision and under-subscribe.

- Quality of Service

  The network infrastructure relies on QoS engineering to provide consistent and predictable end-to-end levels of service for traffic. Neither QoS nor bandwidth alone is the solution; rather, QoS-enabled bandwidth must be engineered into the network infrastructure.

Intra-Cluster Communications

In general, intra-cluster communications means all traffic between servers. There is also a real-time protocol called Intra-Cluster Communication Signaling (ICCS), which provides the communications with the Cisco CallManager Service process that is at the heart of the call processing in each server or node within the cluster.

The intra-cluster traffic between the servers consists of the following:

- Database traffic from the IBM Informix Dynamic Server (IDS) database that provides the main configuration information. The IDS traffic may be re-prioritized in line with Cisco QoS recommendations to a higher priority data service (for example, IP Precedence 1 if required by the particular business needs). An example of this is extensive use of Extension Mobility, which relies on IDS database configuration.
- Firewall management traffic, which is used to authenticate the subscribers to the publisher to access the publisher's database. The management traffic flows between all servers in a cluster. The management traffic may be prioritized in line with Cisco QoS recommendations to a higher priority data service (for example, IP Precedence 1 if required by the particular business needs).
- ICCS real-time traffic, which consists of signaling, call admission control, and other information regarding calls as they are initiated and completed. ICCS uses a Transmission Control Protocol (TCP) connection between all servers that have the Cisco CallManager Service enabled. The connections are a full mesh between these servers. This traffic is priority ICCS traffic and is marked dependant on release and service parameter configuration.
- CTI Manager real-time traffic is used for CTI devices involved in calls or for controlling or monitoring other third-party devices on the Unified CM servers. This traffic is marked as priority ICCS traffic and exists between the Unified CM server with the CTI Manager and the Unified CM server with the CTI device.

Note For detailed information on various types of traffic between Unified CM servers, refer to the TCP and UDP port usage documents at http://www.cisco.com/en/US/products/sw/voicesw/ps556/prod_maintenance_guides_list.html.

Unified CM Publisher

The publisher server replicates a partial read-only copy of the master database to all other servers in the cluster. Most of the database modifications are done on the publisher. If changes such as administration updates are made in the publisher's master database during a period when another server in the cluster is unreachable, the publisher will replicate the updated database when communications are re-established. Database modifications for user-facing call processing features are made on the subscriber servers to which the IP phones are registered. These features include:

- Call Forward All (CFA)
- Message Waiting Indication (MWI)
- Privacy Enable/Disable
- Do Not Disturb (DND) Enable/Disable
- Extension Mobility Login (EM)
- Monitor (for future use; currently no updates at the user level)
- Hunt Group Logout
- Device Mobility
- CTI Certificate Authority Proxy Function (CAPF) status for end users and application users
- Credential hacking and authentication

Each subscriber replicates these changes to every other server in the cluster. Any other configuration changes cannot be made on the database during the period when the publisher is unreachable or offline. Most normal operations of the cluster, including the following, will *not* be affected during the period of publisher failure:

- Call processing
- Failover
- Registration of previously configured devices

Other services or applications might also be affected, and their ability to function without the publisher should be verified when deployed.

Call Detail Records (CDR) and Call Management Records (CMR)

Call detail records and call management records, when enabled, are collected by each subscriber and uploaded to the publisher periodically. During a period that the publisher is unreachable, the CDRs and CMRs are stored on the subscriber's local hard disk. When connectivity is re-established to the publisher, all outstanding CDRs are uploaded to the publisher, which stores the records in the CDR Analysis and Reporting (CAR) database.

Delay Testing

The maximum round-trip time (RTT) between any two servers must not exceed 80 ms. This time limit must include all delays in the transmission path between the two servers. Verifying the round trip delay using the **ping** utility on the Unified CM server will not provide an accurate result. The ping is sent as a best-effort tagged packet and is not transported using the same QoS-enabled path as the ICCS traffic. Therefore, Cisco recommends that you verify the delay by using the closest network device to the Unified CM servers, ideally the access switch to which the server is attached. Cisco IOS provides a extended ping capable to set the Layer 3 type of service (ToS) bits to make sure the ping packet is sent on the same QoS-enabled path that the ICCS traffic will traverse. The time recorded by the extended ping is the round-trip time (RTT), or the time it takes to traverse the communications path and return.

The following example shows a Cisco IOS extended ping with the IP Precedence bits set to 3 (ToS byte value set to 96):

```
Access_SW#ping
Protocol [ip]:
Target IP address: 10.10.10.10
Repeat count [5]:
Datagram size [100]:
Timeout in seconds [2]:
Extended commands [n]: y
Source address or interface:
Type of service [0]: 96
Set DF bit in IP header? [no]:
Validate reply data? [no]:
Data pattern [0xABCD]:
Loose, Strict, Record, Timestamp, Verbose[none]:
Sweep range of sizes [n]:
Type escape sequence to abort.
Sending 5, 100-byte ICMP Echos to 10.10.10.10, timeout is 2 seconds:
!!!!!
Success rate is 100 percent (5/5), round-trip min/avg/max = 1/2/4 ms
```

Error Rate

The expected error rate should be zero. Any errors, dropped packets, or other impairments to the IP network can have an impact to the call processing performance of the cluster. This may be noticeable by delay in dial tone, slow key or display response on the IP phone, or delay from off-hook to connection of the voice path. Although Unified CM will tolerate random errors, they should be avoided to avoid impairing the performance of the cluster.

Troubleshooting

If the Unified CM subscribers in a cluster are experiencing impairment of the ICCS communication due to higher than expected delay, errors, or dropped packets, some of the following symptoms might occur:

- IP phones, gateways, or other devices on a remote Unified CM server within the cluster might temporarily be unreachable.
- Calls might be disconnected or might fail during call setup.
- Users might experience longer than expected delays before hearing dial tone.
- Busy hour call completions (BHCC) might be low.
- The ICCS (SDL session) might be reset or disconnected.
- The time taken to upgrade a subscriber and synchronize its database with the publisher will increase.

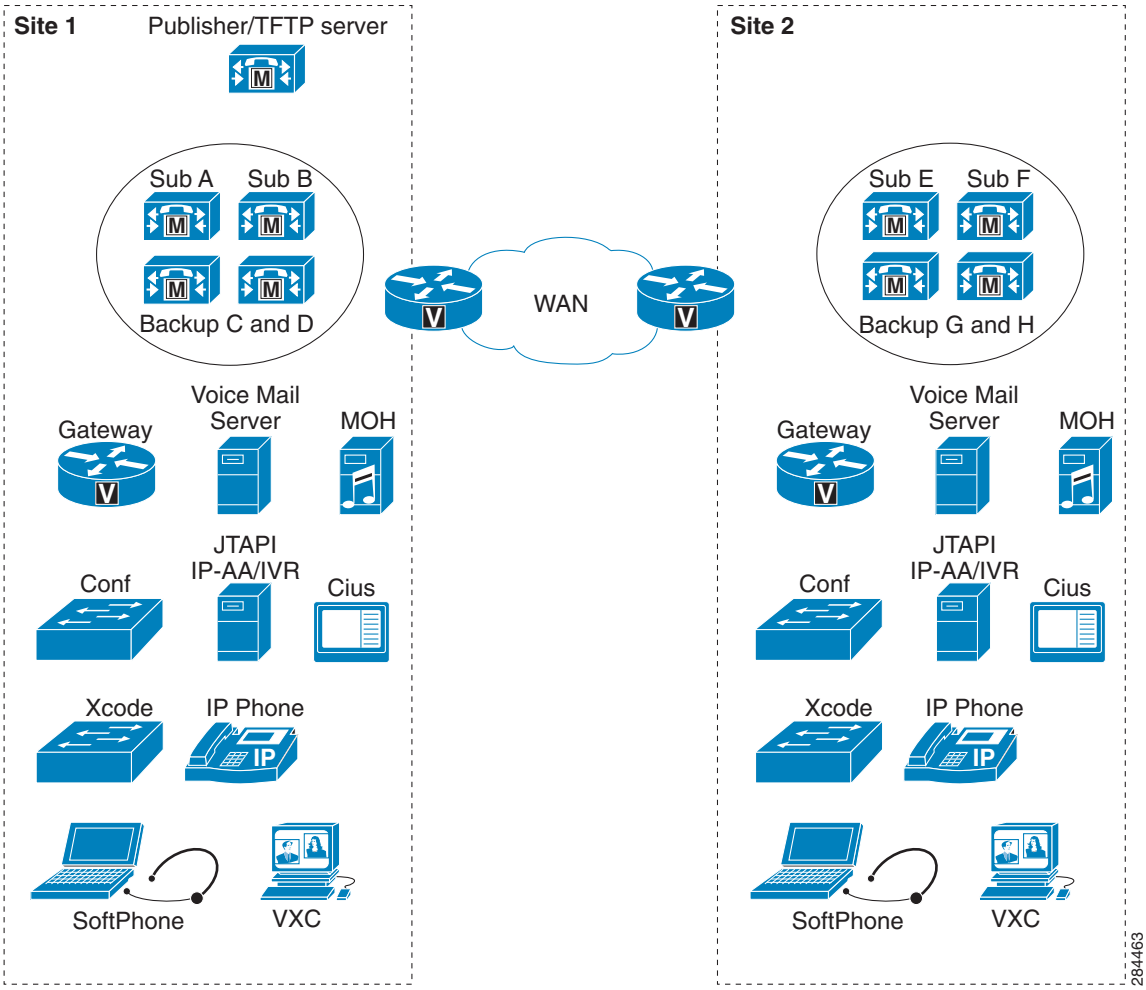
In summary, perform the following tasks to troubleshoot ICCS communication problems:

- Verify the delay between the servers.
- Check all links for errors or dropped packets.
- Verify that QoS is correctly configured.
- Verify that sufficient bandwidth is provisioned for the queues and across the WAN to support all the traffic.

Local Failover Deployment Model

The local failover deployment model provides the most resilience for clustering over the WAN. Each of the sites in this model contains at least one primary Unified CM subscriber and one backup subscriber. This configuration can support up to four sites. The maximum number of phones and other devices will be dependant on the quantity and type of servers deployed. The maximum total number of IP phones for all sites is 40,000. (See Figure 5-8.)

Figure 5-8 Example of Local Failover Model



Observe the following guidelines when implementing the local failover model:

- Configure each site to contain at least one primary Unified CM subscriber and one backup subscriber.
- Configure Unified CM *groups* and *device pools* to allow devices within the site to register with only the servers at that site under all conditions.
- Cisco highly recommends that you replicate key services (TFTP, DNS, DHCP, LDAP, and IP Phone Services), all media resources (conference bridges and music on hold), and gateways at each site to provide the highest level of resiliency. You could also extend this practice to include a voicemail system at each site.
- Under a WAN failure condition, sites without access to the publisher database will lose some functionality. For example, system administration at the remote site will not be able to add, modify, or delete any part of the configuration. However, users can continue to access the user-facing features listed in the section on Unified CM Publisher, page 5-35.
- Under WAN failure conditions, calls made to phone numbers that are not currently communicating with the subscriber placing the call, will result in either a fast-busy tone or a call forward (possibly to voicemail or to a destination configured under Call Forward Unregistered).

- The maximum allowed round-trip time (RTT) between any two servers in the Unified CM cluster is 80 ms.

Note At a higher round-trip delay time and higher busy hour call attempts (BHCA), voice cut-through delay might be higher, causing initial voice clipping when a voice call is established.

- A minimum of 1.544 Mbps (T1) bandwidth is required for Intra-Cluster Communication Signaling (ICCS) for 10,000 busy hour call attempts (BHCA) between sites that are clustered over the WAN. This is a minimum bandwidth requirement for call control traffic, and it applies to deployments where directory numbers are not shared between sites that are clustered over the WAN. The following equation may be used as a guideline to calculate the bandwidth for more than 10,000 BHCA between non-shared directory numbers at a specific delay:

  Total Bandwidth (Mbps) = (Total BHCA/10,000) * (1 + 0.006 * Delay), where
  Delay = RTT delay in ms

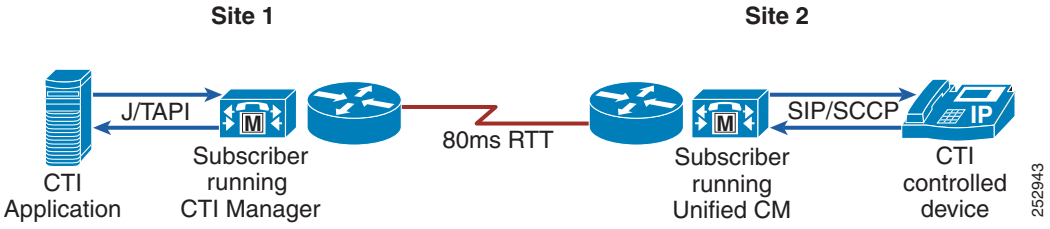
  This call control traffic is classified as priority traffic. Priority ICCS traffic is marked with IP Precedence 3 (DSCP 24 or PHB CS3).

- In addition to the bandwidth required for Intra-Cluster Communication Signaling (ICCS) traffic, a minimum of 1.544 Mbps (T1) bandwidth is required for database and other inter-server traffic for every subscriber server remote to the publisher.
- For customers who also want to deploy CTI Manager over the WAN (see Figure 5-9), the following formula can be used to calculate the bandwidth (Mbps) for the CTI Intra-Cluster Communication Signaling (ICCS) traffic between the Unified CM subscriber running the CTI Manager service and the Unified CM subscriber to which the CTI controlled endpoint is registered:

  With Unified CM 8.6(1) and earlier releases, CTI ICCS bandwidth (Mbps)
  = (Total BHCA/10,000) * 1.25

  With Unified CM 8.6(2) and later releases, CTI ICCS bandwidth (Mbps)
  = (Total BHCA/10,000) * 0.53

Figure 5-9 CTI Over the WAN

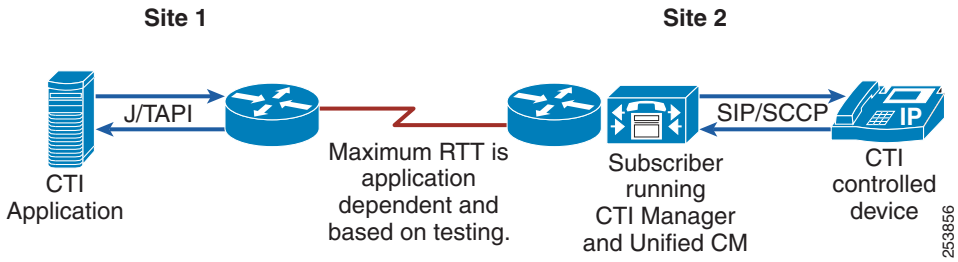


- For deployments where the J/TAPI application is remote from the Unified CM subscriber (see Figure 5-10), the following formula can be used to calculate the Quick Buffer Encoding (QBE) J/TAPI bandwidth for a typical J/TAPI application with Unified CM 8.6(2) and later releases:

  J/TAPI bandwidth (Mbps) = (Total BHCA/10,000) * 0.28

  The bandwidth may vary depending on the J/TAPI application. Check with the application developer or provider to validate the bandwidth requirement.

Figure 5-10 J/TAPI Over the WAN



Example 5-1 Bandwidth Calculation for Two Sites

Consider two sites, Site 1 and Site 2, with Unified CM clustered over the WAN across these two sites that are 80 ms round-trip time apart. Site 1 has one publisher, one combined TFTP and music on hold (MoH) server, and two Unified CM subscriber servers. Site 2 has one TFTP/MoH server and two Unified CM subscriber servers. Site 1 has 5000 phones, each having one DN; and Site 2 has 5000 phones, each having one DN. During the busy hour, 2500 phones in Site 1 call 2500 phones in Site 2, each at 3 BHCA. During that same busy hour, 2500 phones in Site 2 also call 2500 phones in Site 1, each at 3 BHCA. In this case:

Total BHCA during the busy hour = 2500∗3 + 2500∗3 = 15,000

Total bandwidth required between the sites = Total ICCS bandwidth + Total database bandwidth

Because total BHCA is 15,000 (greater than 10,000), we can use the formula to calculate:
Total ICCS bandwidth = (15,000/10,000) ∗ (1 + 0.006∗80) = 2.22 Mbps

Total database bandwidth = (Number of servers remote to the publisher) ∗ 1.544 = 3 ∗ 1.544 = 4.632 Mbps

Total bandwidth required between the sites = 2.22 Mbps + 4.632 Mbps = 6.852 Mbps (Approximately 7 Mbps)

- When directory numbers are shared between sites that are clustered over the WAN, additional bandwidth must be reserved. This overhead or additional bandwidth (in addition to the minimum 1.544 Mbps bandwidth) for 10,000 BHCA between shared DNs can be calculated using the following equation:

  Overhead = (0.012 ∗ Delay ∗ Shared-line) + (0.65 ∗ Shared-line), where:

  Delay = RTT delay over the IP WAN, in ms

  Shared-line = Average number of additional phones on which a directory number is shared across the WAN.

  The following equation may be used as a guideline to calculate the bandwidth for more than 10,000 BHCA between shared directory numbers at a specific delay:

  Total bandwidth (Mbps) = (Total BHCA/10,000) ∗ (1 + 0.006 ∗ Delay + 0.012 ∗ Delay ∗ Shared-line + 0.65 ∗ Shared-line), where:

  Delay = RTT delay in ms

  Shared-line = Average number of additional phones on which a directory number is shared across the WAN.

Example 5-2 Bandwidth Calculation for Two Sites with Shared Directory Numbers

Consider two sites, Site 1 and Site 2, with Unified CM clustered over the WAN across these two sites that are 80 ms round-trip time apart. Site 1 has one publisher, one combined TFTP and music on hold (MoH) server, and two Unified CM subscriber servers. Site 2 has one TFTP/MoH server and two Unified CM subscriber servers. Site 1 has 5000 phones, each having one DN; and Site 2 has 5000 phones, each sharing a DN with the 5000 phones in Site 1. Thus, each DN is shared across the WAN with an average of one additional phone. During the busy hour, 2500 phones in Site 1 call 2500 phones in Site 2, each at 3 BHCA. This also causes the phones in Site 1 to ring. During that same busy hour, 2500 phones in Site 2 call 2500 phones in Site 1, each at 3 BHCA. This also causes the phones in Site 2 to ring. In this case:

Total BHCA during the busy hour = 2500 * 3 + 2500 * 3 = 15,000

Total bandwidth required between the sites = Total ICCS bandwidth + Total database bandwidth

Because total BHCA is 15,000 (greater than 10,000), we can use the formula to calculate:
Total ICCS bandwidth = (15,000/10,000) * (1 + 0.006*80 + 0.012*80*1 + 0.65*1) = 4.635 Mbps

Total database bandwidth = (Number of servers remote to the publisher) * 1.544 = 3 * 1.544 = 4.632 Mbps

Total bandwidth required between the sites = 4.635 Mbps + 4.632 Mbps = 9.267 Mbps (Approximately 10 Mbps)

Note The bandwidth requirements stated above are strictly for ICCS, database, and other inter-server traffic. If calls are going over the IP WAN, additional bandwidth must be provisioned for voice or media traffic, depending on the voice codec used for the calls.

- Subscriber servers in the cluster read their local database. Database modifications can occur in both the local database as well as the publisher database, depending on the type of changes. Informix Dynamic Server (IDS) database replication is used to synchronize the databases on the various servers in the cluster. Therefore, when recovering from failure conditions such as the loss of WAN connectivity for an extended period of time, the Unified CM databases must be synchronized with any changes that might have been made during the outage. This process happens automatically when database connectivity is restored to the publisher and other servers in the cluster. This process can take longer over low bandwidth and/or higher delay links. In rare scenarios, manual reset or repair of the database replication between servers in the cluster might be required. This is performed by using the commands such as **utils dbreplication repair all** and/or **utils dbreplication reset all** at the command line interface (CLI). Repair or reset of database replication using the CLI on remote subscribers over the WAN causes all Unified CM databases in the cluster to be re-synchronized, in which case additional bandwidth above 1.544 Mbps might be required. With longer delays and lower bandwidth between the publisher and subscriber nodes, it can take longer for database replication repair or reset to complete.

  Note Repairing or resetting of database replication on multiple subscribers at the same remote location can result in increased time for database replication to complete. Cisco recommends repairing or resetting of database replication on these remote subscribers one at a time. Repairing or resetting of database replication on subscribers at different remote locations may be performed simultaneously.

- If remote branches using centralized call processing are connected to the main sites via clustering over the WAN, pay careful attention to the configuration of call admission control to avoid oversubscribing the links used for clustering over the WAN.
  - If the bandwidth is not limited on the links used for clustering over the WAN (that is, if the interfaces to the links are OC-3s or STM-1s and there is no requirement for call admission control), then the remote sites may be connected to any of the main sites because all the main sites should be configured as location Hub_None. This configuration still maintains hub-and-spoke topology for purposes of call admission control.
  - If you are using the Multiprotocol Label Switching (MPLS) Virtual Private Network (VPN) feature, all sites in Unified CM locations and the remote sites may register with any of the main sites.
  - If bandwidth is limited between the main sites, call admission control must be used between sites, and all remote sites must register with the main site that is configured as location Hub_None. This main site is considered the hub site, and all other remote sites and clustering-over-the-WAN sites are spokes sites.
- During a software upgrade, all servers in the cluster should be upgraded during the same maintenance period, using the standard upgrade procedures outlined in the software release notes. The software upgrade time will increase for higher round-trip delay time over the IP WAN. Lower bandwidths such as 1.544 Mbps (T1 link) can also cause the software upgrade process to take longer to complete, in which case additional bandwidth above 1.544 Mbps might be required if a faster upgrade process is desired.

Unified CM Provisioning for Local Failover

Provisioning of the Unified CM cluster for the local failover model should follow the design guidelines for capacities outlined in the chapter on Call Processing, page 8-1. If voice or video calls are allowed across the WAN between the sites, then you must configure Unified CM *locations* in addition to the default location for the other sites, to provide call admission control between the sites. If the bandwidth is over-provisioned for the number of devices, it is still best practice to configure call admission control based on locations. If the locations-based call admission control rejects a call, automatic failover to the PSTN can be provided by the automated alternate routing (AAR) feature.

To improve redundancy and upgrade times, Cisco recommends that you enable the Cisco Trivial File Transfer Protocol (TFTP) service on two Unified CM servers. More than two TFTP servers can be deployed in a cluster, however this configuration can result in an extended period for rebuilding all the TFTP files on all TFTP servers.

You can run the TFTP service on either a publisher or a subscriber server, depending on the site and the available capacity of the server. The TFTP server option must be correctly set in the DHCP servers at each site. If DHCP is not in use or if the TFTP server is manually configured, you should configure the correct address for the site.

Other services, which may affect normal operation of Unified CM during WAN outages, should also be replicated at all sites to ensure uninterrupted service. These services include DHCP servers, DNS servers, corporate directories, and IP phone services. On each DHCP server, set the DNS server address correctly for each location.

IP phones may have shared line appearances between the sites. During a WAN outage, call control for each line appearance is segmented, but call control returns to a single Unified CM server once the WAN is restored. During the WAN restoration period, there is additional traffic between the two sites. If this situation occurs during a period of high call volume, the shared lines might not operate as expected during that period. This situation should not last more than a few minutes, but if it is a concern, you can provision additional prioritized bandwidth to minimize the effects.

Gateways for Local Failover

Normally, gateways should be provided at all sites for access to the PSTN. The device pools should be configured to register the gateways with the Unified CM servers at the same site. Call routing (route patterns, route lists, and route groups) should also be configured to select the local gateways at the site as the first choice for PSTN access and the other site gateways as a second choice for overflow. Take special care to ensure emergency service access at each site.

You can centralize access to the PSTN gateways if access is not required during a WAN failure and if sufficient additional bandwidth is configured for the number of calls across the WAN. For E911 requirements, additional gateways might be needed at each site.

Voicemail for Local Failover

Cisco Unity Connection or other voicemail systems can be deployed at all sites and integrated into the Unified CM cluster. This configuration provides voicemail access even during a WAN failure and without using the PSTN. Using Voice Mail Profiles, you can allocate the correct voicemail system for the site to the IP phones in the same location. You can configure a maximum of four voicemail systems per cluster that use the SMDI protocol, that are attached directly to the COM port on a subscriber, and that use the Cisco Messaging Interface (CMI).

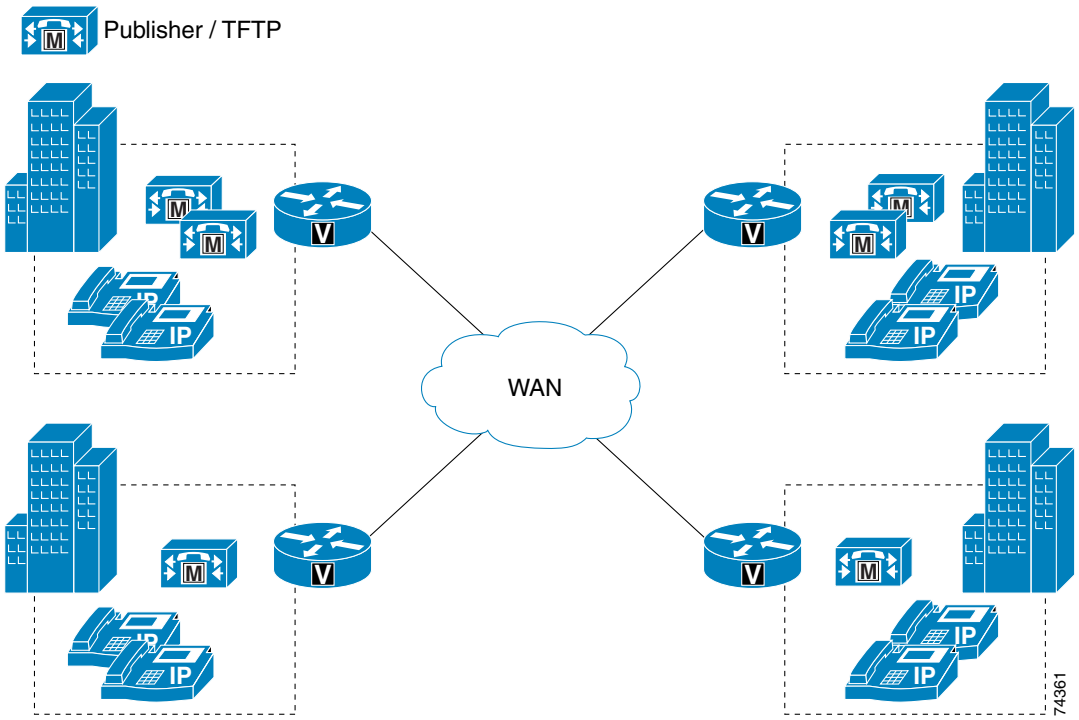
Music on Hold and Media Resources for Local Failover

Music on hold (MoH) servers and other media resources such as conference bridges should be provisioned at each site, with sufficient capacity for the type and number of users. Through the use of media resource groups (MRGs) and media resource group lists (MRGLs), media resources are provided by the on-site resource and are available during a WAN failure.

Remote Failover Deployment Model

The remote failover deployment model provides flexibility for the placement of backup servers. Each of the sites contains at least one primary Unified CM subscriber and may or may not have a backup subscriber. This model allows for multiple sites, with IP phones and other devices normally registered to a local subscriber when using 1:1 redundancy and the 50/50 load balancing option described in the chapter on Call Processing, page 8-1. Backup subscribers are located across the WAN at one or more of the other sites. (See Figure 5-11.)

Figure 5-11 Remote Failover Model with Four Sites



When implementing the remote failover model, observe all guidelines for the local failover model (see Local Failover Deployment Model, page 5-37), with the following modifications:

- Configure each site to contain at least one primary Unified CM subscriber and an optional backup subscriber as desired. If a backup subscriber over the IP WAN is not desired, a Survivable Remote Site Telephony (SRST) router may be used as a backup call processing agent.
- You may configure Unified CM *groups* and *device pools* to allow devices to register with servers over the WAN as a second or third choice.
- Signaling or call control traffic requires bandwidth when devices are registered across the WAN with a remote Unified CM server in the same cluster. This bandwidth might be more than the ICCS traffic and should be calculated using the bandwidth provisioning calculations for signaling, as described in Bandwidth Provisioning, page 3-45.

Note You can also combine the features of these two types of deployments for disaster recovery purposes. For example, Unified CM groups permit configuring up to three servers (primary, secondary and tertiary). Therefore, you can configure the Unified CM groups to have primary and secondary servers that are located at the same site and the tertiary server at a remote site over the WAN.

Cisco Business Edition 6000 Clustering over the WAN

Cisco Business Edition 6000 may be deployed using the clustering-over-the-WAN call processing local failover model. In this type of deployment, two Business Edition 6000 server nodes are deployed at each of two sites to provide geographic redundancy for the Unified CM call processing application. The two Business Edition 6000 server nodes may both be UCS C200 Rack-Mount Servers, or alternatively one of the servers may be a regular Cisco Media Convergence Server (MCS).

Business Edition 6000 call processing clustering over the WAN deployments must observe the same guidelines and requirements as with regular Unified CM clustering over the WAN and as described earlier. Observe the following guidelines when clustering Business Edition 6000 over the WAN with the local failover model:

- Configure Unified CM groups and device pools to allow devices within each site to register with only the servers at that site under all conditions.
- Cisco highly recommends that you replicate key services (TFTP, DNS, DHCP, LDAP, and IP Phone Services), all media resources (conference bridges and music on hold), and gateways at each site to provide the highest level of resiliency.
- Under a WAN failure condition, the site without access to the publisher database will lose some functionality. For example, system administration at the secondary site will not be able to add, modify, or delete any part of the configuration. However, users can continue to access the user-facing features listed in the section on Unified CM Publisher, page 5-35.
- Under WAN failure conditions, calls made to phone numbers that are not currently communicating with the subscriber placing the call, will result in either a fast-busy tone or a call forward (possibly to voicemail or to a destination configured under Call Forward Unregistered).
- The maximum allowed round-trip time (RTT) between the two Business Edition 6000 server nodes at the two sites is 80 ms.
- 1.544 Mbps (T1) bandwidth is required for Intra-Cluster Communication Signaling (ICCS) busy hour call attempts (BHCA) between the two sites that are clustered over the WAN. This is a bandwidth requirement for call control traffic, and it applies to deployments where directory numbers are not shared between sites that are clustered over the WAN. This call control traffic is classified as priority traffic. Priority ICCS traffic is marked with IP Precedence 3 (DSCP 24 or PHB CS3).
- In addition to the bandwidth required for Intra-Cluster Communication Signaling (ICCS) traffic, an additional 1.544 Mbps (T1) of bandwidth is required for database and other traffic between the two Business Edition 6000 server nodes.

More than two UCS C200 Rack-Mount Servers may be clustered for a Business Edition 6000 deployment to provide additional geographic redundancy beyond two sites with the remote failover model for clustering over the WAN (see Remote Failover Deployment Model, page 5-43). However, the total number of users across the Business Edition 6000 cluster may not exceed 1,000 and the total number of configured devices across the cluster may not exceed 1,200. A deployment of UCS C200 Rack-Mount Servers in a cluster exceeding 1,000 users and 1,200 configured devices is considered a regular Unified CM cluster, and as such it is bound by all requirements and design guidance for regular Unified CM clusters.

In deployments of Business Edition 6000 with more than two UCS C200 Rack-Mount Servers in the remote failover model for clustering over the WAN, the following additional guidelines must be observed:

- 1.544 Mbps (T1) bandwidth is required for Intra-Cluster Communication Signaling (ICCS) busy hour call attempts (BHCA) between each site that is clustered over the WAN. This is a bandwidth requirement for call control traffic.
- In addition to the bandwidth required for Intra-Cluster Communication Signaling (ICCS) traffic, an additional 1.544 Mbps (T1) of bandwidth is required for database and other inter-server traffic between any server nodes remote from the Business Edition 6000 publisher node.

Clustering over the WAN for Business Edition 6000 Co-Resident Applications

In addition to clustering call processing services over the WAN, Cisco Business Edition 6000 co-resident applications (Cisco Unity Connection, Cisco IM and Presence, and Cisco Unified Contact Center Express) may also be clustered over the WAN provided that these deployments adhere to the same guidelines and restrictions as apply to these applications running on separate systems.

Each co-resident application must adhere strictly to its maximum delay and bandwidth requirements. Furthermore, it is important to understand that, while maximum delay budget will apply to all applications, the WAN bandwidth required for each clustered application (including call processing) must be added together to derive the appropriate WAN bandwidth requirement.

Observe the following general guidelines when clustering co-resident Cisco Business Edition 6000 applications and services:

- Round-trip delay across the WAN must not exceed 80 milliseconds because this is the maximum round-trip delay supported across all applications, including call processing.
- The bandwidth requirement on the WAN is based on the total of each application's bandwidth requirement for clustering over the WAN. For example, if all applications (Cisco Unified CM, IM and Presence, Unity Connection, and Unified Contact Center Express) are clustered over the WAN, the total bandwidth required on the WAN would be calculated as follows:

  (Total required WAN bandwidth) = (Unified CM required bandwidth) + (IM and Presence required bandwidth) + (Unity Connection required bandwidth) + (Unified Contact Center Express required bandwidth)

For information on clustering delay and bandwidth requirement for each co-resident application, refer to the following information:

- For Cisco IM and Presence, see Clustering Over the WAN, page 23-21.
- For Cisco Unity Connection, see Cisco Unity Connection Redundancy and Clustering Over the WAN, page 21-17.
- For Cisco Unified Contact Center Express, see Clustering Over the IP WAN, page 26-9.

Deploying Unified Communications on Virtualized Servers

Cisco Unified Communications applications can run in a virtualized environment as virtual machines using the VMware ESXi hypervisor. Two hardware options are available:

- Tested Reference Configurations (TRC), which are selected hardware configurations based on Cisco Unified Computing System (UCS) platforms
- Specification-based hardware that provides more hardware flexibility and that, for example, adds support for other Cisco UCS, Hewlett-Packard, and IBM platforms listed in the VMware Hardware Compatibility List (available at (http://www.vmware.com/resources/compatibility/search.php)

This section presents a short introduction of the Cisco Unified Computing System (UCS) architecture, Hypervisor Technology for Application Virtualization, and Storage Area Networking (SAN) concepts, with a simple overview of where each product fits in a Cisco Virtualized Unified Communications solution for enterprises. It also includes design considerations for deploying Unified Communications applications over virtualized servers.

This description is not meant to replace or supersede product-specific detailed design guidelines available at the following locations:

- http://www.cisco.com/en/US/products/ps10265/index.html
- http://www.cisco.com/go/uc-virtualized

For sizing aspects of Unified Communications systems on virtualized servers, use the Cisco Unified Communications Sizing Tool, available to Cisco partners and employees (with valid login authentication) at

http://tools.cisco.com/cucst

Cisco Unified Computing System

Unified Computing is an architecture that integrates computing resources (CPU, memory, and I/O), IP networking, network-based storage, and virtualization, into a single highly available system. This level of integration provides economies of power and cooling, simplified server connectivity into the network, dynamic application instance repositioning between physical hosts, and pooled disk storage capacity.

The Cisco Unified Computing System is built from many components. But from a server standpoint, the UCS architecture is divided into the following two categories:

- Cisco UCS B-Series Blade Servers, page 5-47
- Cisco UCS C-Series Rack-Mount Servers, page 5-50

For more details on the Cisco Unified Computing System architecture, refer to the documentation available at

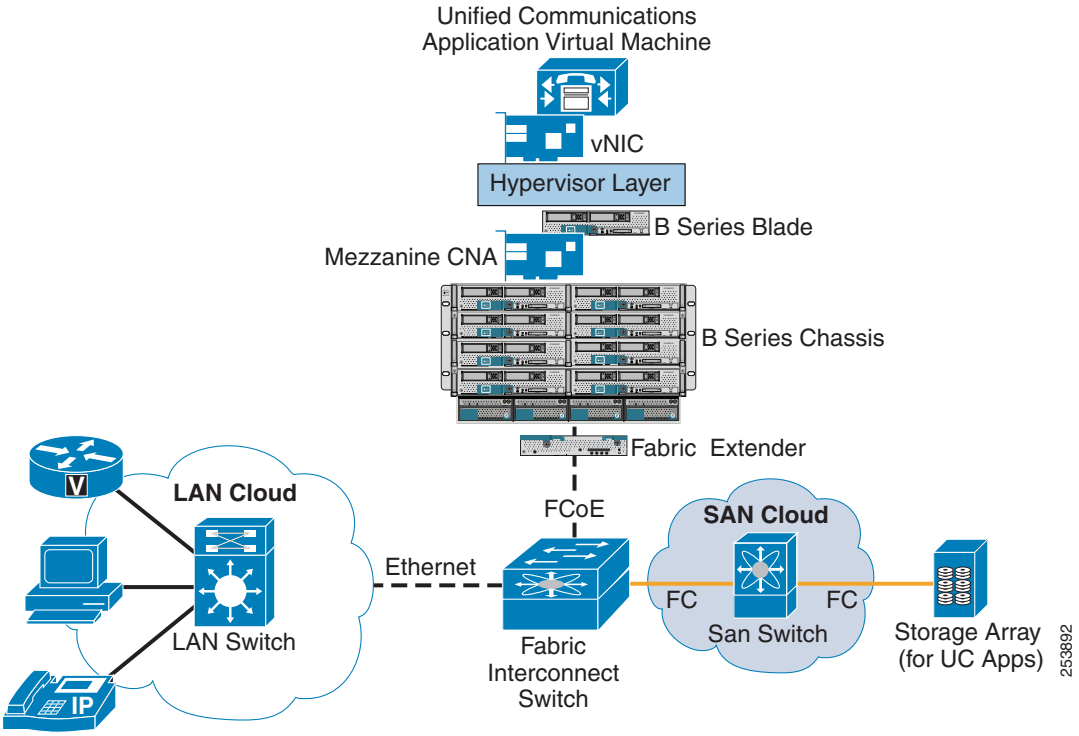
http://www.cisco.com/en/US/netsol/ns944/index.html

Cisco UCS B-Series Blade Servers

The Cisco Unified Computing System (UCS) features blade servers based on x86 architecture. Blade servers provide computing resources (memory, CPU, and I/O) to operating systems and applications. Blade servers have access to the unified fabric through mezzanine form factor Converged Network Adapters (CNA).

The architecture uses a unified fabric that provides transport for LAN, storage, and high-performance computing traffic over a single infrastructure with the help of technologies such as Fibre Channel over Ethernet (FCoE). (See Figure 5-12.) Cisco's unified fabric technology is built on a 10-Gbps Ethernet foundation that eliminates the need for multiple sets of adapters, cables, and switches for LANs, SANs, and high-performance computing networks.

Figure 5-12 Basic Architecture of Unified Communications on Cisco UCS B-Series Blade Servers



This section briefly describes the primary UCS components and how they function in a Unified Communications solution. For details about the Cisco UCS B-Series Blade Servers, refer to the model comparison at

http://www.cisco.com/en/US/products/ps10280/prod_models_comparison.html

Cisco UCS 5100 Series Blade Server Chassis

The Cisco UCS 5100 Series Blade Server chassis not only hosts the B-Series blade servers but also provides connectivity to the uplink Fabric Interconnect Switch by means of Cisco UCS Fabric Extenders.

Cisco UCS 2100 and 2200 Series Fabric Extenders

Cisco UCS 2100 and 2200 Series Fabric Extenders are inserted into the B-Series chassis, and they connect the Cisco UCS 5100 Series Blade Server Chassis to the Cisco UCS Fabric Interconnect Switch. The fabric extender can pass traffic between the blade server's FCoE-capable CNA to the fabric interconnect switch using Fibre Channel over Ethernet (FCoE) protocol.

Cisco UCS 6100 and 6200 Series Fabric Interconnect Switch

A Cisco UCS 6100 and 6200 Series Fabric Interconnect Switch is 10 Gigabit FCoE-capable switch. The B-Series Chassis (and the blade servers) connect to the fabric interconnect, and it connects to the LAN or SAN switching elements in the data center.

Cisco UCS Manager

Management is integrated into all the components of the system, enabling the entire UCS system to be managed as a single entity through the Cisco UCS Manager. Cisco UCS Manager provides an intuitive user interface to manage all system configuration operations.

Hypervisor

A hypervisor is a thin software system that runs directly on the server hardware to control the hardware, and it allows multiple operating systems (guests) to run on a server (host computer) concurrently. A guest operating system (such as that of Cisco Unified CM) thus runs on another level above the hypervisor. Hypervisors are one of the foundation elements in the cloud computing and virtualization technologies, and they consolidate applications onto fewer servers.

Storage Area Networking

Storage area networking (SAN) enables attachment of remote storage devices or storage arrays to the servers so that storage appears to the operating system to be attached locally to the server. SAN storage can be shared between multiple servers.

Design Considerations for Running Virtual Unified Communications Applications on B-Series Blade Servers

This section highlights some design rules and considerations that must be followed for running Unified Communications services on virtualized servers. Many Cisco Unified Communications applications support virtualization on a B-Series Blade server, such as:

- Cisco Unified Communications Manager (Unified CM)
- Cisco Unified CM Session Manager Edition
- Cisco Unity Connection
- Cisco IM and Presence
- Cisco Unified Contact Center Express
- Cisco Unified Contact Center Enterprise

For a full list of supported Cisco Unified Communications applications, refer to the documentation available at

http://www.cisco.com/go/uc-virtualized

Blade Server

The Cisco B-Series Blade Servers support multiple CPU sockets, and each CPU socket can host multiple multi-core processors. For example, one B200 blade has two CPU sockets that can host up to two multi-core processors. This provides the ability to run multiple Unified Communications applications on a single blade server.

Cisco Unified Communications applications should be run on dedicated blades that are not running any non-Unified Communications applications. Each Unified Communications application should be allotted dedicated processing and memory resources, so that the resources are not oversubscribed.

Hypervisor

The VMware ESXi Hypervisor is required to run virtual Unified Communications applications. The local hard drives attached to the Blade Server cannot be used to store virtual machines; they can be used only to install the ESXi hypervisor software. Unified Communications applications must follow the respective guidelines for their virtual machine template and configuration.

VMware vCenter is not mandatory when using a Tested Reference Configuration, but it is strongly recommended to manage multiple ESXi hosts for a large deployment.

For specific configuration and sizing requirements for virtual machines, refer to the respective product documentation available at

http://www.cisco.com/go/uc-virtualized

SAN and Storage Arrays

Tested Reference Configurations based on the Cisco UCS B-Series platform require the virtual machines to run from a Fibre Channel SAN storage array. The SAN storage array must satisfy the requirements of the VMware hardware compatibility list. Other storage options such as iSCSI, FCoE SAN, and NFS NAS are supported with the specification-based hardware support. For more details, refer to the documentation available at

http://www.cisco.com/go/uc-virtualized

Cisco UCS C-Series Rack-Mount Servers

Beside the B-Series Blade Servers, the Cisco Unified Computing System (UCS) also features general purpose rack-mount servers based on x86 architecture. The C-Series Rack-Mount Servers provide computing resources (memory, CPU, and I/O) and optional local storage to operating systems and applications. For more information on C-Series servers, refer to the documentation at

http://www.cisco.com/en/US/products/ps10493/index.html

Design Considerations for Running Virtual Unified Communications Applications on C-Series Rack-Mount Servers

Tested Reference Configurations are also available with Cisco UCS C-Series Rack Mount Servers such as the Cisco UCS C200, C210 and C260.

Many Cisco Unified Communications applications support virtualization on a C-Series Rack Mount Server, such as:

- Cisco Unified Communications Manager (Unified CM)
- Cisco Unified CM Session Manager Edition
- Cisco Unity Connection
- Cisco IM and Presence
- Cisco Unified Contact Center Express
- Cisco Unified Contact Center Enterprise

For a full list of supported Cisco Unified Communications applications, refer to the documentation available at

http://www.cisco.com/go/uc-virtualized

Unlike with the UCS B-Series, the Tested Reference Configurations based on the high-end UCS C-Series Rack Mount Servers (for example, C210 and C260) support storage for virtual machines either locally on the directly attached storage drives or on an FC SAN storage array. Multiple Unified Communications applications can reside on the same C-Series server. Low-end UCS C-Series Rack Mount Servers (for example, C200) allow only local storage of Cisco Unified Communications virtual machines.

UCS C210 servers support more user capacity than UCS C200 servers.

There are specific requirements that must be met in order to run Cisco Unified Communications applications as virtual servers on the UCS C-Series Rack-Mount Servers. These requirements are mentioned in the following document:

http://www.cisco.com/en/US/prod/collateral/voicesw/ps6790/ps5748/ps378/solution_overview_c22-597556.html

Impact of Virtual Servers on Deployment Models

Deploying Cisco Unified Communications applications on virtualized servers supports the same deployment models as when physical servers are used. The chapter on Network Infrastructure, page 3-1, offers some design guidance on how to integrate the QoS capabilities of Cisco UCS B-Series virtualized servers into the network. Also, the integration of physical servers (such as Cisco MCS servers) and Cisco UCS virtual servers is supported in many cases. As an example, music on hold (MoH) servers can run on Cisco MCS server platforms, while also being part of a cluster whose other member servers are run on Cisco UCS virtual servers.

All the call processing deployment models described in this chapter are supported on Cisco UCS virtual server platforms.

Design Considerations for Section 508 Conformance

Regardless of which deployment model you choose, you should consider designing your Cisco Unified Communications network to make the telephony features more accessible to users with disabilities, in conformance with Section 255 of the Telecommunications Act and U.S. Section 508.

Observe the following basic design guidelines when configuring your Cisco Unified Communications network to conform to Section 508:

- Enable Quality of Service (QoS) on the network.
- Configure only the G.711 codec for phones that will be connected to a terminal teletype (TTY) device or a Telephone Device for the Deaf (TDD). Although low bit-rate codecs such as G.729 are acceptable for audio transmissions, they do not work well for TTY/TDD devices if they have an error rate higher than 1% Total Character Error Rate (TCER).
- Configure TTY/TDD devices for G.711 across the WAN, if necessary.
- Enable (turn ON) Echo Cancellation for optimal performance.
- Voice Activity Detection (VAD) does not appear to have an effect on the quality of the TTY/TDD connection, so it may be disabled or enabled.
- Configure the appropriate *regions* and *device pools* in Unified CM to ensure that the TTY/TDD devices always use G.711 codecs.

- Connect the TTY/TDD to the Cisco Unified Communications network in either of the following ways:
  - Direct connection (Recommended method)

    Plug a TTY/TDD with an RJ-11 analog line option directly into a Cisco FXS port. Any Cisco voice gateway with an FXS port will work. Cisco recommends this method of connection.
  - Acoustic coupling

    Place the IP phone handset into a coupling device on the TTY/TDD. Acoustic coupling is less reliable than an RJ-11 connection because the coupling device is generally more susceptible to transmission errors caused by ambient room noise and other factors.
- If stutter dial tone is required, use an analog phone in conjunction with an FXS port on the Cisco VG224 or ATA 187. In addition, most Cisco IP Phones support stutter dial tone, which is sometimes referred to as audible message waiting indication (AMWI).

Call Routing and Dial Plan Distribution Using Call Control Discovery for the Service Advertisement Framework

When multiple call processing agents are present in the same system, each can be configured manually to be aware of the others. This configuration can be time consuming and error prone. Call routing between the various call processing agents requires the configuration of static routes on the call agents and updating them when changes occur.

Instead, the Cisco Service Advertisement Framework (SAF) can be used to share call routing and dial plan information automatically between call agents. SAF allows non-Cisco call agents (such as TDM PBXs) to partake in the Service Advertisement Framework when they are interconnected through a Cisco IOS gateway.

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

- SAF Clients — Advertise and consume information about services.
- SAF Forwarders — Distribute and maintain SAF service availability information.
- The SAF Client Protocol — Used between SAF Clients and SAF Forwarders.
- The SAF Forwarder Protocol — Used between SAF Forwarders.

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

Services that SAF Can Advertise

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

The dynamic nature of SAF and the ability for call agents to advertise the availability of their hosted DN ranges and To PSTN prefixes to other call agents in a SAF network, provides distinct advantages over other static and more labor-intensive methods of dial plan distribution.

This chapter discusses the deployment of Call Control Discovery (CCD) in SAF-enabled Unified Communications networks. For more information on SAF itself, see Service Advertisement Framework (SAF), page 3-69.

The following Cisco products support the Call Control Discovery (CCD) service for SAF:

- Cisco Unified Communications Manager (Unified CM) Release 8.0(1) or higher
- Cisco Unified Communications Manager Express (Unified CME) on a Cisco Integrated Services Router (ISR)
- Survivable Remote Site Telephony (SRST) on a Cisco ISR platform
- Cisco Unified Border Element on a Cisco ISR platform
- Cisco IOS Gateways on a Cisco ISR platform

CCD is supported on Cisco ISR platforms running Cisco IOS Release 15.0(1)M or higher. For more information on Cisco IOS Release 15.0(1)M, refer to the following websites:

- http://wwwin.cisco.com/ios/release/15mt
- http://www.cisco.com/en/US/products/ps10621/index.html

For information on the use of CCD with Unified CM, refer to the *Cisco Unified Communications Manager Features and Services Guide*, available at

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

SAF Service IDs

CCD is the first SAF service. SAF services are identified to a network of SAF Forwarders and Clients by their SAF Service ID. CCD for Unified Communications uses a SAF Service ID of 101:2:*x.x.x.x*, where:

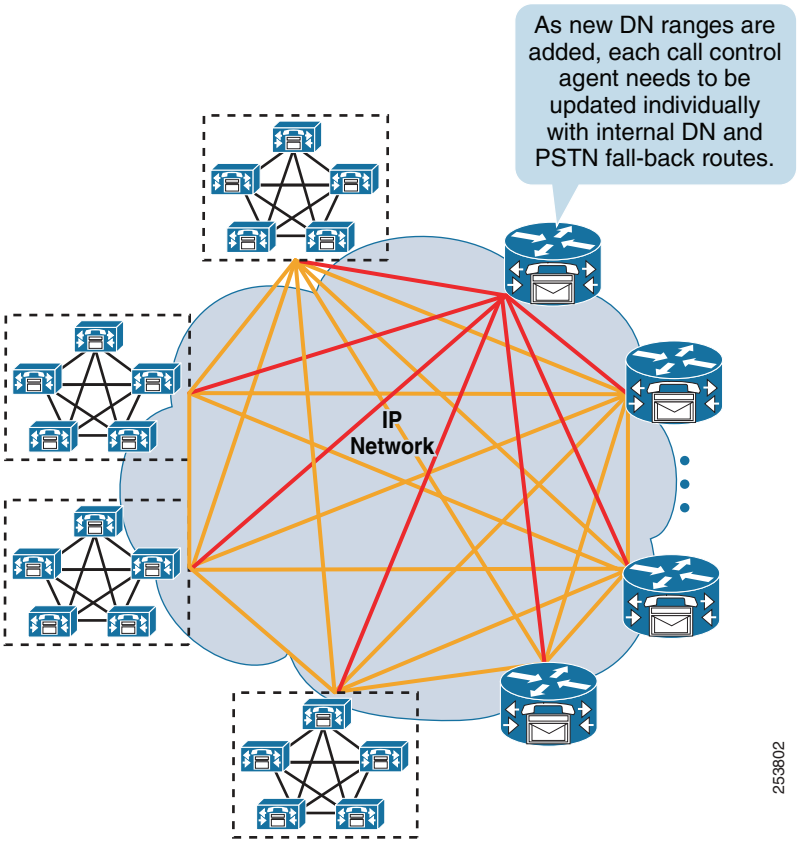
- Service ID 101 = Unified Communications
- Sub-Service ID 2 = CCD
- Instance ID *x.x.x.x* = ID of Unified CM cluster (PKID) or Cisco IOS device

Deploying SAF CCD Within Your Network

The SAF CCD service allows information about the location and availability of directory number ranges hosted by call control agents, such as Unified CM and Unified CME, to be propagated dynamically within a SAF-enabled Unified Communications network.

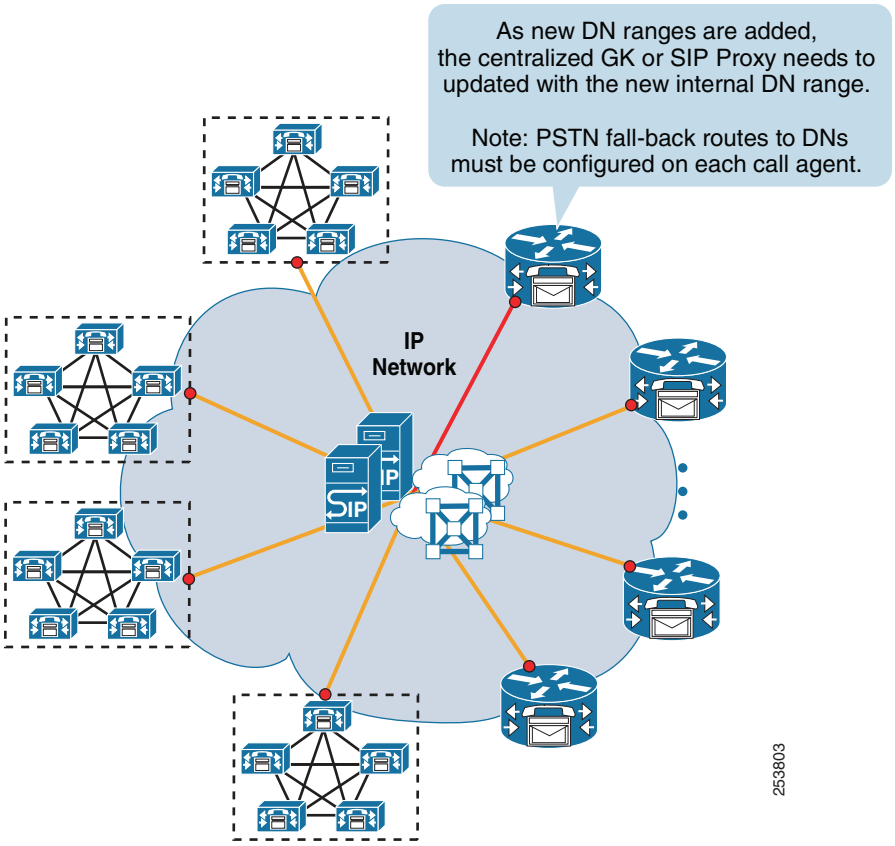
The advantages of deploying SAF to distribute and maintain DN information can be understood by considering the management of the dial plan in an example Unified Communications network consisting of four Unified CM clusters and 40 Unified CMEs. In a statically configured network, as new directory number ranges are introduced within the Unified Communications system, details of how those new number ranges can be reached must be made available to all other call control applications within the Unified Communications network. In the worst case, with a full mesh of connections between all call control applications, each call control application must be updated with information about each new number range and how it can be reached (see Figure 5-13). This cascade of configuration changes is time consuming, error prone, and requires significant ongoing management.

Figure 5-13 A Full Mesh of Connections Between Call Control Applications



The dial plan can be centralized on a Session Management Edition cluster, H.323 gatekeeper, or SIP proxy (see Figure 5-14). This reduces configuration overhead, but it allows only the internal dial plan to be centralized. If access to the centralized dial plan is unavailable, alternative routes such as PSTN routes can be used only if they are configured as backup routes in each call control application.

Figure 5-14 A Centralized Internal Dial Plan



SAF CCD enables each call control application to advertise its directory number ranges and their corresponding "To PSTN" prefixes to all other call control applications in the SAF network. (See Figure 5-15 and Figure 5-16.) In doing so, SAF CCD removes the following restrictions:

- The need for a centralized application that hosts the internal system-wide dial plan.
- The requirement to configure each call control application individually as new DN ranges and their corresponding "To PSTN" prefixes are added to the Unified Communications network.

Furthermore, SAF CCD is dynamic rather than static in nature. When DN ranges are deleted or IP connectivity is lost to the call control application, the SAF network automatically updates all other call control applications by withdrawing the routes to the unavailable DNs. Likewise, when connectivity is reestablished (or DN ranges are reconfigured), the SAF network updates all other call control applications, thus reinstating the routes to the DN ranges.

Figure 5-15 Advertising Unified CM Internal DN Ranges and Corresponding "To PSTN" Prefixes to the SAF Network

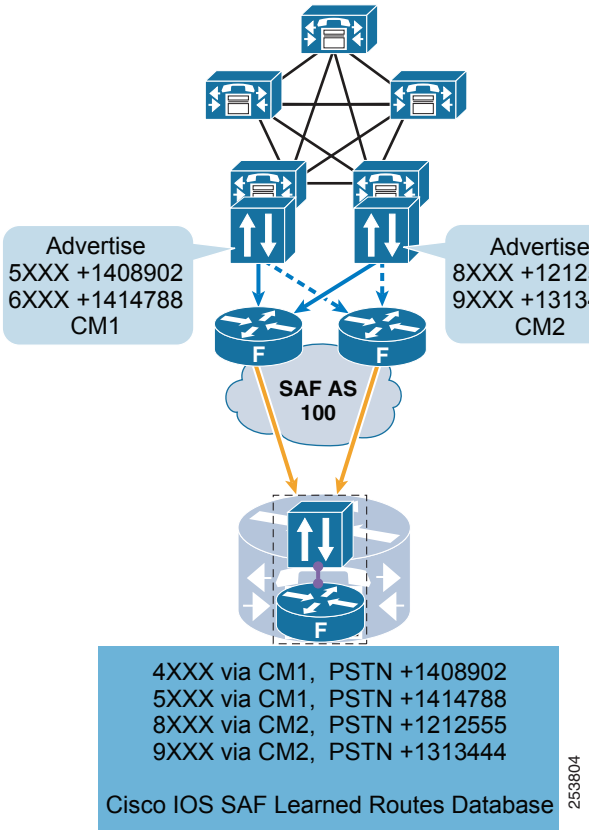
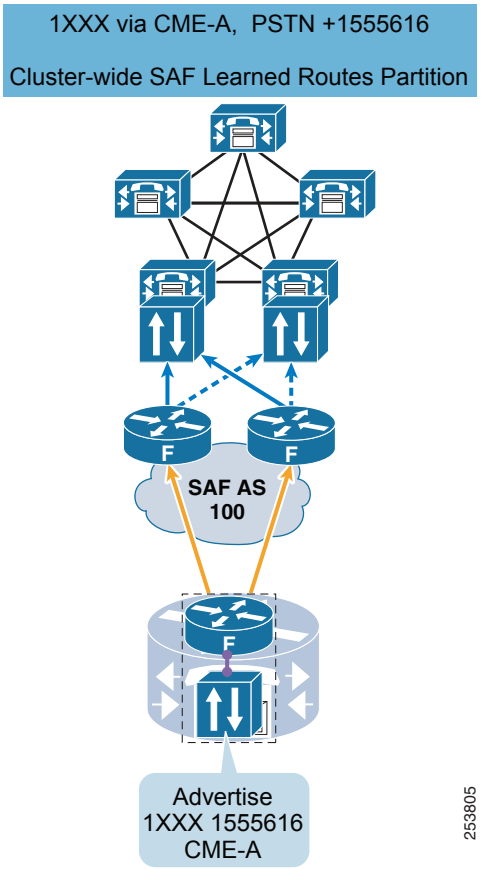


Figure 5-16 Advertising Unified CME Internal DN Ranges and Corresponding "To PSTN" Prefixes to the SAF Network



Comparison of SAF CCD Operation and Standard Unified CM Call Routing

Call routing using SAF CCD is fundamentally different than standard Unified CM call routing, which uses route patterns, route lists, and route groups that are not used by SAF CCD. Instead, the directory numbers, directory number ranges, and "To PSTN" prefixes to remote endpoints are learned dynamically by a SAF CCD-enabled cluster rather than being configured statically (see Figure 5-17). With SAF CCD, each Unified CM cluster (or other SAF-enabled call control application) configures which directory numbers, DN ranges, and so forth, that it wishes to advertise to the SAF network. SAF CCD also advertises the means by which to reach these numbers, by advertising the IP addresses and port numbers of the SAF-enabled SIP or H.323 trunks in the cluster.

Each SAF-enabled cluster also listens for advertisements from other clusters about their DNs, DN ranges, associated "To PSTN" Prefixes, and trunk information. These SAF learned routes are placed into a single partition. Any device that has access to this partition can reach any device advertised within SAF. Cisco recommends SAF CCD for the distribution of internal DN ranges only and their To PSTN routes.

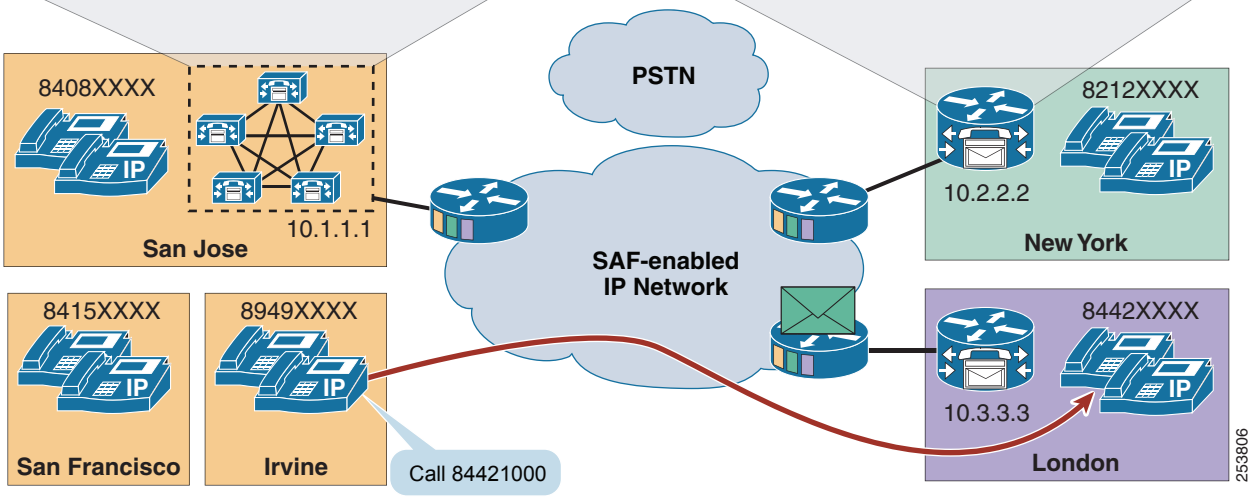
Figure 5-17 Dynamic Call Routing with SAF CCD

New York Unified CME Routing Table

DN Pattern	"to DID"rule	IP address	Protocol
8408XXXX	+1408555 /4	10.1.1.1	SIP
8415XXXX	+1415777 /4	10.1.1.1	SIP
8949XXXX	+1949222 /4	10.1.1.1	SIP
8442XXXX	**4:+442077111**	**10.3.3.3**	**H.323**

San Jose Unified CM Routing Table

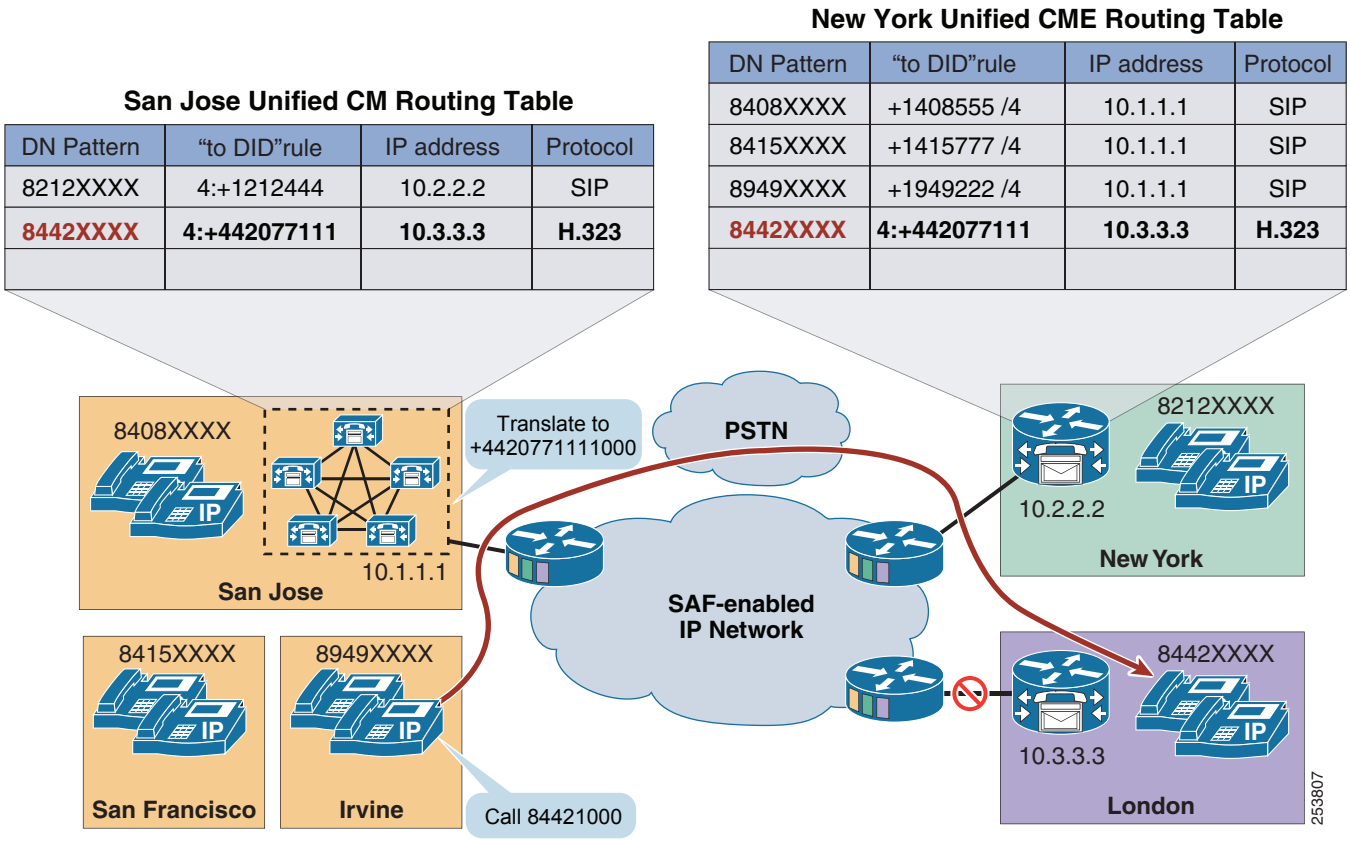
DN Pattern	"to DID"rule	IP address	Protocol
8212XXXX	4:+1212444	10.2.2.2	SIP
8442XXXX	**4:+442077111**	**10.3.3.3**	**H.323**



Any call made using SAF learned routes has automatic PSTN failover if the IP path to the called number is not available (see Figure 5-18). The call is routed according to the following order:

- Take the selected IP path to reach the called number.
- If the IP path is not available, use the PSTN prefix to modify the called number and route the call through the PSTN.

Figure 5-18 Automatic PSTN Failover with SAF CCD



SAF CCD is different than standard call routing in that only a single IP route can be chosen for a given SIP or H.323 call, whereas with standard call routing, multiple IP paths may be defined and consecutively attempted for a single call by using route lists and route groups.

CCD and Unified CM

CCD enables Unified CM to advertise multiple directory numbers, directory number ranges, and their corresponding "To PSTN" prefixes to a SAF-enabled network. CCD introduces several new configurable components in Unified CM:

- SAF Forwarder Configuration (the external SAF Client on Unified CM)
- SAF Enabled Trunks
- Hosted DN Patterns
- Hosted DN Groups
- CCD Advertising Service
- CCD Requesting Service

SAF Forwarder Configuration (External SAF Client on Unified CM)

The SAF Forwarder Configuration on Unified CM represents the configuration of the External SAF Client to a SAF Forwarder in a Unified Communications network. The Unified CM SAF Forwarder configuration defines the following items:

- The destination IP address and port number of the remote SAF Forwarder
- The Security Profile (username and password) used to authenticate with the SAF Forwarder
- The Client Label

  This is a string that the SAF Forwarder uses to map the Unified CM external client into a specific SAF Autonomous System. Cisco IOS supports bulk provisioning of the Client Label, whereby a client-label string that ends with an @ is considered as a base name or label. A base label configured on a router will accept any character following the @ in the base name as a valid client-label to identify a client in the REGISTER message sent by an external client.

For example, Unified CM cluster A can use CUCM-A as the base name for the cluster and can append a number after the @ following the base name for each configured SAF Forwarder (external SAF Client in Unified CM). By defining the external client CUCM-A as a base name in Cisco IOS, the Cisco IOS forwarder will accept any client label beginning with CUCM-A@, such as any of the following labels:

- CUCM-A@Client-1
- CUCM-A@Client-2
- CUCM-A@Client-3
- CUCM-A@Client-4

This allows SAF Clients 1 through 4 to register with the same SAF Forwarder and SAF autonomous system (AS).

External SAF Client Instance Creation and Activation within the Unified CM Cluster

By default, an instance of the external SAF client configured through the SAF Forwarder Configuration page in Unified CM is created on every call processing node within the cluster (see Figure 5-19). The external SAF client is activated only if an instance of a CCD Advertising Service or the CCD Requesting Service is also active on the call processing node. The activation of Advertising and Requesting Services on call processing nodes is determined by the SAF trunks associated with each service. (For details, see CCD Advertising and Requesting Services, page 5-63.)

Figure 5-19 Single SAF Forwarder Defined in Unified CM

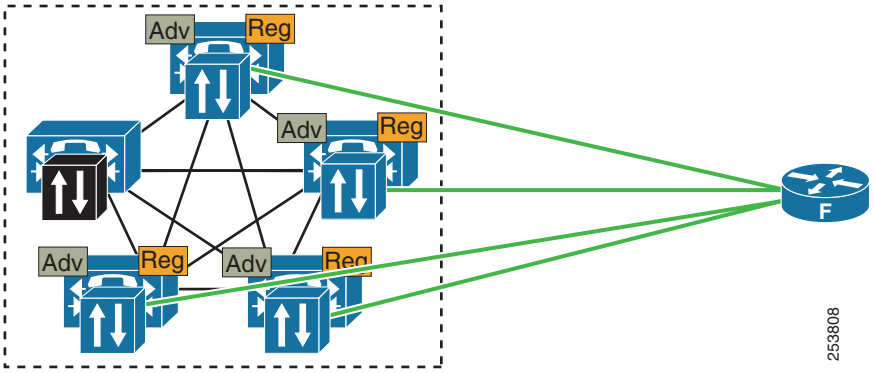
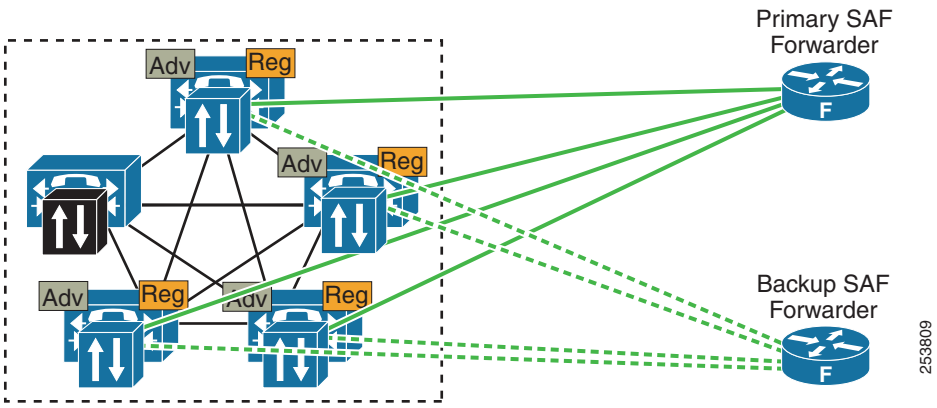


Figure 5-19 shows four active External SAF Clients connecting to a single SAF Forwarder. (The greyed-out SAF client is not activated because there is no active Advertising or Requesting Service associated with that Unified CM node). Each active External SAF client establishes a connection to the SAF Forwarder, registers with the SAF network, publishes its associated Services, and subscribes to the SAF CCD service active in the SAF AS. Such duplication can be useful for resilience and redundancy, but it can also create overhead within the cluster and the SAF Forwarder. By carefully selecting where the Advertising and Requesting Services run within the cluster, you can fine-tune this duplication and redundancy. For more information, see the CCD Advertising and Requesting Services, page 5-63.

Multiple SAF Forwarders

You can configure multiple SAF Forwarders within a cluster for redundancy. The SAF Client establishes a secure connection to the primary and the backup SAF Forwarders, registers with the SAF Forwarders, and sends a publish request for the HostedDN service to the primary SAF Forwarder. The SAF Client makes an arbitrary decision on selecting one SAF Forwarder as primary and another as backup at system startup time, based on the first SAF Forwarder to respond to a registration request from the client. The SAF Client publishes and subscribes services to the primary SAF Forwarder only. The SAF Client maintains the connection to the SAF Forwarder by sending keepalives to the SAF Forwarder at regular intervals. If the connection to the primary SAF Forwarder fails, the SAF Client switches to the backup SAF Forwarder, sending all the publish and subscription requests to the backup SAF Forwarder that it had sent to the primary SAF Forwarder.

Figure 5-20 Two SAF Forwarders Defined in Unified CM



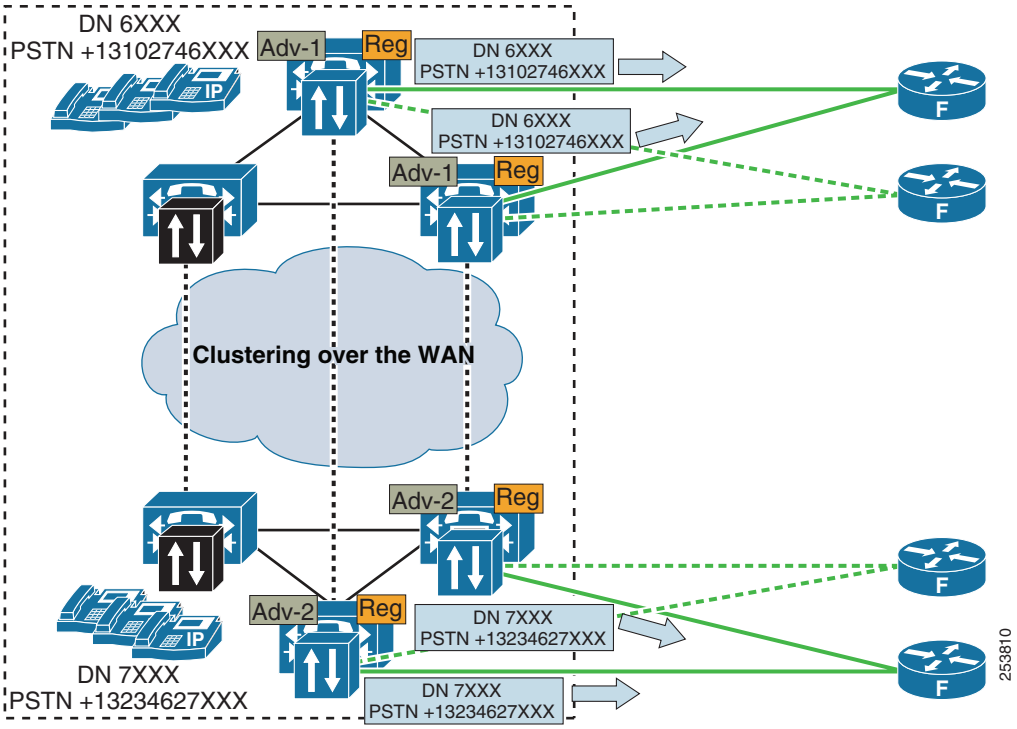
Advanced SAF Client Configuration

By default, for each configured SAF Forwarder a corresponding instance of the SAF client is created on every call processing node within the Unified CM cluster. Using the advanced SAF Forwarder configuration option, the administrator can create the SAF client on selected call processing nodes within the cluster. This configuration option enables the administrator to create the SAF client on specific nodes within the cluster and to configure SAF CCD with spatial distribution of CCD services for systems that employ clustering over the WAN.

SAF CCD and Clustering over the WAN

By creating multiple SAF client instances and multiple Advertising Services and associating them with specific Unified CM nodes within a cluster that uses clustering over the WAN, you can advertise CCD Hosted Directory Number ranges into the SAF network, with a geographical association to their local Unified CM trunks and nodes within the cluster.

Figure 5-21 SAF CCD-Selected SAF Client Configuration for Clustering over the WAN



SAF-Enabled Trunks

SAF-enabled trunks are used solely to route calls between SAF-enabled call control applications. They cannot be used with standard route patterns, route lists, and route groups. You cannot configure the destination address of a SAF-enabled trunk because this destination address is learned through SAF; however, you can configure all other trunk parameters.

You can enable SAF on the following trunk types:

- SIP trunks — Enabled by selecting **Call Control Discovery** as the Trunk Service Type when creating a new SIP trunk.
- H.323 Non-Gatekeeper controlled intercluster trunks — Enabled by checking the **Enable SAF** check box on the Trunk configuration page.

Both of these trunk types may be used between Unified CM clusters and between Unified CM and Cisco IOS gateways.

CCD uses SAF-enabled trunks for two purposes:

- To originate calls — These SAF-enabled trunks are associated with the CCD Requesting Service.
- To accept incoming calls — These SAF-enabled trunks are associated with the CCD Advertising Service. The IP addresses and port numbers of these SAF-enabled trunks are published with the DN ranges associated with the Advertising Service.

A SAF-enabled trunks can be used by both the Advertising and Requesting Service.

When the CCD Advertising Service publishes the trunk details for a hosted DN range, it sends the IP address and port number of each Unified CM node in the SAF trunk's Cisco Unified Communications Manager Group in separate SAF advertisements. For example, to advertise hosted DN range 5*XXX* from SIP trunk A, which has CUCM1 and CUCM2 in its Cisco Unified Communications Manager Group, the CCD Advertising Service would publish two advertisements:

- 5XXX via SIP trunk IP address (CUCM1) port number 5060
- 5XXX via SIP trunk IP address (CUCM2) port number 5060

The Requesting Service of the cluster receiving this advertisement would place two routes to 5XXX in its SAF learned routes partition:

- 5XXX via SIP trunk IP address (CUCM1) port number 5060
- 5XXX via SIP trunk IP address (CUCM2) port number 5060

Calls to 5XXX from this cluster would select the two available SIP trunk destinations in round-robin order.

SAF trunks support TCP or UDP transport protocols. Because a SAF trunk can accept incoming calls from multiple call control applications, TLS-based Signalling Authentication and Encryption is not supported over SAF-enabled trunks.

Hosted DN Patterns and Hosted DN Groups

Hosted DN groups represent groups of hosted DN patterns. The hosted DN patterns in a hosted DN group typically represent the range of directory numbers associated with a physical site. Digit strip and prepend information for "To PSTN" failover routing can be configured for each hosted DN group. The same DN pattern cannot be associated with multiple hosted DN groups.

A hosted DN pattern can define a single directory number (for example, 5000), or a range of directory numbers (for example, 5*XXX*). Every DN pattern must be unique. Each hosted DN pattern can be configured with digit strip and prepend information for PSTN failover routing. The PSTN failover configuration on the hosted DN pattern takes precedence over the PSTN failover configuration at the hosted DN group level.

CCD Advertising and Requesting Services

CCD uses two Unified CM services to communicate with the SAF network: the Advertising Service, which is used to publish DN ranges and their associated trunks to the SAF network, and the Requesting Service, which is used to learn about the reachability of DN ranges from other call agents in the SAF network. The following sections describe these two services.

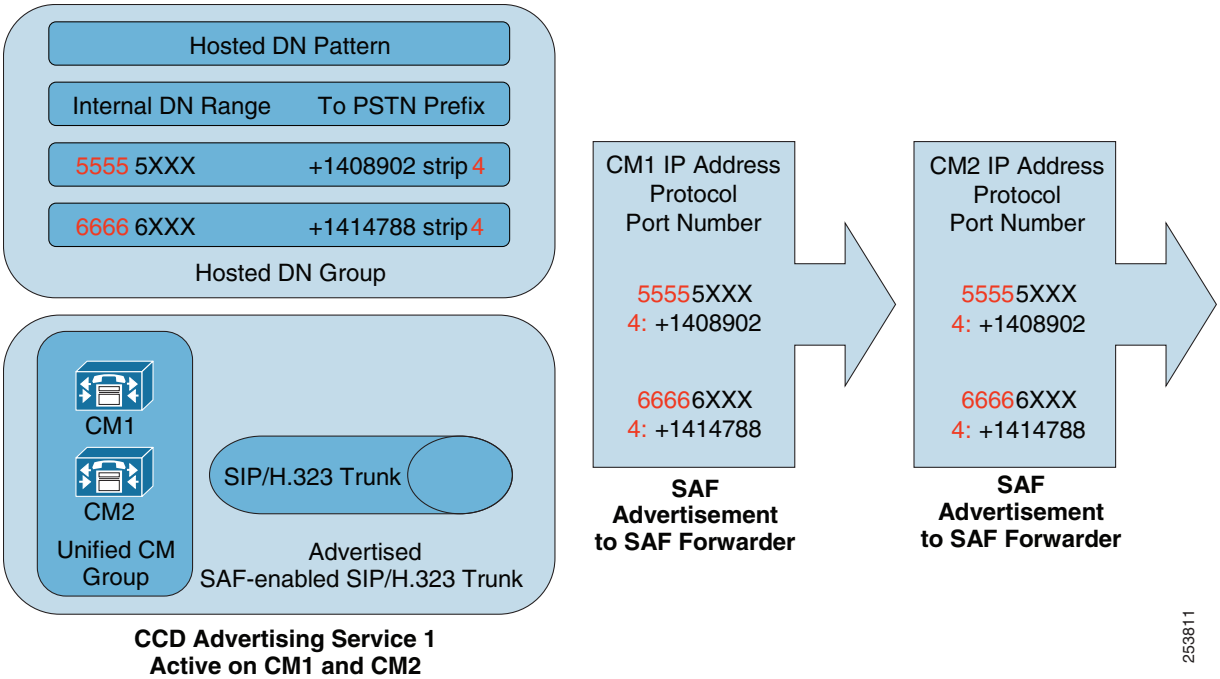
CCD Advertising Service

The CCD Advertising Service associates one hosted DN group with a SAF-enabled SIP and/or H.323 trunk. The Advertising Service is created and activated on each server in the Cisco Unified Communications Manager Group (Unified CM Group) of its associated SAF-enabled trunk(s). The

Advertising Service uses the SAF Client on each of the servers in the Unified CM Group of each trunk to publish information about the group of hosted DNs and associated trunk nodes to the client's SAF Forwarder. (See Figure 5-22.)

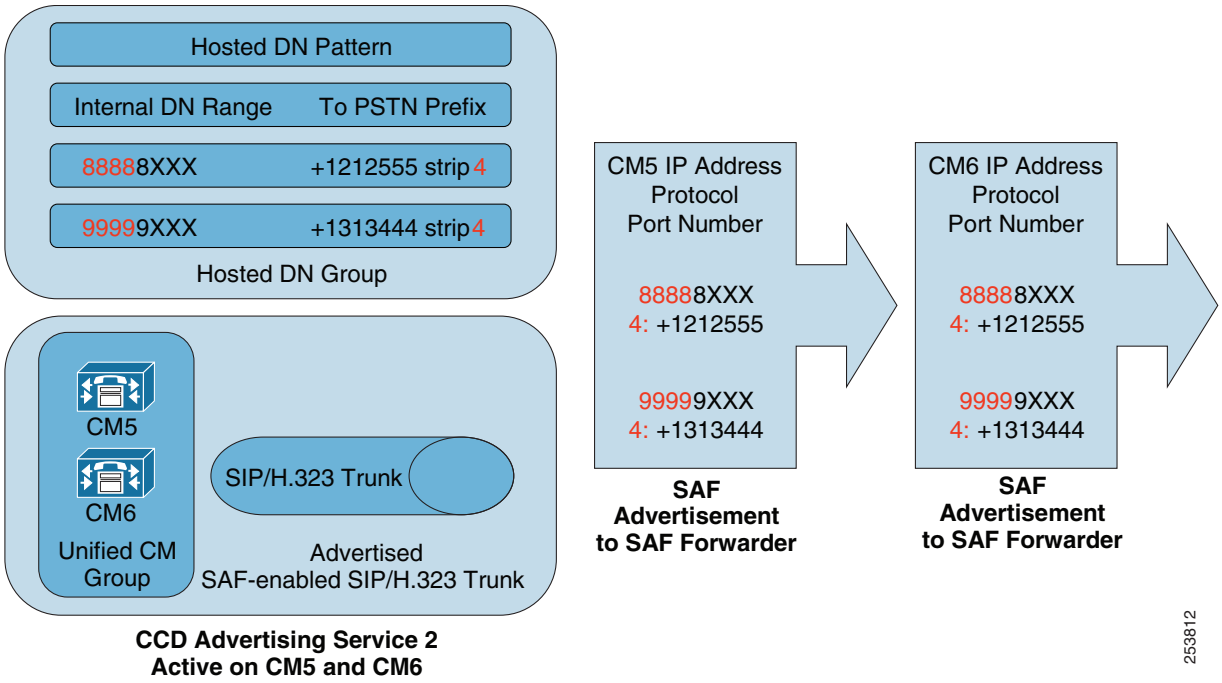
Because SIP and H.323 trunks support different feature sets (for example, H.323 trunks support QSIG over Annex M1), it is typical to select only one trunk type per Advertising Service. If both an H.323 and a SIP trunk are selected, calls to the hosted DN ranges associated with this Advertising Service will be distributed in a round-robin fashion across both the SIP and H.323 trunks.

Figure 5-22 CCD Advertising Service 1 Active on CM1 and CM2



You can create multiple advertising services within a Unified CM cluster. An Advertising Service can use the same (or different) SAF-enabled trunks as other Advertising Services. However, each Advertising Service must be associated with a unique hosted DN group, and the same hosted DN pattern cannot be advertised by multiple Advertising Services within a cluster. Creating multiple Advertising Services allows inbound calls to be distributed by DN range across multiple trunk servers within a cluster. (See Figure 5-23.)

Figure 5-23 CCD Advertising Service 2 Active on CM5 and CM6

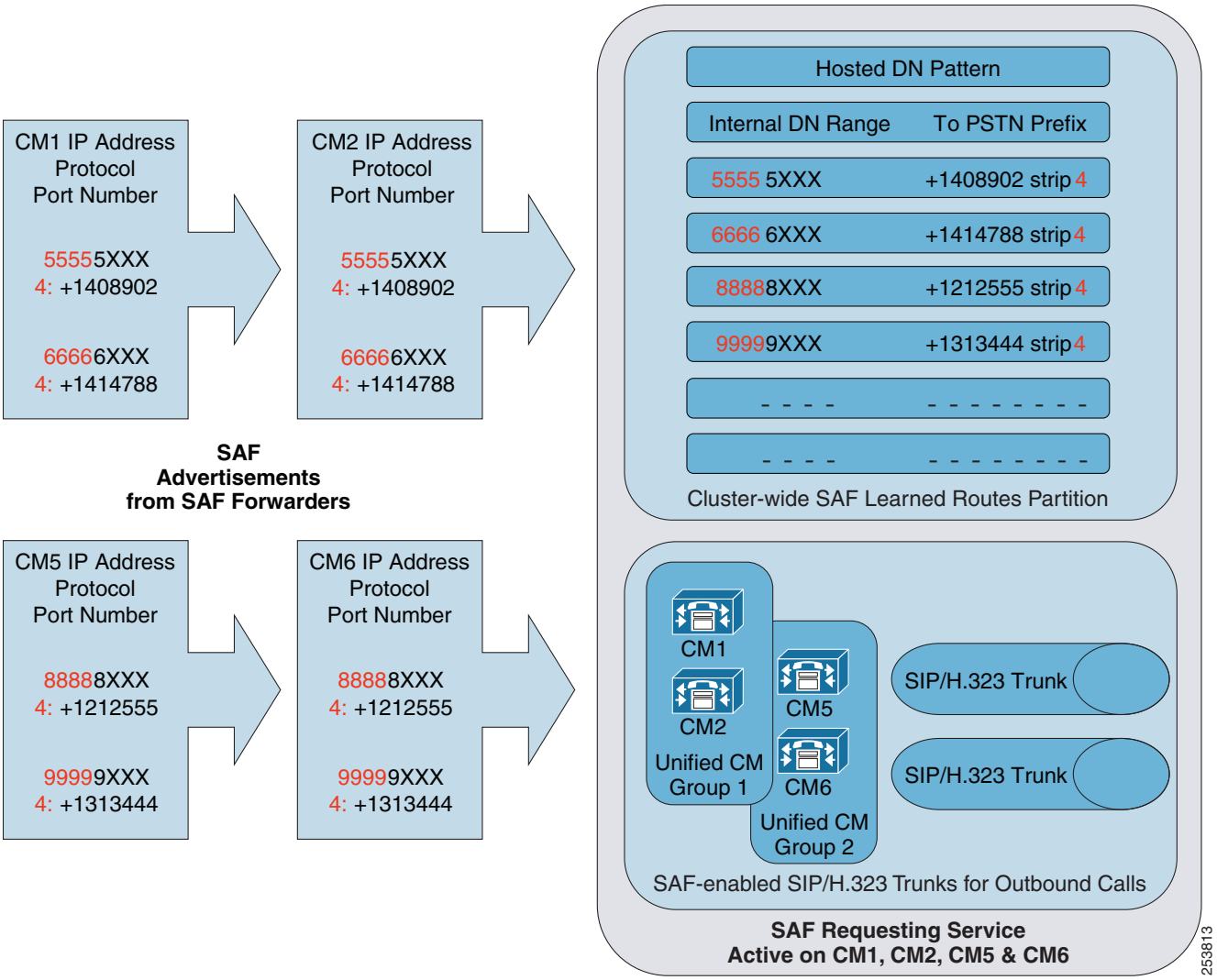


CCD Requesting Service

The CCD Requesting Service collects information about hosted DN routes advertised in the SAF AS and places them into a partition for SAF learned routes. (See Figure 5-24.) The Requesting Service is also used to select which SAF trunks will be used to initiate outbound SAF calls. More than one SAF-enabled trunk can be selected. If multiple trunks are selected, these SAF trunks and their corresponding Unified CM Group server nodes are selected on a round-robin basis for outbound calls. Similar to the Advertising Service, trunks of the same protocol type are usually associated to the Requesting Service. The Requesting service also allows digits to be prefixed to learned DN patterns and learned "To PSTN" patterns.

Only a single Requesting Service can be configured in the Unified CM cluster, and the Requesting Service is activated on all of the nodes in the Unified CM Groups of its associated SAF trunks.

Figure 5-24 Unified CM CCD Requesting Service



Blocking CCD Learned Patterns

Unified CM enables the SAF CCD administrator to purge and block learned route information from the SAF CCD learned routes partition. Routes can be blocked based on whether they match one or more of the following entries:

- Learned Pattern (for example, 500*X*)
- Learned Pattern Prefix (for example, +1408)
- Remote Call Control Entity Name (This is the Unified CM Cluster ID in Enterprise Parameters.)
- Remote Call Control IP Address (This could be the address of a Cisco IOS SAF CCD router or one or more Unified CM servers in a Unified CM cluster.)

If required, these entries can be used in a logical AND combination such as the following:

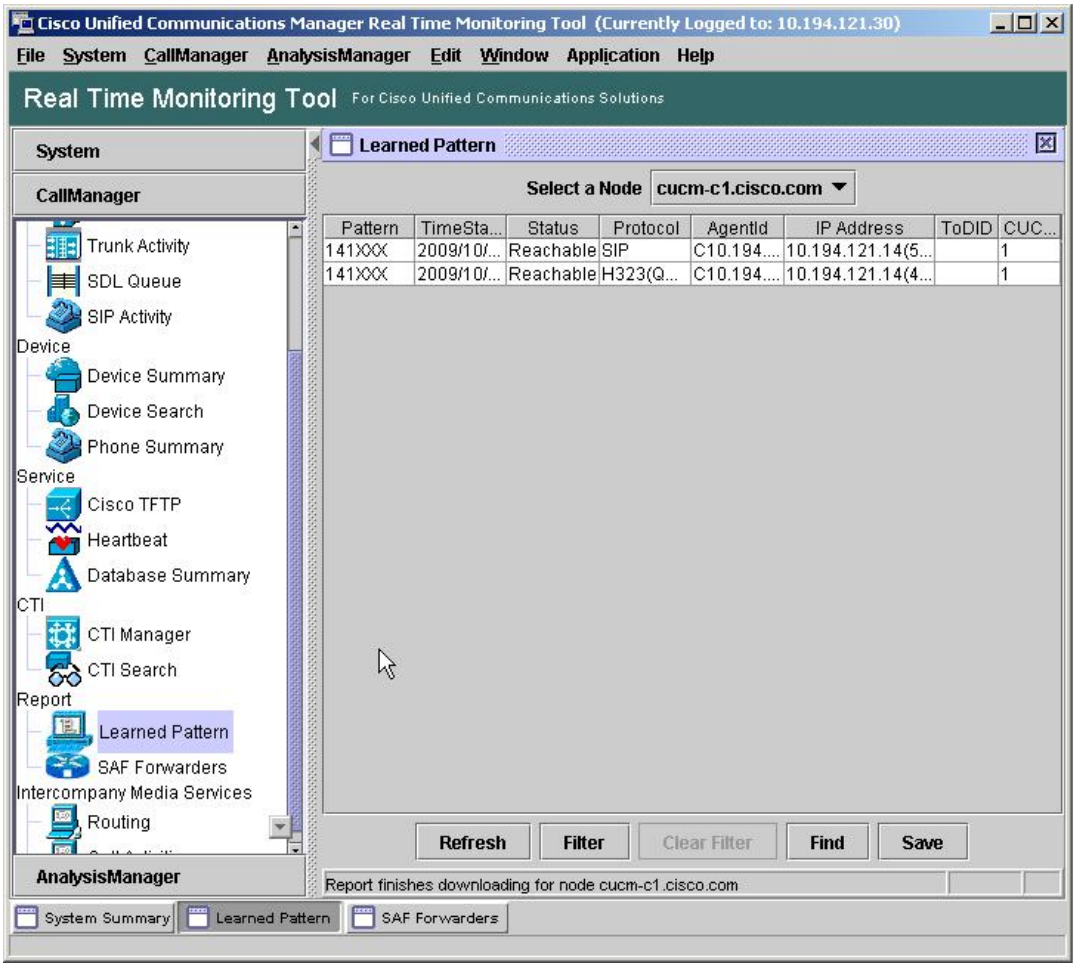
Pattern = "5XXX" AND Prefix = "+1408" AND Remote Call Control Address = "10.10.1.1"

Blocking CCD learned patterns can be particularly useful in SAF CCD deployments where a Unified CM cluster connects to multiple SAF ASs and wishes to advertise DN route information to an AS but does not wish to receive some or all of the DN route information being sent by the AS.

Displaying SAF Learned Routes in Unified CM

Because SAF learned routes are dynamic in nature, they are not held in the Unified CM database but are stored in memory. Use the Cisco Unified Communications Manager Real-Time Monitoring Tool (RTMT) to display SAF learned routes and to monitor SAF Forwarders (see Figure 5-25).

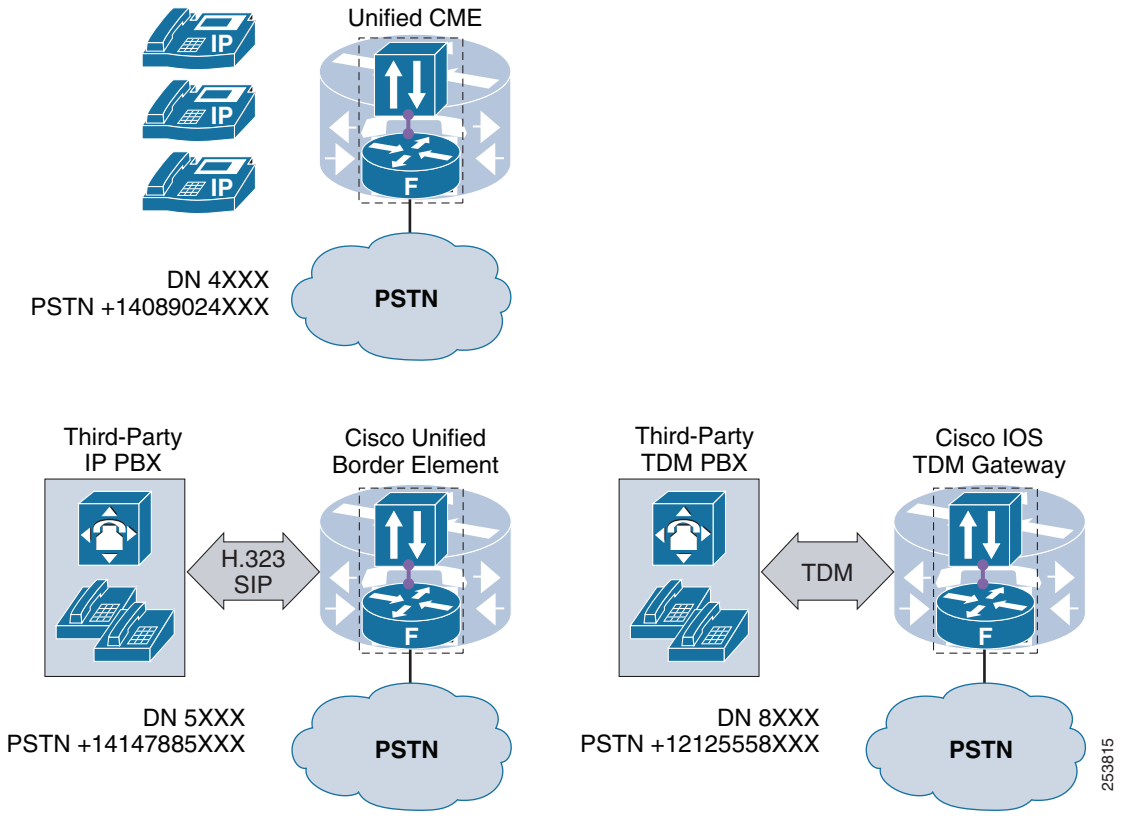
Figure 5-25 Real-Time Monitoring Tool (RTMT) for SAF CCD



Cisco IOS-Based SAF CCD

Cisco IOS-based SAF CCD is supported by Unified CME, SRST, Cisco Unified Border Element, and Cisco IOS Gateways on the Integrated Services Router (ISR) platform with Cisco IOS Release 15.0(1)M. (See Figure 5-26.) The configuration of Cisco IOS SAF CCD is the same across all of these products. SRST, however, is a special case of CCD and is discussed in the section on SAF CCD and SRST, page 5-71.

Figure 5-26 Cisco IOS-Based SAF CCD Call Agents



For Unified CME, Cisco IOS TDM gateways, and Cisco Unified Border Element, SAF CCD can be used to advertise the internal directory number ranges and "To PSTN" prefixes of the endpoints associated with each of these products and also to subscribe to SAF advertisements from other SAF CCD-enabled call control applications.

For both Cisco IOS and Unified CM, Cisco does *not* recommend using SAF CCD to advertise external PSTN number ranges (for example, for tail-end hop off) for the following key reasons:

- SAF CCD provides no information about the capacity of IP, PSTN, or TDM trunks. (For example, an ISDN BRI with two DS0s and a T1 TDM interface with 24 DS0s would be weighted equally by SAF CCD.)
- All SAF CCD routes are placed into a single partition. This means that any SAF CCD user has access to all learned SAF CCD routes and that no SAF CCD classes of service can be created.

Although the principles of Cisco IOS SAF CCD configuration are the same as those for Unified CM, the naming conventions and commands are different.

Internal SAF Clients

For Cisco IOS-based SAF CCD applications, the SAF Client and Forwarder are co-resident within Cisco IOS. Configuration and authentication is not required between the internal SAF Client and internal SAF Forwarder.

External SAF Clients

To enable the authentication of an external SAF Client to a Cisco IOS SAF Forwarder, use the **external-client** Cisco IOS command to define the external client's label or base name, username, password, and keepalive timer.

SAF-Enabled Trunks

SAF trunks are defined under the **profile trunk-route** Cisco IOS command. The trunk-route profile defines the IP address, port number, protocol (SIP or H.323), and transport protocol (UDP or TCP) for the SAF trunk.

DN Patterns, DN Blocks, and DN Service

The definition and configuration of directory numbers, DN ranges and "To PSTN" prefixes is slightly different in Cisco IOS when compared with Unified CM configuration. Cisco IOS uses the concept of DN blocks to group DN numbers and DN ranges. A DN block can contain more than one DN pattern. The "To PSTN" failover rules for stripping and prefixing digits are also defined at the DN block command line. The PSTN failover rule is known as an **alias** in Cisco IOS. (The PSTN failover rule is applied to the concatenated Site Code and Extension DN Pattern.) The following example shows the Cisco IOS configuration for a DN block:

```
profile dn-block 1 alias 1408902 strip 3
    pattern 1 extension 5xxx
    pattern 2 extension 6xxx
```

Call Control Profile, DN Service, and Site Code

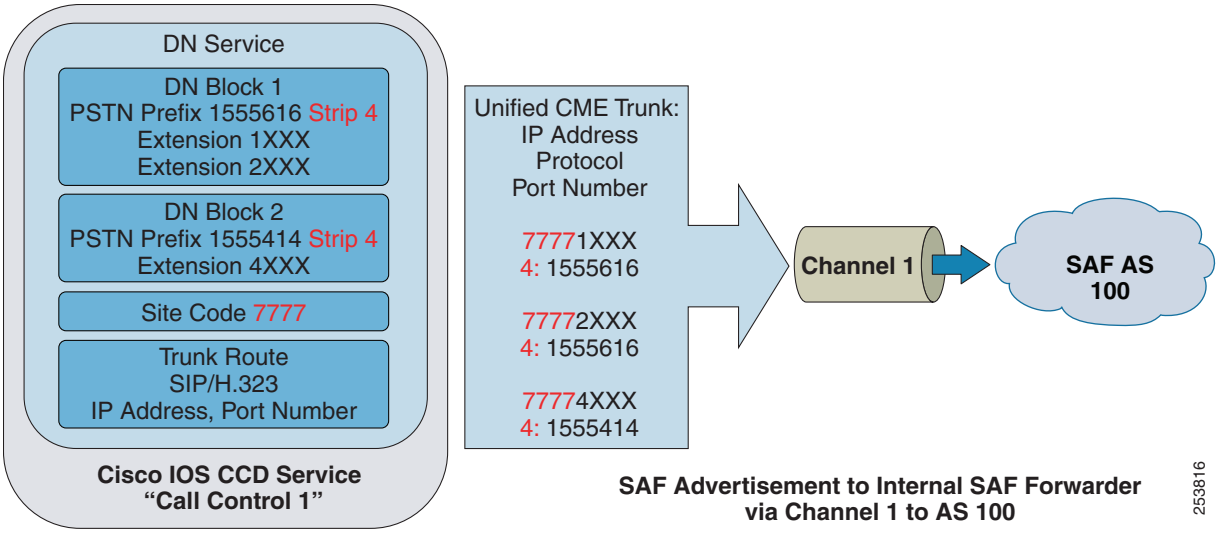
The CCD call control profile is associated with a DN service. A DN service in Cisco IOS can be considered to be equivalent to an Advertising Service in Unified CM. The DN service is used to group one or more DN blocks, one trunk route, and one site code. If present, the site code consists of one or more digits that are prepended to the advertised extension DN patterns.

Multiple call control profiles can be created. The same DN blocks, trunk routes, and site codes can be reused in multiple call control profiles, but only one profile can be associated with a SAF AS.

Publishing and Subscribing to SAF Services within a SAF AS

Call control profiles advertise their associated DN ranges, "To PSTN" failover rules, and trunk route to one SAF AS by means of a configured SAF "channel." A SAF channel can publish the CCD service information contained in only one call control profile to a single SAF AS. (See Figure 5-27.)

Figure 5-27 Cisco IOS CCD Service Call Control 1 Advertising Through Channel 1 to SAF AS 100



A SAF Channel can subscribe to all CCD services within a SAF AS using a wildcard service ID, or up to two selected SAF CCD services that are identified by the instance values in the SAF service ID. (The instance value for Unified CM is the cluster PKID.) For example:

Wildcard SAF Service ID =

Service:	Sub-service:	Instance.	Instance.	Instance.	Instance.
101:	2:	FFFFFFFF.	FFFFFFFF.	FFFFFFFF.	FFFFFFFF.

Tip Use the Cisco IOS command **show eigrp service-family ipv4 [***AS number***] events** to display the Service ID for the Cisco IOS SAF CCD service on the router. The Service ID will be displayed as "connected" (for example, 101:2:59F8412.0.0.6F0100).

Outbound SAF CCD Calls in Cisco IOS

Cisco IOS adds SAF as a configurable session target to standard Cisco IOS voice dial peers. Dial peers can also be assigned a preference setting to control the order in which standard and SAF dial peers are selected.

SAF CCD and SRST

SRST CCD is a special type of SAF deployment. SRST CCD does not advertise any number ranges into SAF; it only listens to the advertisements from other SAF CCD services such as Unified CM, Unified CME, and so forth. SRST CCD does not use SAF learned IP routes at any time; only PSTN routes are used and only when the router and associated phones are in SRST mode.

You can use SAF for SRST CCD to avoid the labour-intensive task of updating every SRST router with a new number expansion rule every time a new SRST router is added to the Unified Communications network.

With standard (non-SAF) SRST operation, if Unified CM becomes unavailable, phones register their extension numbers to their SRST reference router. (See Figure 5-28.) In SRST mode, calls can be made to other phones registered to the SRST router by dialing their extension number as normal. When a phone in SRST mode is used to call a phone in another site, the PSTN number of the called phone must be dialed. (See Figure 5-29.) The number expansion command in Cisco IOS, much like the PSTN failover rule in SAF CCD, allows the dialed extension number to be expanded to the full PSTN number in SRST mode.

In a Unified Communications deployment with many SRST routers, when a new SRST router is added to the Unified Communications network, every SRST router must add a number expansion rule that corresponds to the PSTN access prefix for this new SRST site.

SAF for SRST CCD allows the PSTN failover rules for every SRST site to be distributed to every SRST router within the SAF AS.

Figure 5-28 Normal (Unified CM) Operation of a Unified CM Deployment with SAF SRST CCD

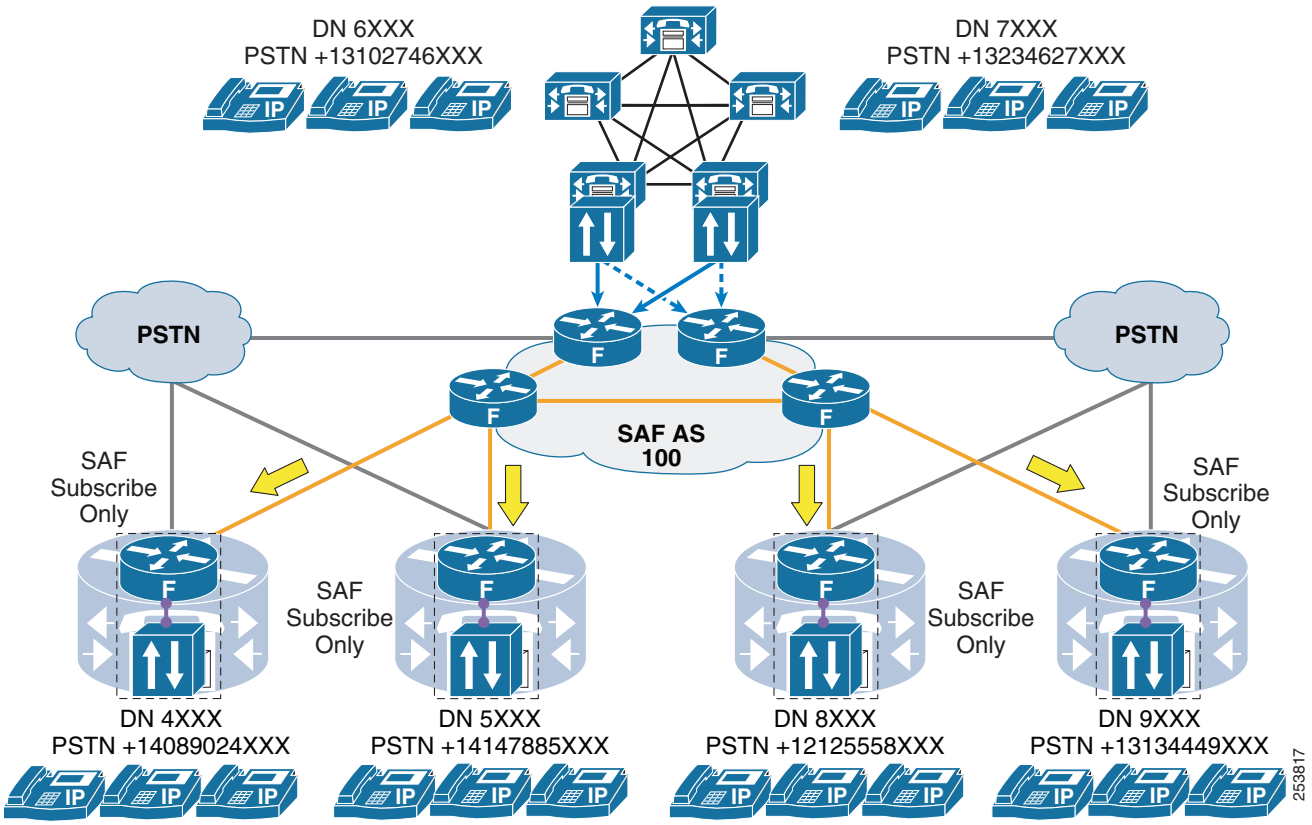
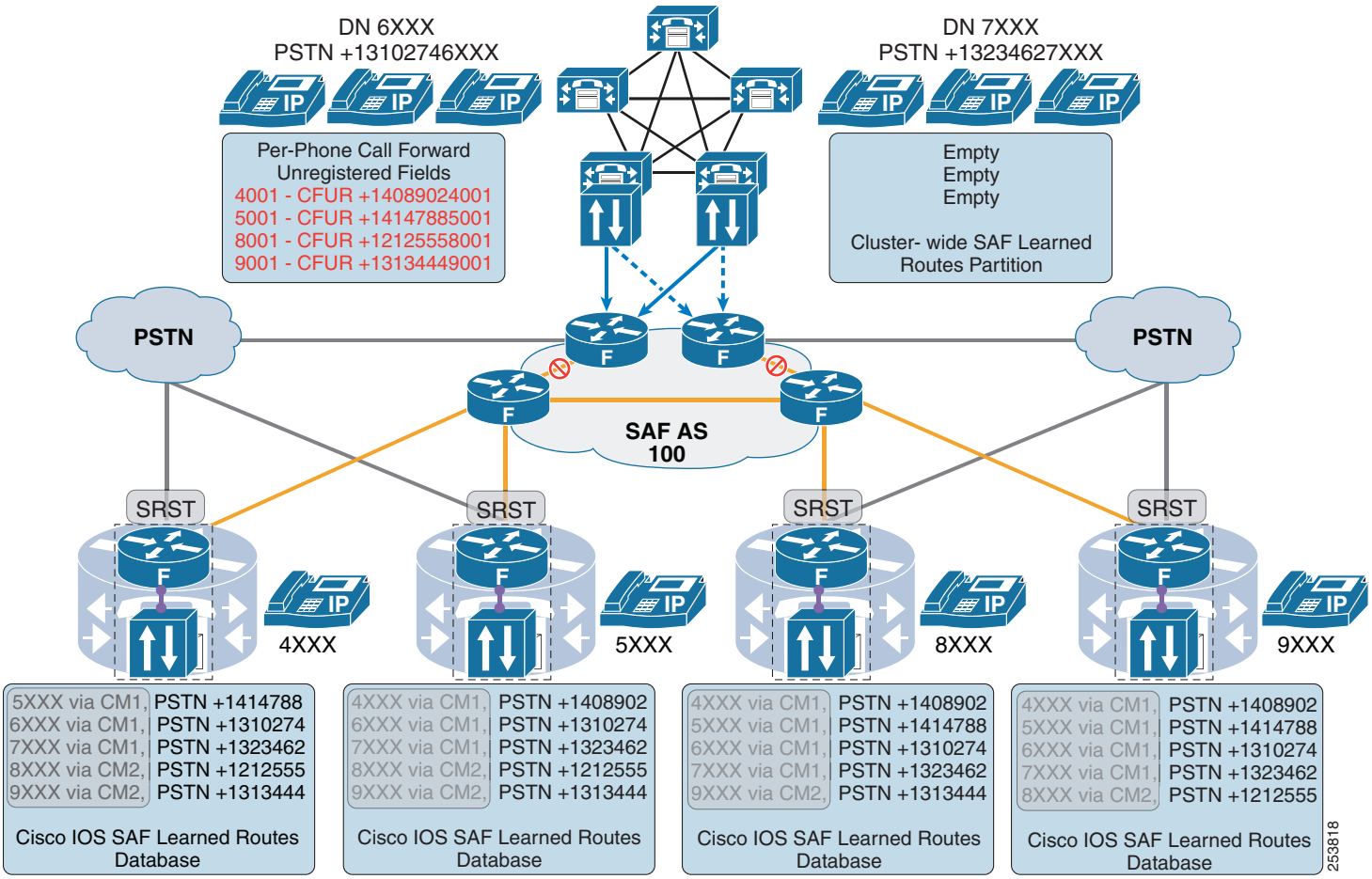


Figure 5-29 SRST Operation of a Unified CM Deployment with SAF SRST CCD



Typical SAF CCD-Based Unified Communications Deployments

Figure 5-30 show a typical SAF CCD network deployment.

Figure 5-30 A Global SAF Network with Regional Call Agents and SAF Clients and Forwarders

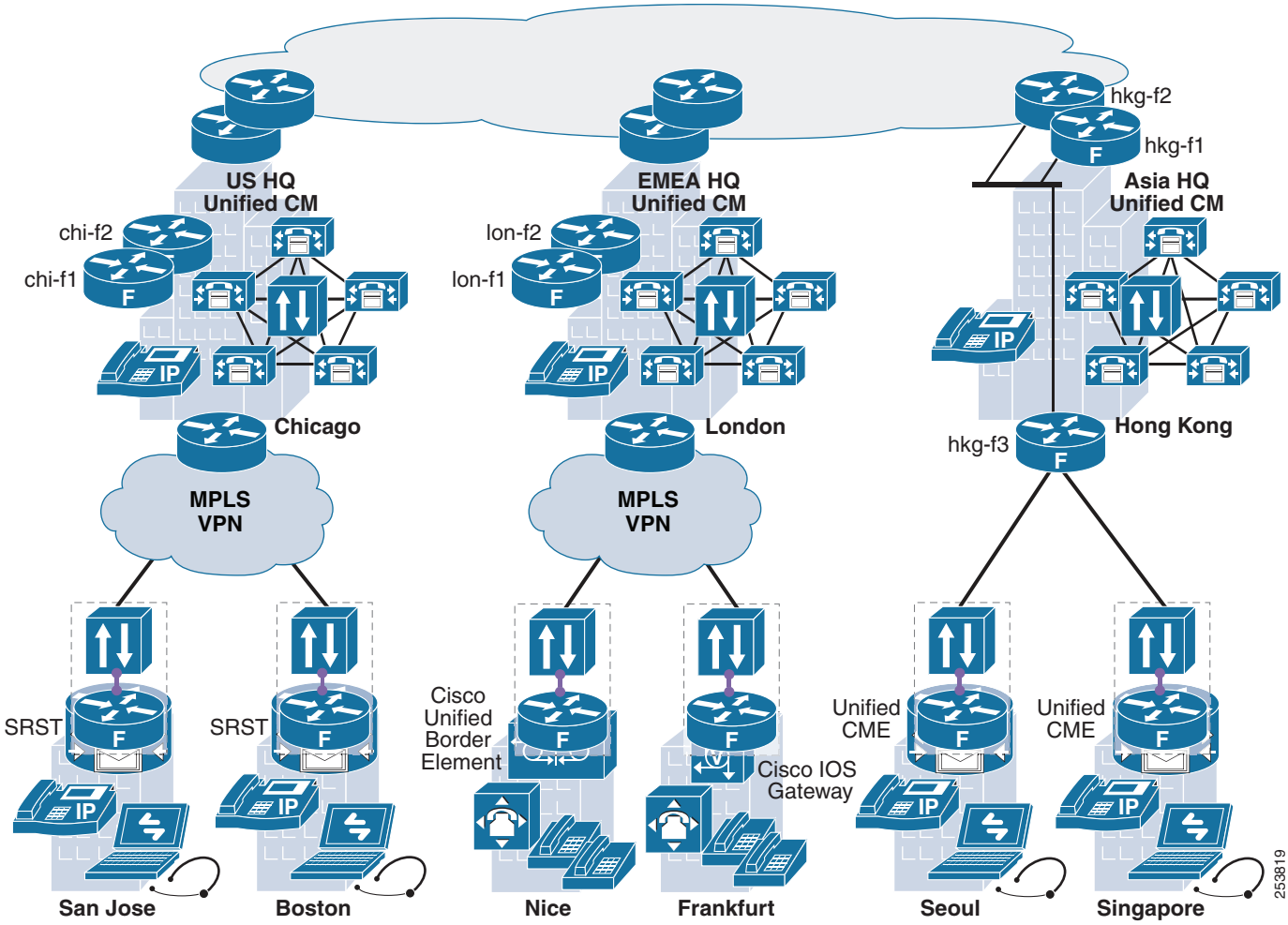
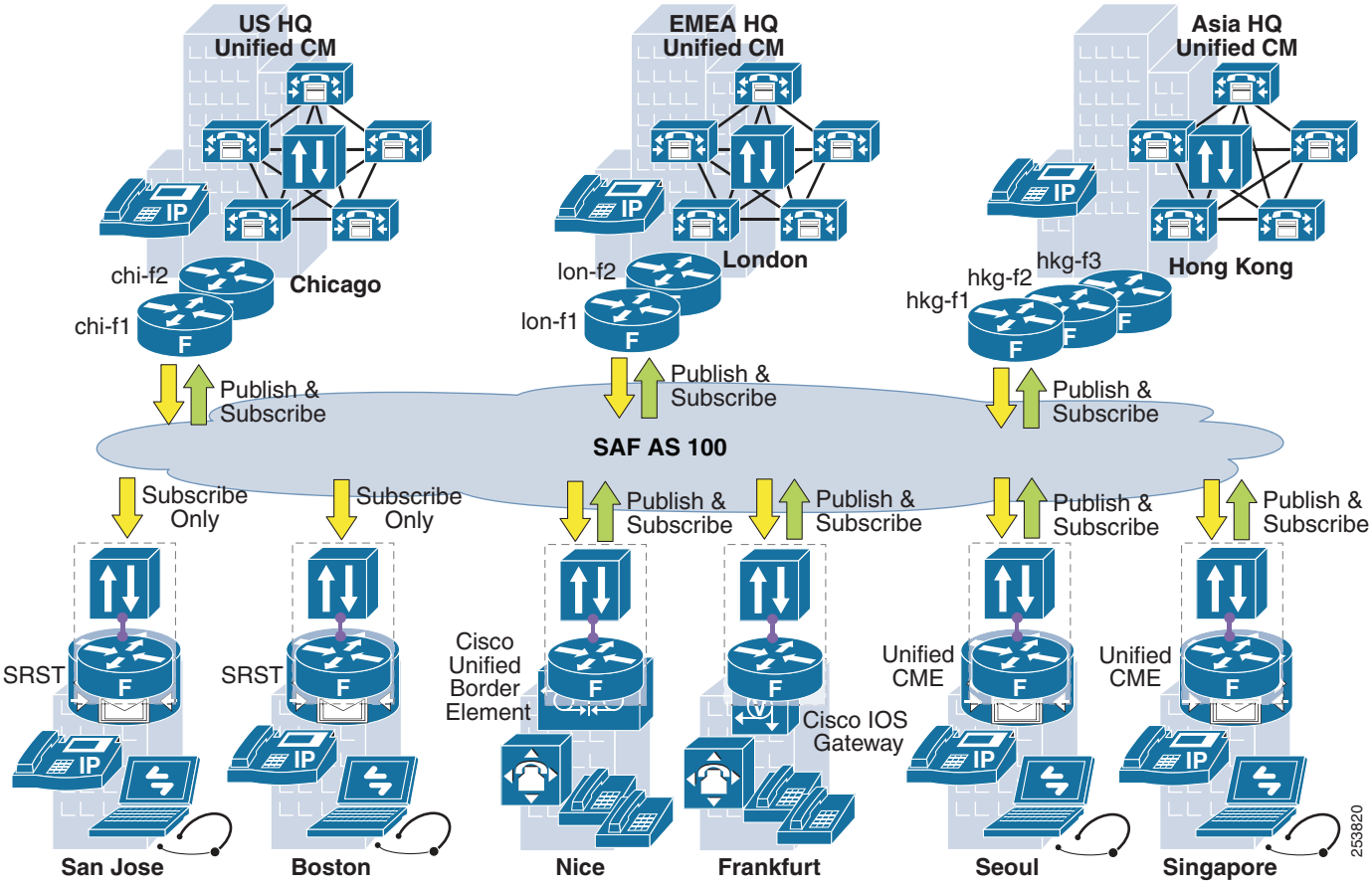


Figure 5-31 shows a logical diagram of the same global SAF network with regional call agents and SAF Clients and Forwarders

Figure 5-31 Logical Representation of Global SAF Network with Regional Call Agents and SAF Clients and Forwarders



SAF CCD Deployment Considerations

Migration to SAF CCD is relatively risk free. The SAF CCD network can be built and tested for basic operation and scalability before any devices that use SAF are enabled in the network. Unified CM users can be given the capability to use the SAF CCD network by adding the SAF Learned Routes Partition to their device or profile. In Cisco IOS the preference for SAF dial peers can be prioritized above standard dial peers. This allows SAF to be enabled incrementally throughout the network.

The following scalability limits apply to Unified CM and Cisco IOS SAF CCD products:

- Up to 2,000 advertised DN patterns per Unified CM cluster
- Up to 20,000 learned DN patterns per Unified CM cluster
- Up to 125 advertised DN patterns per Unified CME, Cisco Unified Border Element, or Cisco IOS Gateway
- Up to 6,000 learned DN patterns per Unified CME, Cisco Unified Border Element, Cisco IOS Gateway, or SRST (platform-dependant)

Note For SAF deployments using a single SAF AS and consisting of Cisco Unified CM and Cisco IOS SAF CCD systems, SAF CCD system-wide scalability is limited to 6,000 learned DN patterns.

In very large SAF CCD networks, multiple SAF ASs can be used to limit the distribution of SAF advertised DN patterns. Unified CM and/or Cisco Unified Border Element may also be used to manually summarize SAF advertisements from one SAF AS and statically advertise them into another SAF AS.

SAF CCD Port Numbers

SAF CCD uses the following port numbers:

- SAF EIGRP — IP Protocol 88
- Unified CM SAF Client to Cisco IOS SAF Forwarder — TCP port 5050 (configurable)
- Advertised SIP trunks — port 5060
- Advertised H.323 — ephemeral port number

Note The Cisco Adaptive Security Appliance (ASA) firewall uses standard SIP inspection and fix-up to open pinholes in the firewall for the RTP media streams of SAF enabled SIP trunk calls. H.323 inspection and fix-up of SAF-enabled H.323 trunk calls are not supported.

Cisco Intercompany Media Engine

Cisco Intercompany Media Engine (IME) is another variation of a multisite deployment with distributed call processing; however, with IME the sites are separate enterprise organizations. The term *boundary-less* Unified Communications is used to describe this technology because it allows for the business-to-business extension of Unified Communications capabilities such as high-fidelity codecs, enhanced caller ID, and video telephony outside the corporate networks. The solution learns routes in a dynamic, secure manner and provides for secure communications between organizations across the internet. Organizations that work closely together and have high levels of intercompany communications will benefit most from the enhanced communications offered by IME. This section discusses the components of the solution and the high-level architecture, with relevant design considerations for deploying IME.

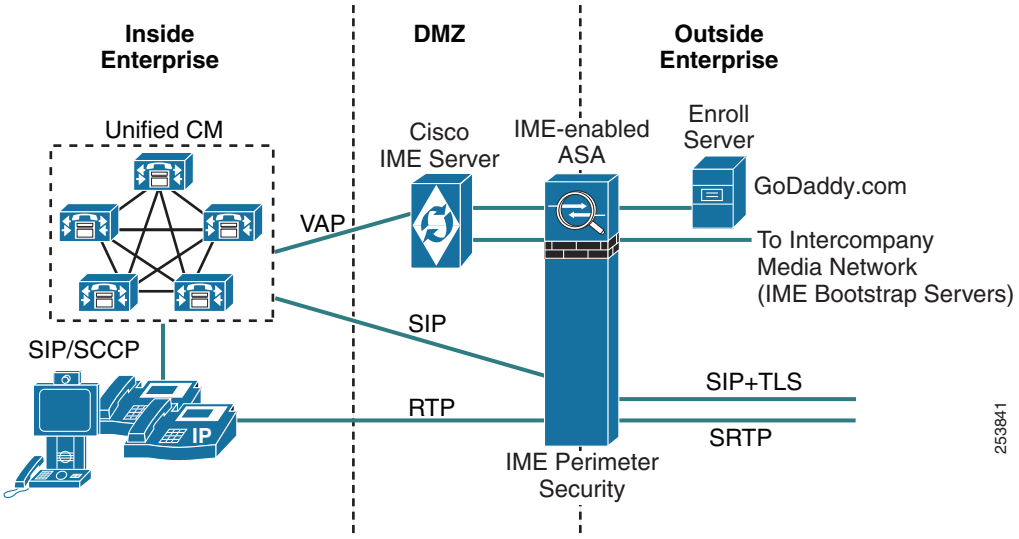
IME Components

The IME solution consists of several components to allow for the dynamic learning of IME routes and the secure encryption of call signaling and media between organizations. Two elements of the solution are hosted on the internet: the GoDaddy.com Enrollment Servers and the Intercompany Media Engine Bootstrap Servers, hosted by GoDaddy.com and Cisco, respectively. The following additional integral components are deployed on-premises:

- Cisco Intercompany Media Engine Server
- Cisco Unified Communications Manager (Unified CM)
- Cisco Adaptive Security Appliance (ASA)

Figure 5-32 illustrates a high-level view of the deployed components.

Figure 5-32 Cisco Intercompany Media Engine Components



GoDaddy.com Enrollment Server

The GoDaddy.com Enrollment Server validates all IME servers before they enter the ring of IME Servers formed over the Internet. Only IME servers installed and enrolled with the proper GoDaddy.com certificates are allowed to participate in the ring. This enrollment server is accessed only prior to entry into the ring or when certificates expire and an IME server must re-enroll.

Intercompany Media Engine Bootstrap Servers

The IME bootstrap servers are a collection of globally accessible IME servers owned and operated by Cisco. Each IME server participating in the ring (also known as the distributed cache ring) joins the network by first connecting to an IME bootstrap server. The peer-to-peer certificate obtained in the enrollment process is used for all peer-to-peer TLS connections, including the initial connection to the bootstrap server.

Intercompany Media Engine Servers

Each organization owns and operates one or more IME servers on their network. The IME server is responsible for publishing directory numbers owned by the organization to the distributed cache ring, validating call records, learning routes to remote enterprises, and pushing IME learned routes to Unified CM. It is involved only with the IME learning cycle of the solution and does not play a role in the real-time signaling or media communications.

Unified Communications Manager and Session Management Edition

Cisco Unified CM 8.*x* or Unified CM Session Management Edition 8.*x* is required for any organization to participate in IME. Unified CM communicates with IME servers to upload the IME designated directory numbers to the distributed cache ring and sends call records to IME for PSTN calls made by these directory numbers. Unified CM also receives IME learned routes that are validated by IME servers and initiates dynamic SIP trunk calls to the remote directory numbers in these IME learned routes. SIP trunk signaling always flows through an IME-enabled Adaptive Security Appliance (ASA).

Adaptive Security Appliance

All IME calls must flow through an IME-enabled Adaptive Security Appliance (ASA), which provides perimeter security for the solution. The IME-enabled ASA is responsible for receiving SIP signaling communications (outbound from Unified CM or inbound from remote enterprises), validating IME tickets, performing address translation, and providing SIP to SIP+TLS conversion for secure signaling across the Internet. Audio and video media between organizations also flow through the IME-enabled ASA, where it provides RTP-to-Secure RTP (sRTP) conversion and voice quality monitoring of the audio stream incoming from the Internet. There are off path and basic (inline) deployment options. For more information on these deployment options, see ASA Intercompany Media Engine Proxy, page 4-25.

IME Architecture

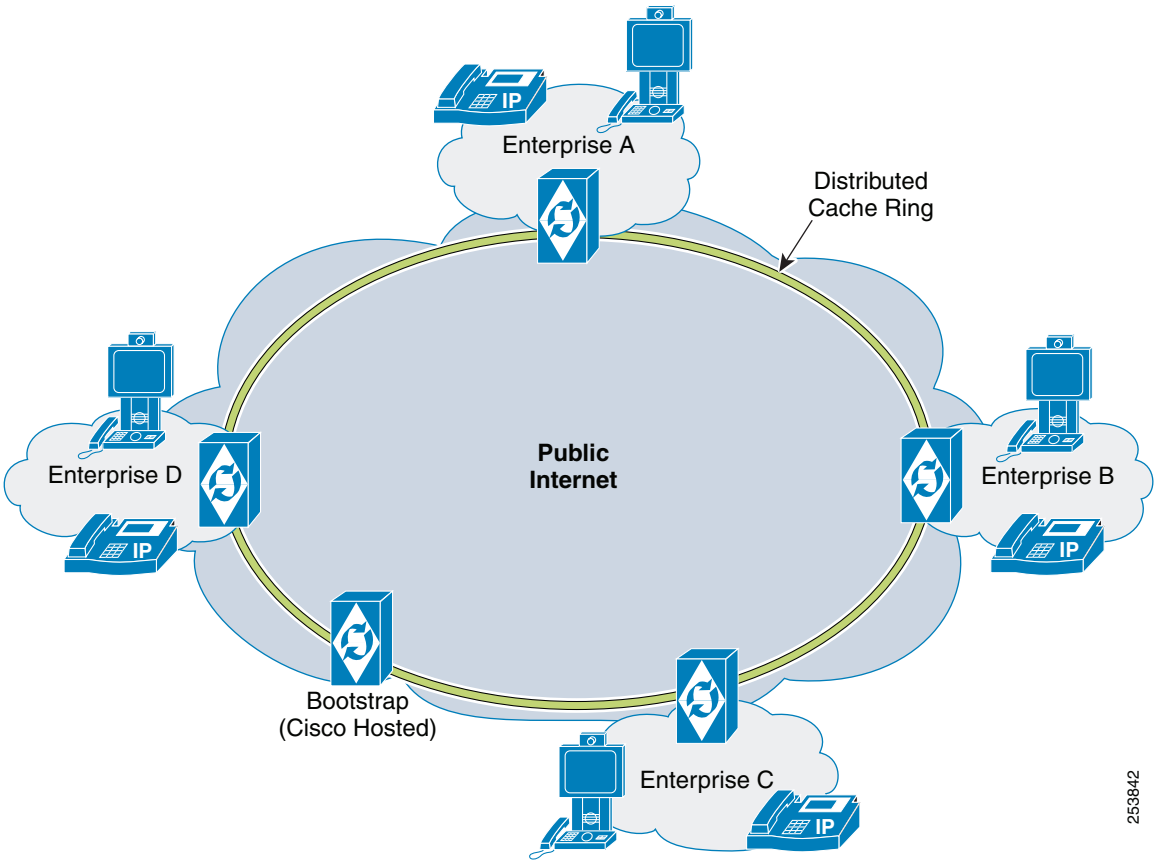
The architecture of IME is reflected in the way IME operates. The operation of IME involves the following high-level phases:

- IME Learned Routes, page 5-77
- IME Call Processing, page 5-80

IME Learned Routes

After enrolling with the GoDaddy.com Enrollment Server and being validated by the IME Bootstrap Server, an IME server becomes an active server on the peer-to-peer ring. IME servers from all organizations participating in IME join the ring on the Internet and communicate using a secure peer-to peer technology based on the Resource Location And Discovery (RELOAD) protocol. The IME servers create a distributed hash table that stores one IME-specific piece of information: a one-way hash of all published +E.164 directory numbers and the IME server peer ID that owns them. This information is distributed across all IME servers, and the architecture of the peer-to-peer technology is such that IME servers can dynamically join or leave the ring without degradation of the ring's functionality. Establishing the IME server on the ring and publishing the enterprise's IME-enrolled directory numbers is the first step toward learning IME routes. Figure 5-33 shows a logical view of the distributed cache ring (DCR).

Figure 5-33 Intercompany Media Engine Distributed Cache Ring

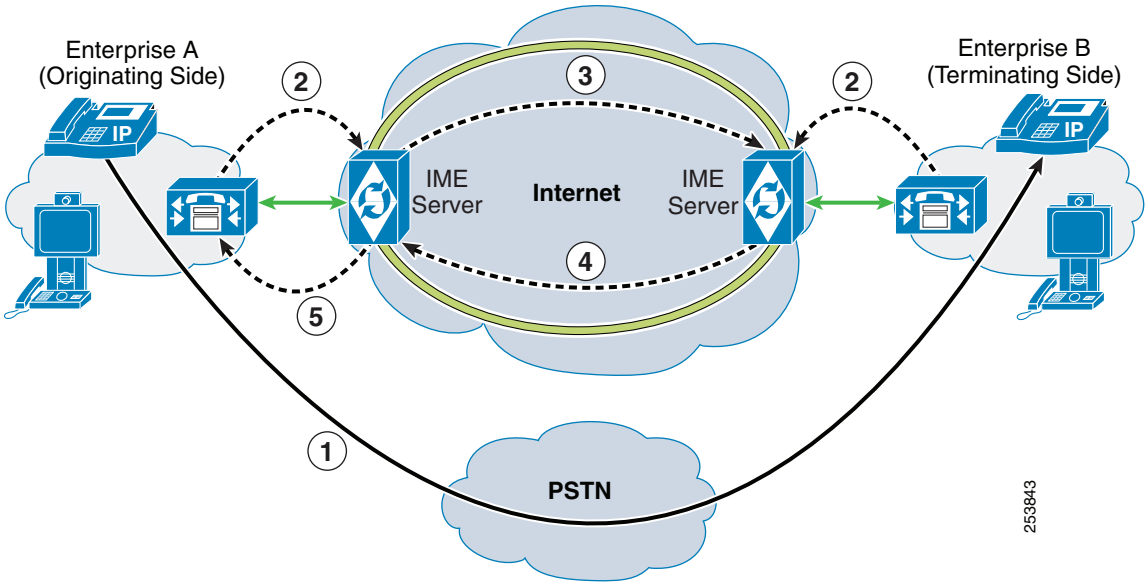


Note A draft has been submitted to the IETF governing body to standardize the IME server peer-to-peer protocol. More information is available at http://datatracker.ietf.org/doc/draft-rosenberg-dispatch-vipr-reload-usage/.

Note IME requires all directory numbers associated with the Intercompany Media Network to be in E.164 format, including the international + prefix (such as +14085551212). This is referred to as +E.164 format throughout this document.

Figure 5-34 illustrates the IME Learned Route process.

Figure 5-34 Intercompany Media Engine Learned Route Process



After the IME solution is deployed in an organization, select directory numbers can be enrolled administratively for IME, and these +E.164 numbers will be published to the distributed cache ring. The first call from an IME directory number uses the PSTN as it did before (Step 1 in Figure 5-34). Because it is an IME directory number, after the call is completed, information about that call in the form of a Voice Call Record (VCR), a CDR-like record specific to IME, is uploaded to the IME server by means of Validation Access Protocol (VAP) (Step 2 in Figure 5-34).

Voice call records contain information such as the called and calling numbers in +E.164 format and the start and stop time of the call. At some point later (it is not real time), the IME server of the enterprise that originated the call will query its peers on the DCR in an attempt to find the enterprise that owns this +E.164 called number (Step 3 in Figure 5-34). When the owner of this called number is discovered (which implies that this directory number has been enrolled in IME by another enterprise), the validation process begins. All communication between IME servers occurs over 128-bit AES TLS. The terminating-side IME server verifies that the called/calling number and start/stop times of the originating IME server's VCR match a corresponding VCR on the terminating side. If verified, the terminating IME server sends a successful reply to the originating IME server that includes a “ticket” (a security hash that only the terminating-side ASA can decipher, as described in IME Call Processing, page 5-80) and the external IP address to which IME SIP trunk calls should be directed for this +E.164 number (Step 4 in Figure 5-34). This constitutes an IME learned route. The originating IME server receives this learned route and at some point later publishes it to Unified CM by means of VAP (Step 5 in Figure 5-34). When Unified CM receives this IME learned route, it inserts the route into the Unified CM database. At this point, when any IME-enabled directory number in the originating enterprise makes a call to a number in the IME learned routes list, it will be an IME call. Note there are no real-time communications involved in learning IME routes. For a detailed example of learned routes, refer to the *Cisco Intercompany Media Engine Installation and Configuration Guide*, available at

http://www.cisco.com/en/US/products/ps10669/prod_maintenance_guides_list.html

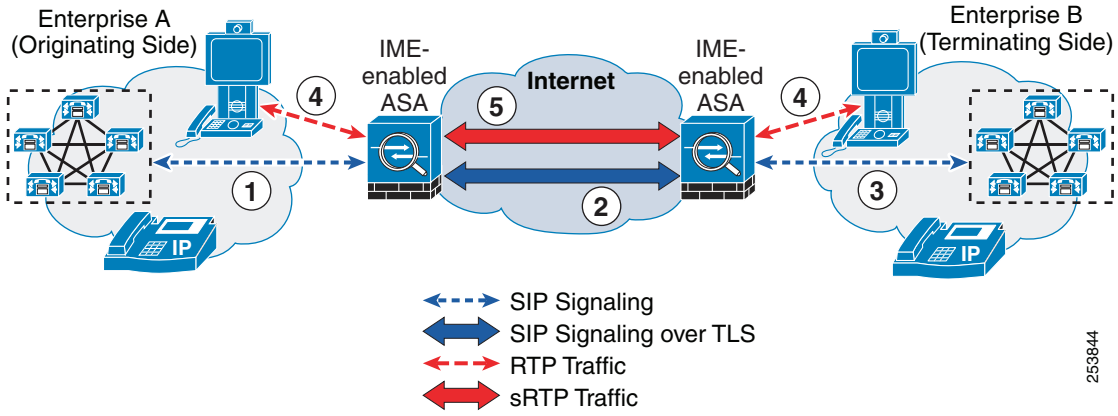
Note Unified CM provides a facility for blocking IME learned routes insertion into the Unified CM database for defined prefixes or domains.

Note Unified CM has a method of transforming called and calling numbers to a globalized +E.164 format specifically for IME VCRs, even if a globalized dial plan is not implemented. For more information on +E.164 transformations for VCRs, see Dial Plan Considerations for the Intercompany Media Engine, page 9-33.

IME Call Processing

Once an IME learned route exists in the Unified CM database, the information in the route is used to set up an IME call. However, the IME server itself is no longer involved in the call processing phase. Figure 5-35 illustrates a high-level view of IME call processing.

Figure 5-35 Intercompany Media Engine Call Processing



Note IME also supports the use of secure SIP signaling via TLS between IME-enabled ASA and Unified CM.

To initiate an IME call, the called number must match an IME learned route pattern in the database and the directory number of the calling endpoint must be enrolled in IME. If these criteria are met, Unified CM dynamically invokes an IME SIP trunk addressed to the external IP address or fully qualified domain name (FQDN) of the terminating enterprise that was included in the IME learned route. IME learned route patterns are in +E.164 format; however, even if a globalized Unified CM dial plan is not deployed, the same process available for converting called numbers to +E.164 format for IME-specific VCRs is used to analyze outgoing called numbers on gateways and to convert them to +E.164 format for IME-specific analysis only. For more information on E.164 transformation profiles, see Dial Plan Considerations for the Intercompany Media Engine, page 9-33.

An IME-enabled ASA serves as a proxy for all IME communications with remote organizations. The ASA provides network address translation (NAT) and SIP application layer gateway (ALG) functionality to translate addressing inside the SIP messaging itself. There are two deployment options for the IME-enabled ASA: basic (inline) or offpath. Offpath is the recommended method because it provides the capability for Unified CM to direct IME traffic to an IME-enabled ASA in a DMZ. This allows use of an existing ASA already deployed in the network that all Unified CM internet-bound traffic would otherwise travel through. For more information about basic and offpath ASA deployments, see ASA Intercompany Media Engine Proxy, page 4-25.

The originating Unified CM initiates a SIP Invite that reaches the IME-enabled ASA (Step 1 in Figure 5-35) and that includes the security hash ticket from the learned route as an attribute in the SIP header. The ASA will fix-up the packets at the SIP level so that its externally facing IP address appears as the source of the Invite and extends it to the external IP address of the remote enterprise over a secure (256 bit AES) TLS connection (Step 2 in Figure 5-35). The external IP address listed in an IME learned route correlates with the outside address of the IME-enabled ASA that receives inbound SIP signaling on behalf of a Unified CM cluster. The terminating ASA receives the SIP Invite, decrypts it, and validates the ticket. Any request without a valid ticket is blocked. After the ticket as been verified, the ASA performs NAT and ALG functions before forwarding the ticket on to the terminating Unified CM (Step 3 in Figure 5-35).

Note The IME Server and IME-enabled ASA do not have direct communications; however, they are both configured with an identical **epoch ticketpassword**, which allows for successful ticket validations.

Once successful SIP signaling has been negotiated, each IME-enabled ASA instructs its respective Unified CMs to have the endpoints stream RTP media directly to its internal media termination address (Step 4 in Figure 5-35). The ASAs take in this RTP stream, encrypt it, perform NAT, and send it across to the remote ASA as sRTP sourced from its external media termination address, including audio and video media (Step 5 in Figure 5-35). At this point, the two endpoints have an active IME call.

The IME solution also provides a mechanism to allow calls to fall-back to the PSTN if the voice quality of the audio stream degrades below an acceptable level. Advanced features such as video are lost, but the audio portion of the call remains intact and the change is otherwise unapparent to the user.

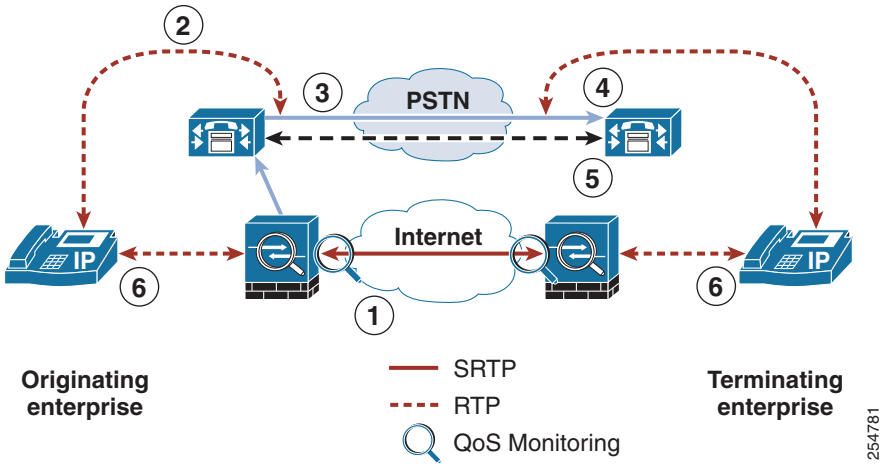
For more details regarding the IME fallback feature and IME-enabled ASA configuration, refer to the information on configuring Cisco Intercompany Media Engine Proxy in the *Cisco ASA 5500 Series Configuration Guide using the CLI, 8.3*, available at

http://www.cisco.com/en/US/docs/security/asa/asa83/configuration/guide/config.html

PSTN Failover

Cisco IME uses the public internet to carry business-to-business traffic. To avoid negative user experience caused by possible packet loss in the public network, the IME-enabled ASA involved in the call continuously monitors the inbound SRTP stream and calculates RTP statistics in real-time. The IME-enabled ASA looks for random packet loss, small bursts of lost packets, and bursts of large packet loss. All three packet impairment conditions are used to decide whether the call meets the defined minimum quality limits. If the measured quality as defined by the calculated real-time RTP statistics falls below certain pre-defined quality limits, the IME-enabled ASA initiates PSTN fallback for the affected call as shown in Figure 5-36. The QoS management algorithm maintains five sensitivity levels with varying packet impairment thresholds to allow granular control of the PSTN fallback sensitivity. The Fallback QoS Sensitivity Level can be set on a global basis or per IME Enrolled Group.

Figure 5-36 IME PSTN Failover



After an IME call is established, the ASA inspects RTP packets as they flow through it from outside to inside. The ASA inspects the sequence numbers and timestamps, and based on the observed packet loss, an algorithm decides whether a fallback is required. The algorithm uses the fallback QoS sensitivity levels to create a set of packet loss thresholds for each QoS sensitivity level. If the algorithm indicates a fallback is needed (step 1 in Figure 5-36), the ASA sends an out-of-dialog REFER to Cisco Unified Communications Manger, asking for it to fallback to the PSTN (step 2 in Figure 5-36).

As the terminating Unified CM receives the REFER, it issues a mid-dialog REFER to the originating Unified CM over the existing dialog. This REFER is required to inform the originating Unified CM of the required fallback. The PSTN call required for the PSTN fallback is always initiated by the Unified CM originating the IME call, regardless of whether the IME-enabled ASA on the originating or terminating side triggers the fallback

The originating Unified CM then places a PSTN call to the Fallback Directory E.164 Number advertised by the terminating Unified CM as part of the SIP call setup (step 3 in Figure 5-36). The Fallback Directory E.164 Number can be configured both in the global Fallback Feature Configuration Settings and in the Fallback Profile Configuration Settings, thus allowing different Fallback E.164 Numbers per IME Enrolled Group.

The call to the Fallback Directory E.164 Number by default is routed using the calling device´s AAR calling search space. The global Fallback Feature Configuration Settings and the Fallback Profile Configuration Settings also allow you to use the calling device´s Reroute Calling Search Space.

When a PSTN call is routed to a configured Fallback Directory E.164 Number, Unified CM has to associate the incoming call with the correct IME call. The first step is to match the caller ID of this PSTN call against the caller ID signalling from the originating Unified CM in the SIP INVITE of the initial VoIP call (step 4 in Figure 5-36). If sufficient digits (as defined by the Number of Digits for Caller ID Partial Match in the global Fallback Feature Configuration Settings or in the Fallback Profile Configuration Settings) are matched, the Unified CM terminating the PSTN fallback call informs the Unified CM originating the PSTN call by sending a single DTMF digit 1. The originating Unified CM immediately splits the VoIP call, connects the PSTN leg with the phone, and terminates the VoIP leg. If the caller IDs do not match, the terminating Unified CM sends a single DTMF digit 2 (step 5 in Figure 5-36).

The originating Unified CM waits for DTMF digits. If a 1 is received, the originating Unified CM immediately splits the VoIP call, connects the PSTN leg with the phone, and terminates the VoIP leg (step 6 in Figure 5-36).

If a 2 is received, indicating that the terminating Unified CM was not able to associate the PSTN fallback call with a unique existing IME VoIP call, the originating Unified CM out-pulses a DTMF sequence uniquely identifying the call. This DTMF sequence is learned from the terminating Unified CM as part of the SIP exchange during the initial IME VoIP call establishment. After sending the DTMF sequence, the originating Unified CM splits the VoIP call, connects the PSTN leg with the phone, and terminates the VoIP leg (step 6, in Figure 5-36).

Unified CM expects to receive the DTMF digits received as part of the PSTN fallback procedure to be delivered out-of-band.

Capacity Planning

IME servers are sized according to how many enrolled DIDs will be published on them. Table 5-5 provides the current supported capacity limits per platform.

Table 5-5 IME Server Supported Capacities

Platform	Maximum Number of Enrolled DIDs
Cisco MCS 7825-H2/I2 and 7825-H4/I4	20,000
Cisco MCS 7845-H2/I2 and 7845-I3	40,000

Because all IME call media (audio and video) flow through the IME-enabled ASA, capacity depends on the type and number of calls flowing through it. The IME-enabled ASA monitors only the audio stream incoming from the internet for voice quality. The video media is not monitored for voice quality, but it does flow through the IME-enabled ASA for RTP-to-sRTP conversion, and the bandwidth of the video directly affects the number of sessions each can handle. Table 5-6 provides capacity limits for the ASA-5550 and ASA-5580. Performance limits of other ASA models have not been validated yet.

Table 5-6 Maximum Number of Calls per Type and ASA Model

ASA Model	Voice G.711	Video 300 kbps	Video 800 kbps	Video 1 Mbps
ASA-5550 4 GB	480 calls	240 calls	120 calls	80 calls
ASA-5580-20 4 GB	900 calls	600 calls	300 calls	200 calls

Unified CM does not have a limit on the number of IME calls it can handle, but IME calls should be factored into the overall call capacity provided by the cluster. Your Cisco Partner or Cisco Systems Engineer should use the Cisco Unified Communications Sizing Tool (http://tools.cisco.com/cucst) to validate all designs that handle large call traffic volumes. The Sizing Tool can accurately determine the number of servers or clusters required to meet your design criteria.

High Availability

The IME route learning phase involves several aspects of high availability. The distributed cache ring (DCR) itself has a high degree of redundancy built into the peer-to-peer technology, where the information stored on the DCR peers is adjusted as IME servers join or leave the ring. Multiple IME bootstrap servers are also hosted by Cisco to guarantee that valid IME servers can join the ring at any time. These aspects are inherent to the solution.

In Unified CM, each IME Service (which defines a set of enrolled DIDs, excluded DIDs, and IME servers, among other parameters) can consist of a primary and secondary IME Server. Both servers are up and active, and Unified CM uploads the enrolled DIDs and any terminating call VCRs to both servers. However, originating call VCRs are uploaded to the primary IME server only; therefore, only the primary will initiate validation requests, while either can process validation requests received from other enterprises because both have terminating call VCRs. There is no direct communication between primary and secondary IME servers regarding VCRs, so originating call VCRs stored for validation on the primary would be lost in the event of an outage. A recommended option is to split the enrolled DIDs into two ranges and to create two IME Services whereby the primary IME server for service A is the secondary IME server for service B, and vice versa. This balances the originating call validation load across the IME Servers and further minimizes the number of originating VCRs that are lost in the event of an outage.

On the Unified CM side, once an IME Service is configured, the Unified CM responsible for initiating VAP communications with the primary IME server is determined by the device pool of the IME SIP trunk associated with the IME Service. The Unified CM Group attribute associated with the device pool determines the primary, secondary, and tertiary Unified CM responsible for the service. In the event that the primary is down, the secondary Unified CM picks up VAP communications with the active IME server.

With respect to call processing, the Unified CM Group associated with the IME SIP trunk in the IME Service also determines which Unified CM subscribers initiate IME calls. This allows for IME calls to continue in the event that the primary Unified CM for the IME SIP trunk is offline. For receiving calls, each IME Service can configure external IP address and port pairs for Unified CM call processing subscribers in the cluster. Each external IP address and port pair is actually an IP address and port that is configured on the IME-enabled ASA and that has a 1:1 correlation to a Unified CM call processing node. When there are multiple external IP addresses and ports in an IME route, Unified CM sends calls for this IME route in a circular fashion so that calls are load-balanced across Unified CM servers at the remote enterprise. If a remote Unified CM is offline, the originating Unified CM tries the next external IP address and port in the list. If no response is received and this list is exhausted, the call is sent to the PSTN as it would have been without IME.

Dual IME-enabled ASAs may also be deployed in active-standby mode; however, this does not provide stateful failover. In the event of a failover, active calls are disconnected, but subsequent IME calls connect through the standby (now active) ASA. For deployments with an offpath IME-enabled ASA, the IME Service configuration in Unified CM allows for a single IME firewall to be associated. Multiple IME-enabled ASAs can be deployed to handle IME calls for different enrolled DID ranges, thus offering a mechanism of load balancing IME calls in addition to increasing capacity.

Note Active-active failover mode is not supported for the IME-enabled ASA.

While an IME call is connected, the IME-enabled ASA is capable of monitoring the quality of the call. If the quality falls below a certain sensitivity level, the call is moved back to the PSTN. For more information, see ASA Intercompany Media Engine Proxy, page 4-25.

Design Considerations

The IME solution requires that IME servers and the IME-enabled ASA have publicly reachable IP addressing; therefore, they are most commonly placed in an organization's DMZ. This may require close coordination between groups responsible for security and Unified Communications within the

organization. It is important for both the security and Unified Communications teams to be involved from the early design stages of an IME project. In addition, observe the following guidelines and considerations when designing an IME solution for your enterprise:

- Cisco requires the use of Network Time Protocol for all Unified CM servers, IME servers, and IME-enabled ASAs. They must be synchronized to a dependable, high-stratum clock source. It is vital to Voice Call Record start and stop times during the IME route learning phase.
- A hosted IME solution deployment model is also supported. In a hosted IME deployment, an IME server publishes enrolled directory numbers and validates VCRs on behalf of multiple Unified CM or Unified CM Session Management Edition clusters. For more information, refer to the information on hosted IME solutions in the *Cisco Intercompany Media Engine Installation and Configuration Guide*, available at

  http://www.cisco.com/en/US/products/ps10669/prod_maintenance_guides_list.html

IME Servers

- By default, VAP communications between Unified CM and the IME server are authenticated only. When an IME server is located in the DMZ, Cisco recommends configuring VAP communications as authenticated and encrypted, which will force the communications to occur over TLS. This requires additional configuration to share security certificates.

Unified CM and Unified CM Session Management Edition

- The Intercompany Media Network requires all published numbers to be in +E.164 format to ensure their global uniqueness. Calling and called numbers must be converted to +E.164 format so that, when IME-specific Voice Call Records (VCRs) are uploaded to IME servers, they are in the proper format. Unified CM provides a facility to transform calling and called numbers to +E.164 format solely for IME purposes, which will not affect normal dial plan digit analysis. For more information, refer to the E.164 transformation profile information in the *Cisco Intercompany Media Engine Installation and Configuration Guide*, available at

  http://www.cisco.com/en/US/products/ps10669/prod_maintenance_guides_list.html
- Gateways or trunks used for PSTN connectivity must have the PSTN Access checkbox checked in order for calling and called numbers to be analyzed for IME participation. Upon upgrade to Unified CM 8.*x*, this parameter is enabled by default for all gateways and trunks. You can unchecked it if it is not required.
- Configuration settings in Unified CM for the regions between internal endpoints and the IME SIP trunk determine the audio and video capabilities allowed for IME calls.
- To limit capacity through the IME-enabled ASAs, Cisco recommends applying Unified CM locations-based call admission control to the IME SIP trunk to control the number of audio and video calls sent through the ASA. When the bandwidth limits are reached, subsequent calls will be routed through the PSTN as they were prior to IME deployment.
- Cisco recommends explicitly trusting the domains of remote enterprises with which your organization intends to communicate through IME. Once a trust group is configured, there is a default deny on all other domains that try to validate VCRs.
- Unicast music on hold (MoH) is supported during user-initiated hold and transfer scenarios. To work properly through the firewalls, the MoH full-duplex streaming service parameter must be enabled.
- Cisco recommends excluding analog and fax station directory numbers from the enrolled group of DIDs for an IME server because they will not benefit from enhanced Unified Communications and because fax calls are not supported over IME.

IME-Enabled ASA

- For more information on basic and offpath ASA deployments as well as security considerations for other firewalls in the network, see ASA Intercompany Media Engine Proxy, page 4-25.
- High-bandwidth video (greater than 384 kbps) is supported; however, it directly affects the capacity of calls flowing through the IME-enabled ASA.
- Fallback sensitivity levels should be left at the default settings for the initial IME deployment. Fallback should be monitored during the first few months of use and then adjusted accordingly. Cisco recommends viewing call detail records to find calls generated on behalf of IME or fallback. Appropriate fallback sensitivity levels will vary from enterprise to enterprise.
- When endpoints with IME-enrolled DIDs are remotely located with VPN connectivity into the enterprise, latency and jitter characteristics for calls with these endpoints will be amplified and could result in the IME-enabled ASAs triggering more frequent fallbacks to the PSTN. If fallbacks occur too frequently for a specific endpoint, it might be necessary either to configure these devices with a device pool that has a fallback profile with no fallback enabled, to lower the fallback sensitivity levels, or to remove the enrolled DID from IME.