Collaboration Deployment Models

Revised: November 19, 2013; OL-30952-01

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

Earlier versions of this chapter based the deployment models discussion exclusively on the call processing deployment models for Cisco Unified Communications Manager (Unified CM). The current version of this chapter offers design guidance for the entire Cisco Unified Communications and Collaboration 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 10-1 lists the topics that are new in this chapter or that have changed significantly from previous releases of this document.

Table 10-1 New or Changed Information Since the Previous Release of This Document

New or Revised Topic	Described in	Revision Date
Collaboration edge	Deployments for the Collaboration Edge, page 10-33, and various other sections of this chapter	November 19, 2013
Intercluster Lookup Service (ILS) and Global Dial Plan Replication (GDPR)	Intercluster Lookup Service (ILS) and Global Dial Plan Replication (GDPR), page 10-29, and various other sections of this chapter	November 19, 2013
Virtualized deployment	Deploying Unified Communications on Virtualized Servers, page 10-55	November 19, 2013
Recommendation to register SIP-based VCS endpoints on Unified CM	Various sections of this chapter	November 19, 2013
Other updates for Cisco Collaboration Systems Release 10.0	Various sections of this chapter	November 19, 2013

Deploying Unified Communications and Collaboration

From its beginnings with Voice over IP (VoIP) and IP Telephony 10 to 15 years ago, and continuing today with Unified Communications and Collaboration, users expect to be able to meet and communicate in a variety ways using a range of devices with differing capabilities. Today, a Unified Communications and Collaboration system could start with the deployment of Jabber Clients for IM and Presence only and incrementally add voice, video, web conferencing, mobile voice applications, social media, video conferencing, and telepresence as required. A tightly integrated Unified Communications architecture is required as the number of devices and forms of communication available to a single Unified Communications user increases. Cisco's Unified Communications and Collaboration architecture has the flexibility and scale to meet the demands of a rapidly changing and expanding Unified Communications environment that will become more URI-centric as users with multiple Unified Communications devices wish to be identified by a single user name irrespective of the form of communication.

Collaboration as a Service

As Unified Communications has become more commonplace, the deployment options for Cisco Unified Communications and Collaboration have increased to address the demand for "collaboration as a service" as well as traditional on-premises Unified Communications and Collaboration deployments. The main focus of this chapter is to provide the reader with design guidance for on-premises Unified Communications and Collaboration deployments, but it also includes a description of systems such as Cisco Hosted Collaboration Solution (HCS) and Cisco WebEx, which can be deployed as managed cloud-based Unified Communications and Collaboration services. The choice of using an on-premises, cloud-based, or hybrid solution may be determined by many factors. For example, cloud-based solutions require less on-site expertise but might lack the deployment flexibility that many enterprises need. Hybrid designs can also be deployed, where some Unified Communications functions such as call control are provided on-premises and others are provided as a cloud-based service.

Cisco offers the following "collaboration as a service" products that can be used to augment or replace functionality provided in an on-premises Unified Communications deployment:

- Cisco WebEx, offering the following cloud-based services:
  - Cisco WebEx Meetings
  - Cisco WebEx Telepresence
- Cisco WebEx Messenger

WebEx Meetings, WebEx Telepresence and WebEx Messenger can be deployed as WebEx cloud services or as on-premises services using Cisco Unified CM, Cisco WebEx Meetings Server, Cisco Expressway, and Cisco IM and Presence.

For more information on Cisco Webex, refer to the documentation at

http://www.cisco.com/en/US/products/ps10352/index.html

Cisco Hosted Collaboration Solution provides the following range of applications and services:

- Cisco Unified Communications Manager (call control for voice and video)
- Cisco Unity Connection (voicemail)
- Cisco Jabber (instant messaging, presence, and soft endpoint)
- Cisco Unified Mobility for Nokia, iPhone, and Google Android clients
- Cisco Unified Attendant Console Enterprise Edition
- Cisco Contact Center Enterprise (customer collaboration)

- Cisco WebEx (web conferencing)
- Cisco TelePresence (point-to-point and multipoint videoconferencing)

For more information on Cisco Hosted Collaboration Solution, refer to the documentation available at

http://www.cisco.com/en/US/products/ps11363/index.html

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 9-28.

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 with up to eight call processing subscriber nodes.

Summary of Unified Communications Deployment Models

This chapter discusses three basic on-premises deployment models for Unified Communications and Collaboration:

- Campus deployment model

  Where the Unified Communications and Collaboration services, their associated endpoints, gateways, border controllers, media resources, and other components are all located on a single high speed LAN or MAN.

- Centralized deployment model

  Where the Unified Communications and Collaboration services are located in a central campus site or data center, but the endpoints, gateways, media resources, and other components are distributed across multiple remote sites that are interconnected by a QoS-enabled WAN.

- Distributed deployment model

  Where multiple campus and/or centralized deployments are interconnected by means of a trunk and dial plan aggregation platform, such as a Cisco Unified Communications Manager Session Management Edition cluster, over a QoS-enabled WAN.

There are an infinite number of variations on these three basic deployment models, such as deployments with centralized or distributed PSTN access and services, but the basic design guidance provided in this chapter still applies to the majority of them.

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.

Common Design Criteria

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

  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 Survivable Remote Site Telephony (SRST) or Enhanced SRST. Likewise, a centralized voice messaging service such as that of Cisco Unity Connection can be provisioned to allow remote sites operating under SRST to access local voicemail services using Unity Connection Survivable Remote Site Voicemail (SRSV).

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 SIP or H.323 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. Depending on the design and features in use, this can provide the possibility for continued service during site disruptions such as loss of power, network outages, or even compromises in the physical integrity of a site by catastrophic events such as a 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 10-2 summarizes the ability of each Cisco Unified Communications service to be deployed in the manners outlined above.

Table 10-2 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 Unity Express	No	Yes	Yes, through Voice Profile for Internet Mail (VPIM) networking	No
Cisco Unity Connection	Yes	Yes (One Cisco Unity Connection per site)	Yes, through VPIM networking	Yes
Cisco Emergency Responder	Yes	Yes (One Emergency Responder group per site)	Yes, through Emergency Responder clustering	Yes
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
Cisco Expressway	Yes	Yes	Yes	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 9-1.

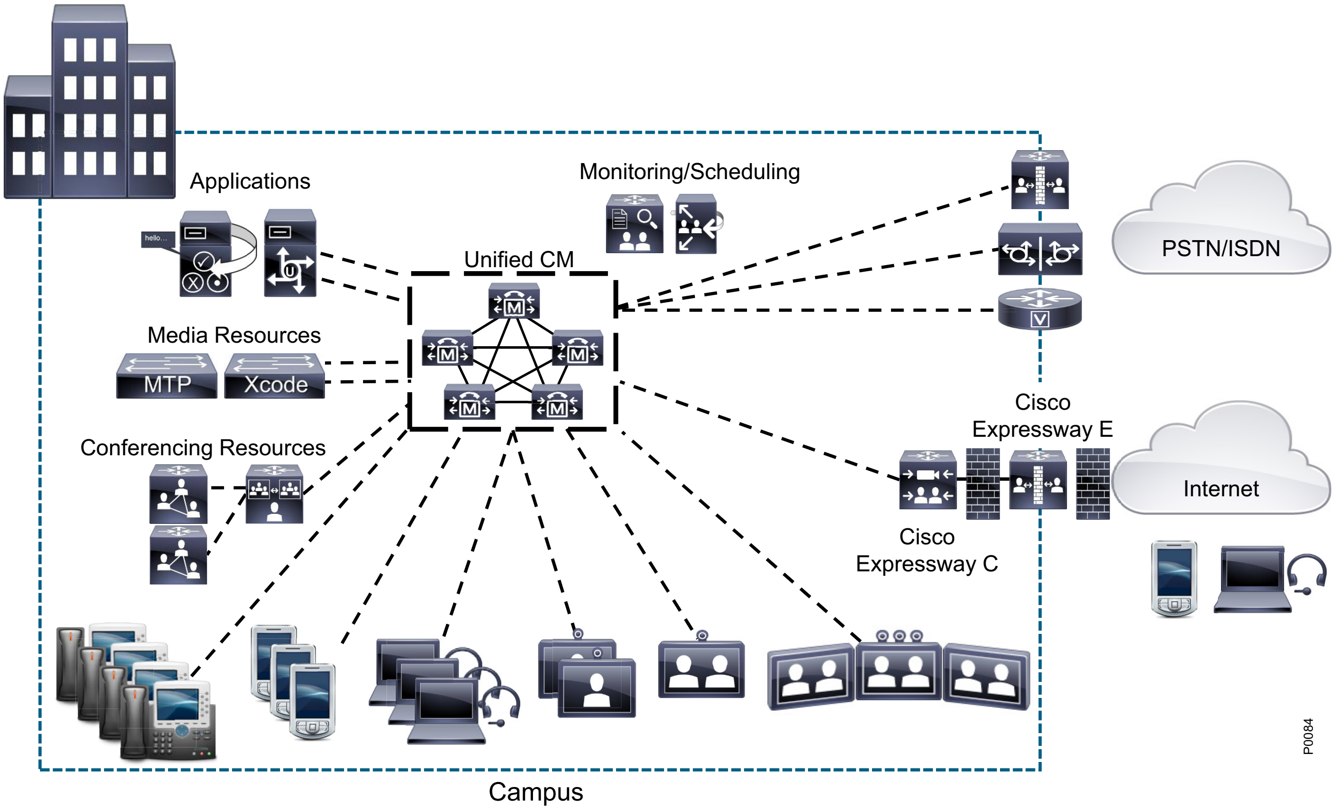
Design Characteristics and Best Practices for Deployment Models

This section describes the fundamental deployment models for Cisco Collaboration and Unified Communications systems, and it lists best practices for each model.

Campus Deployments

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 bandwidth in the gigabit range 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. (See Figure 10-1.)

Figure 10-1 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 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, softphones, analog ports, video endpoints, SIP-based TelePresence endpoints and room-based TelePresence conferencing systems, 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.
- Multipoint control unit (MCU) resources are required for multipoint conferencing.
- 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.

- SIP-based video ISDN gateways are needed to communicate with videoconferencing devices on the public ISDN network.
- Cisco Expressway C and Cisco Expressway E provide a collaboration edge function that enables secure business-to-business telepresence and video communications, and enterprise access for remote and mobile workers over the internet.
- Cisco TelePresence Video Communication Server (VCS) may also be used to register legacy H.323 and third-party telepresence endpoints. However, to avoid the dial plan and call admission control complexities that dual call control introduces (see Design Considerations for Dual Call Control Deployments, page 10-38), Cisco recommends using SIP to register all TelePresence endpoints and room-based TelePresence conferencing systems with Cisco Unified Communications Manager
- High-bandwidth audio is available (for example, G.711 or G.722) between devices within the site.
- High-bandwidth video (for example, 1.5 Mbps with 4CIF or 720p, to 2 Mbps with 1080p) 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 9-1.

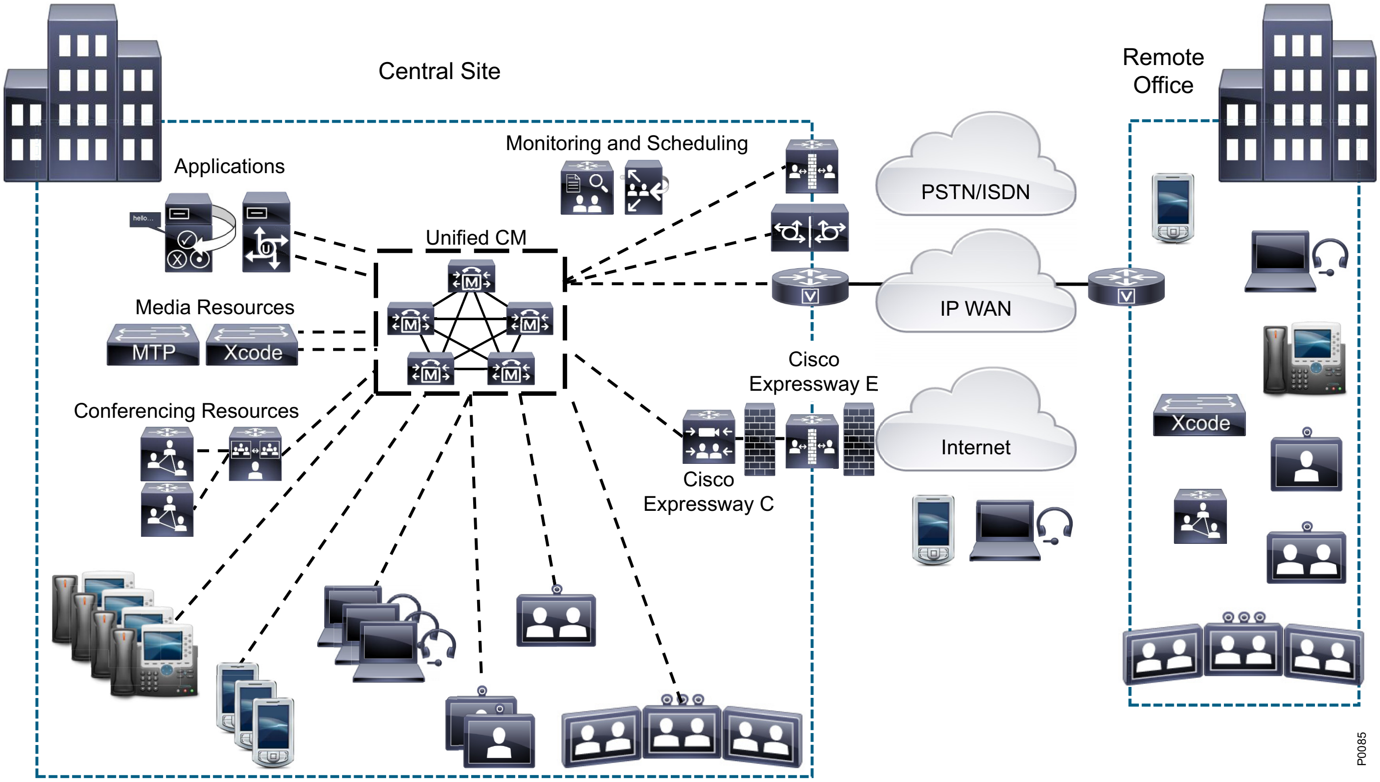
Multisite Deployments with Centralized Call Processing

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 10-2 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 10-2 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.
- 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, softphones, analog ports, video endpoints, SIP-based TelePresence endpoints and room-based TelePresence conferencing systems, 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.
- Multipoint control unit (MCU) resources are required for multipoint conferencing. These resources may all be located at the central site or may be distributed to the remote sites if local conferencing resources are required.
- Capability to integrate with legacy private branch exchange (PBX) and voicemail systems. Connections to legacy voice services such as PBXs and voicemail systems can be made 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.
- SIP-based video ISDN gateways are needed to communicate with videoconferencing devices on the public ISDN network. ISDN video gateways can be centralized and/or deployed at each remote site.
- Cisco Expressway C and Cisco Expressway E provide a collaboration edge function that enables secure business-to-business telepresence and video communications, and VPN-less enterprise access for remote and mobile workers over the internet.
- Cisco TelePresence Video Communication Server (VCS) may also be used to register legacy H.323 and third-party telepresence endpoints. However, to avoid the dial plan and call admission control complexities that dual call control introduces (see Design Considerations for Dual Call Control Deployments, page 10-38), Cisco recommends using SIP to register all TelePresence endpoints and room-based TelePresence conferencing systems with Cisco Unified Communications Manager
- 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, 1.5 Mbps with 4CIF or 720p, to 2 Mbps with 1080p) between devices in the same site, and low-bandwidth video (for example, 384 kbps with 448p or CIF) between devices at different sites.
- A minimum of 1.5 Mbps 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 between endpoints within a cluster 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. Video endpoints located at remote sites become audio-only devices if the WAN connection fails. Starting with Cisco IOS release 15.3(3)M, using phone load firmware 9.4.1 or later, Enhanced SRST enables video survivability on SIP video endpoints (Cisco Unified IP Phone 9900, for example) during WAN failure. For SRST video support with a particular phone model, refer to the respective Cisco Unified IP Phone Administration Guide available at http://www.cisco.com.

- Cisco Unified Communications Manager Express (Unified CME) may be used for remote site survivability (Enhanced SRST) 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 Enhanced 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.
- With multisite centralized call processing model, PSTN routing through both central and remote site gateways is supported. Providing a local gateway at a remote site for local PSTN breakout might be a requirement for countries that provide emergency services for users located at remote sites. In this case, the local gateway at the remote site provides call routing to the local PSAP for emergency calls. Local PSTN breakout at remote sites might also be required for countries having strict regulations that require the separation of the IP telephony network 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).

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 and video 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 and/or video traffic and deteriorating the quality of established calls. For centralized call processing deployments, Enhanced Location CAC or RSVP-enabled locations configured within Unified CM provide call admission control (CAC). (Refer to the chapter on Call Admission Control, page 13-1, 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 is available on Cisco IOS gateways or on Cisco Unified CME running Enhanced SRST. Unified CME running Enhanced SRST 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 13-1, 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 Enhanced SRST 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 or Enhanced SRST.
  - For SIP phones, use SIP SRST or Enhanced SRST.
  - For deployments with centralized voicemail, use Survivable Remote Site Voicemail (SRSV).

  SRST, Enhanced SRST, SIP SRST, SRSV, 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 10-3 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 10-3 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 Enhanced SRST	No	Yes

The first two solutions listed in Table 10-3 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 10-3 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 Enhanced SRST 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.

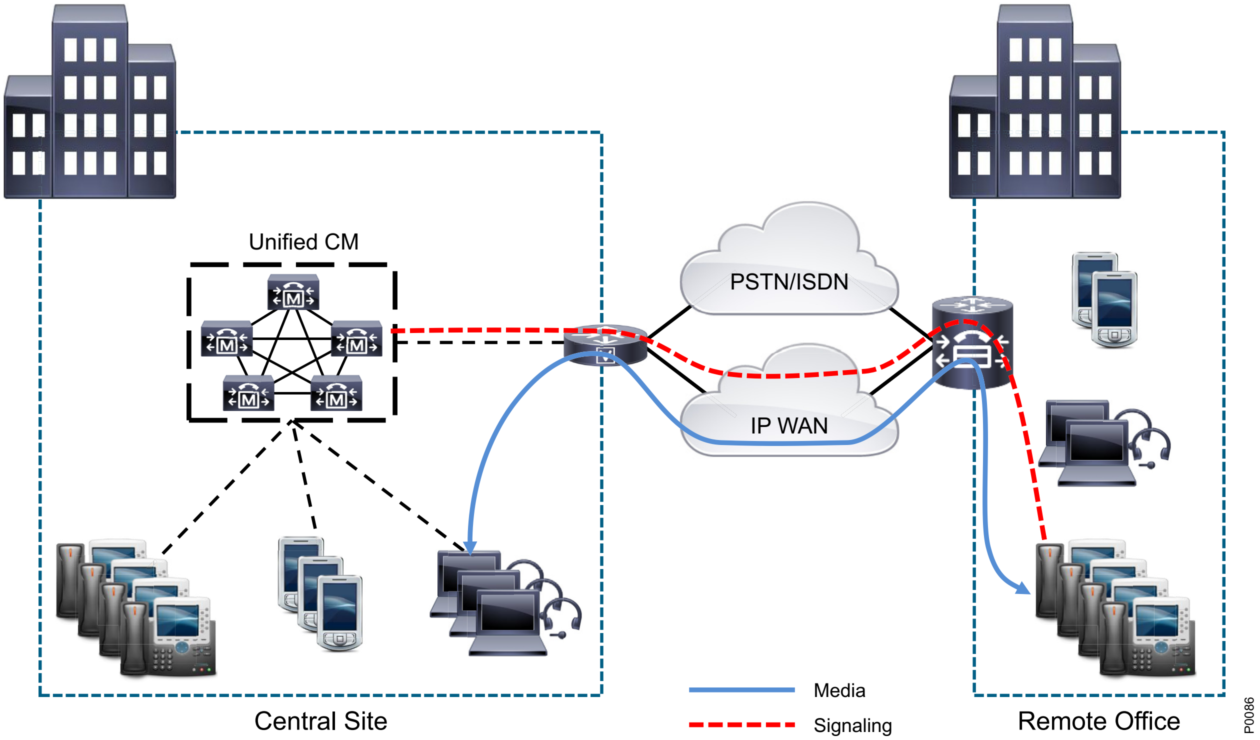
- Data and voice ISDN backup

  With this option, ISDN is used for both data and voice survivability. In this case, SRST or Enhanced SRST 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 10-3, Survivable Remote Site Telephony (SRST) or Enhanced SRST, 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 10-3 illustrates a typical call scenario with SRST or Enhanced SRST.

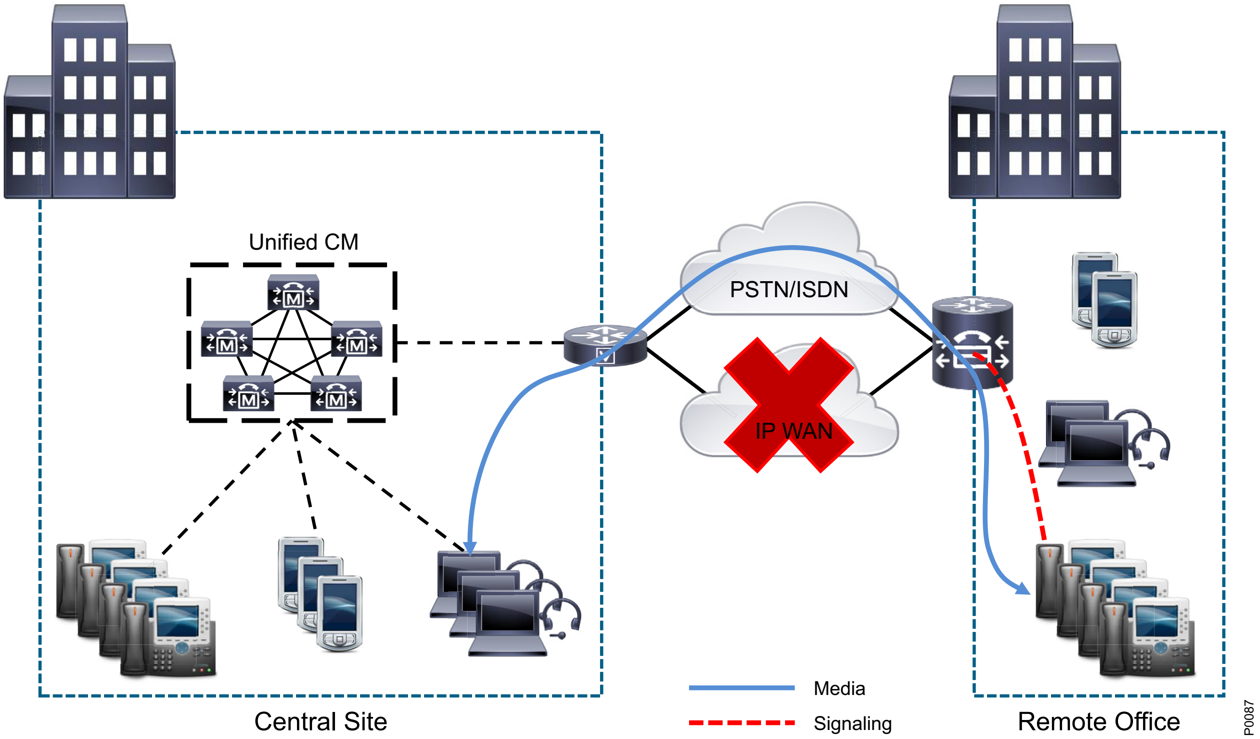
Figure 10-3 Survivable Remote Site Telephony (SRST) or Enhanced SRST, Normal Operation



Under normal operations shown in Figure 10-3, the remote office connects to the central site via an IP WAN, which carries data traffic, voice traffic, and call signaling. The IP phones at the remote office exchange call signaling information with the Unified CM cluster at the central site and place their calls across the IP WAN. The remote office 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 remote office fails, as shown in Figure 10-4, or if some other event causes loss of connectivity to the Unified CM cluster, the remote office IP phones re-register with the remote office router in SRST mode. The remote office router, using SRST or Enhanced SRST, queries the IP phones for their configuration and uses this information to build its own configuration automatically. The remote office IP phones can then make and receive calls either within the remote office 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.

Figure 10-4 Survivable Remote Site Telephony (SRST) or Enhanced SRST, WAN Failure



When WAN connectivity to the central site is reestablished, the remote office IP phones automatically re-register with the Unified CM cluster and resume normal operation. The remote office SRST router deletes its information about the IP phones and reverts to its standard routing or gateway configuration. Routers using Enhanced SRST at the remote office 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.

Enhanced SRST

Enhanced SRST 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, Enhanced SRST also provides most of the Unified CME telephony features for 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 Jabber Clients
- Cisco Unified Video Advantage
- Endpoint video calls
- Integration with Cisco Unity with MWI support at remote sites, with distributed Microsoft Exchange or IBM Lotus Domino server

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

Best Practices for Enhanced SRST

- 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 Enhanced SRST, 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 MGCP Gateway fallback, refer to the information on MGCP gateway fallback in the *Cisco Unified Communications Manager and Interoperability Configuration Guide, Cisco IOS Release 15M&T*, available at

http://www.cisco.com/en/US/docs/ios-xml/ios/voice/cminterop/configuration/15-mt/cminterop-15-mt-book.html

Best Practices for SRST Router

Use a Cisco Unified SRST router, rather than Enhanced SRST, for the following deployment scenarios:

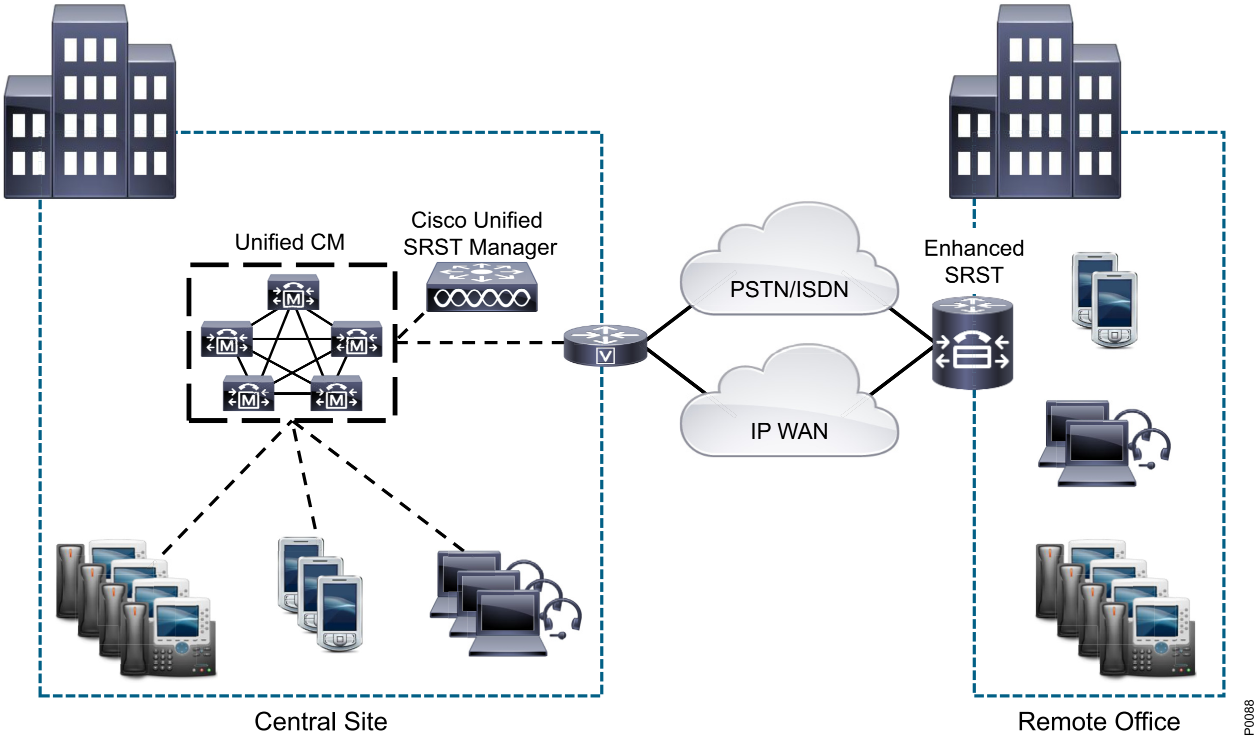
- For supporting a maximum of 1,500 phones on a single SRST router. (Enhanced SRST supports a maximum of 450 phones.)
- For up to 3,000 phones, use two SRST routers. Dial plans must be properly configured to route the calls back and forth between the SRST routers.
- For simple, one-time configuration 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 Enhanced SRST as well as traditional SRST in the branch. (See Figure 10-5.) Cisco Unified SRST Manager 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 10-5 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 10-5 Cisco Unified SRST 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 remote office 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.
- The remote office 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.

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.

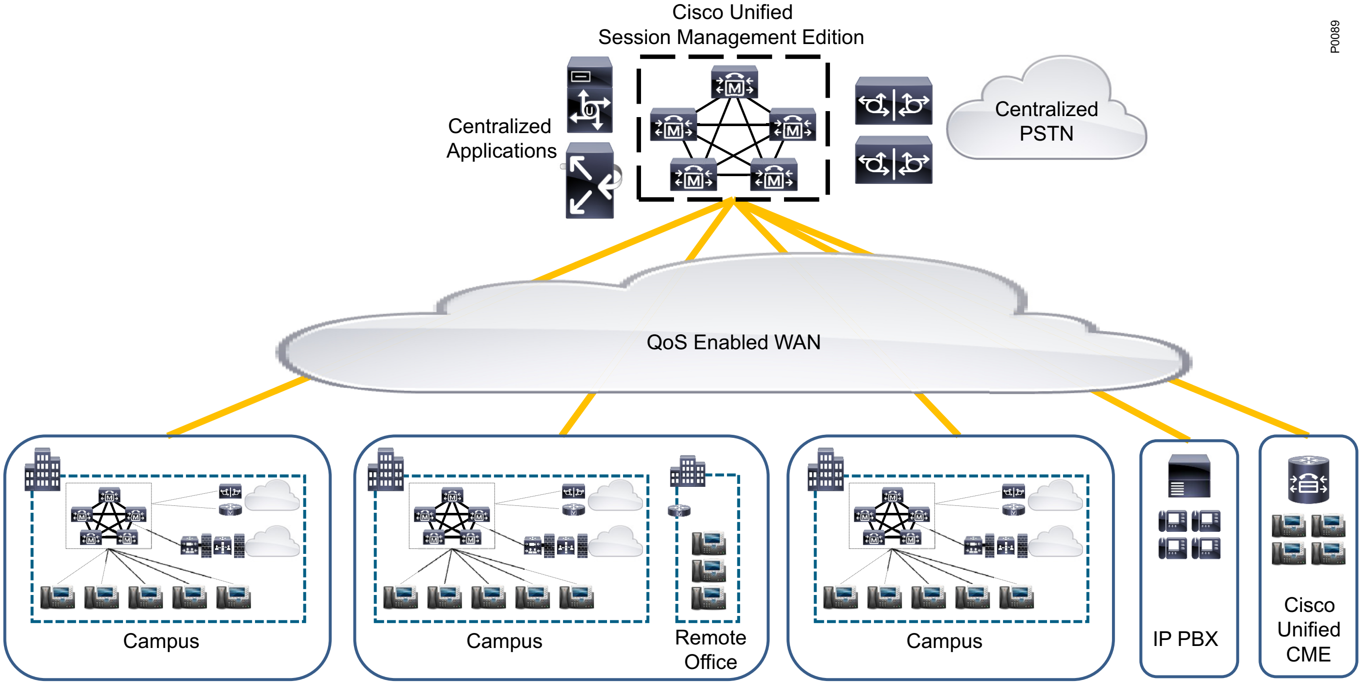
For more information on VoPSTN deployment options and design guidance, refer to the VoPSTN sections in the *Unified Communications Deployment Models* chapter of the *Cisco Unified Communications System 9.0 SRND*, available at

http://www.cisco.com/en/US/docs/voice_ip_comm/cucm/srnd/9x/models.html

Multisite Deployments 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 10-6 illustrates a typical distributed call processing deployment.

Figure 10-6 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 6000
  - Cisco Unified Communications Manager Express (Unified CME)
  - A third-party IP PBX
  - A legacy PBX with Voice over IP (VoIP) gateway
- A centralized call processing site and all of its associated remote sites

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

- A centralized platform for trunk and dial plan aggregation is commonly deployed. This platform is typically a Cisco Unified Communications Manager Session Management Edition (SME) cluster, although a Session Initiation Protocol (SIP) Proxy Server could also be used to provide intercluster call routing and dial plan aggregation in multisite distributed call processing deployments.
- Centralized services such as:
  - Centralized PSTN access
  - Centralized voicemail
  - Centralized conferencing

  These services can be deployed centrally, thus benefiting from centralized management and economies of scale. Services that need to track end-user status (for example, Cisco IM and Presence) must connect to the Unified CM cluster for the users that they serve.
- 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, 1.5 Mbps with 4CIF or 720p, to 2 Mbps with 1080p) between devices in the same site, but low-bandwidth video (for example, 384 kbps with 448p or CIF) between devices at different sites.
- Minimum of 1.5 Mbps or greater WAN link speeds. Video is *not* recommended on WAN connections that operate at speeds lower than 1.5 Mbps.
- 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 Deployments, page 10-8.)

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 Deployments, page 10-8, and Multisite Deployments with Centralized Call Processing, page 10-10.)

Dial Plan Aggregation Platforms for Distributed Call Processing Deployments

A Cisco Unified Communications Manager Session Management Edition (SME) cluster 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 trunk and 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 or URI based using the Intercluster Lookup Service (ILS) and Global Dial Plan Replication (GDPR) (see Global Dial Plan Replication, page 14-46). Unified CM Session Management Edition uses exactly the same code and user interface as Unified CM but leverages support for multiple trunk protocols (SIP, H.323, and MGCP) as well as sophisticated trunk, digit manipulation, and call admission control features. Unified CM Session Management Edition cluster deployments typically consist of many trunks (SIP trunks are recommended; see Cisco Unified CM Trunks, page 6-1) and no Unified Communications endpoints. Unified CM Session Management Edition clusters can use all of the high availability features (such as clustering over the WAN, and Run on all Unified CM Nodes) that are available to Unified CM clusters.

SIP Proxy Deployments

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.

Note Because Session Management Edition (SME) uses exactly the same code and GUI as Unified CM and can also share intercluster features such as ILS, GDPR, and Enhanced Locations Call Admission Control (ELCAC), SME is the recommended trunk and dial plan aggregation platform in multi-site distributed call processing deployments.

Leaf Unified Communications Systems 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 may have 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 10-4 lists recommended call processing agents.

Table 10-4 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 • SIP trunks are recommended
Cisco Business Edition 6000	Up to 2,500 phones	• For small to medium sites • Supports centralized call processing • Supports distributed call processing • SIP trunks are recommended
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 call processing • Supports distributed call processing • SIP trunks are recommended
IP PBX	Depends on the PBX	• IP PBXs commonly use SIP trunks, which can be used to connect to SME
Legacy PBX with VoIP gateway	Depends on the PBX	• Number of IP WAN calls and functionality depend on the PBX-to-VoIP gateway protocol and the gateway platform • SIP trunks are recommended between the VoIP gateway and SME

Unified CM Session Management Edition

Cisco Unified CM Session Management Edition (SME) is the recommended trunk and dial plan aggregation platform in multi-site distributed call processing deployments. SME 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.

Cisco Unified CM Session Management Edition supports the following trunk protocols:

- H.323 Annex M1 intercluster trunks
- SIP intercluster trunks
- SIP trunks
- H.323 trunks to gateways
- MGCP trunks to gateways

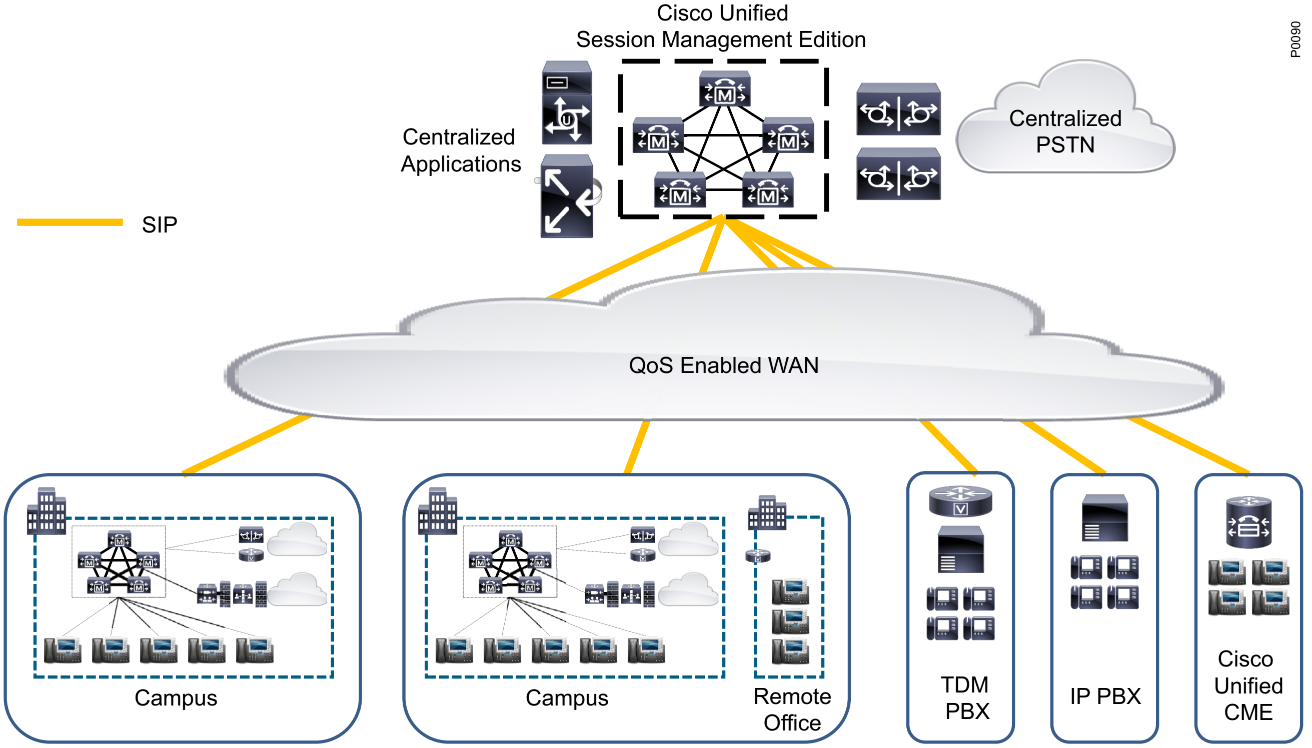
SIP trunks are recommended for SME and leaf Unified Communications systems because SIP offers additional features and functionality over H.323 and MGCP trunks. (For more information, see the chapter on Cisco Unified CM Trunks, page 6-1.)

Cisco Unified CM Session Management Edition supports the following call types:

- Voice calls
- Video calls
- Encrypted calls
- Fax calls

Unified CM Session Management Edition may also be used to connect to the PSTN and third-party unified communications systems such as PBXs and centralized unified communications applications. (See Figure 10-7.) 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 10-7 Multisite Distributed Call Processing Deployment with Unified CM Session Management Edition



When to Deploy Unified CM Session Management Edition

Cisco recommends deploying Unified CM Session Management Edition (SME) 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.

  Note Running Intercluster Lookup Service (ILS) and Global Dial Plan Replication (GDPR) on SME and Unified CM leaf clusters further simplifies dial plan administration because individual directory numbers, E.164 numbers corresponding to DNs, route patterns (for internal and external number ranges), and URIs can be distributed using the ILS service. This approach simplifies dial plan administration by reducing the required number of route patterns to one SIP route pattern per call control system (Unified CM cluster, for example), instead of a route pattern for each unique number range. For more information on ILS and GDPR, see Intercluster Lookup Service (ILS) and Global Dial Plan Replication (GDPR), page 10-29.

- 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 voicemail 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. If ILS GDPR is deployed, the number ranges and/or URIs supported by each third-party system can also be imported into ILS GDPR and reached through a SIP route pattern and corresponding SIP trunk.

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 to satisfy the requirements of this 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, between 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. For more information on SME sizing, see the chapter on Collaboration Solutions Design and Deployment Sizing Considerations, page 27-1.

Trunks

Although SME supports SIP, H.323, and MGCP trunks, Cisco recommends using SIP as the trunk protocol for SME and Unified CM leaf clusters running Cisco Unified CM 8.5 and later releases.

SIP trunks provide a number of unique features that greatly simplify trunk designs and Unified Communications deployments, such as:

- Run on All Unified CM Nodes
- OPTIONS Ping
- Accept Codec Preference in Received Offer
- Lua scripts, which allow SIP Message and Session Description Protocol (SDP) content modification for interoperability

Using only SIP trunks in the SME cluster allows you to deploy a "media transparent" cluster where media resources, when required, are inserted by the end or leaf Unified Communications system and never by SME. Using only SIP trunks also allows you to use extended round trip times (RTTs) between SME nodes when clustering over the WAN.

SME SIP trunks should be configured as **MTP-less Early Offer** trunks. Leaf Unified CM clusters can be configured as **SIP Delayed Offer** or **SIP Early Offer for Voice and Video (insert MTP if needed)**. For more details, see the chapter on Cisco Unified CM Trunks, page 6-1.

Media Resources

When a media resource such as an MTP or transcoder is needed to allow a call to proceed successfully, these resources should ideally be allocated by the leaf Unified Communications systems. If SME trunk media resources are deployed and used for a call traversing the SME cluster, the media path call will hairpin through the SME media resource. By using SIP trunks only and **MTP-less Early Offer**, you can deploy an SME cluster without media resources. If or when media resources are required, they can be allocated by the leaf Unified Communications system.

Clustering over the WAN

SME deployments support round-trip times (RTTs) of up to 500 ms between SME cluster nodes. (See Figure 10-8.) This extended RTT applies only to SME clusters (80 ms is the maximum RTT for a standard Unified CM cluster designs) and is subject to the following design restrictions:

- Extended round-trip times for SME deployments with clustering over the WAN are supported with SIP trunks only. All SIP trunks must be configured as **MTP-less Early Offer** and must use the **Run on all Unified CM Nodes** feature so that calls are not routed between nodes within the SME cluster. (For more information, see the chapter on Cisco Unified CM Trunks, page 6-1.) MGCP, SCCP, and H.323 protocols do not support extended round-trip times for SME deployments with clustering over the WAN.
- No endpoints or CTI devices are configured or registered to the SME cluster.
- No media resources such as MTPs, Trusted Relay Points (TRPs), RSVP agents, or transcoders are configured or registered to the SME cluster. (To disable media resources hosted on Unified CM nodes, deactivate the IPVMS service on each node within the cluster.)
- A minimum of 1.544 Mbps (T1) bandwidth is required for Intra-Cluster Communication Signaling (ICCS) traffic between sites.

- 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 between the publisher and every remote subscriber node.

  Like all other SME designs, your SME design must be reviewed and approved by the Cisco SME team prior to deployment.

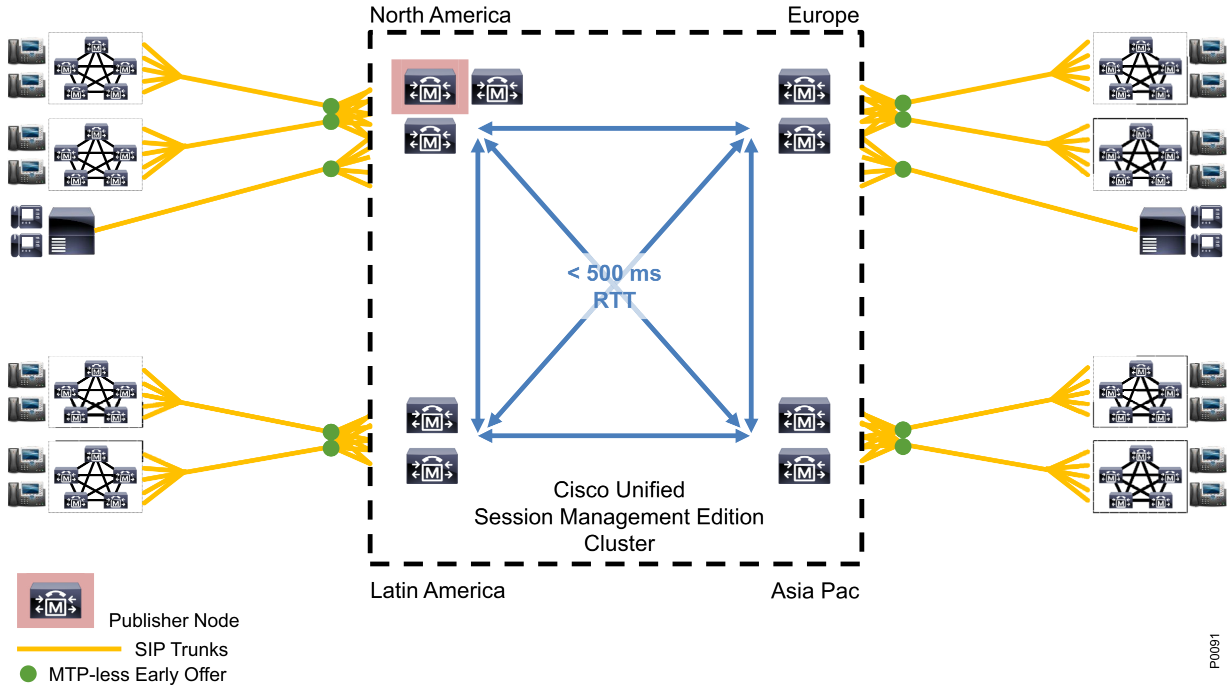
Note The upgrade process for an SME cluster consists of two key parts: Version switch-over, where the call processing node is re-booted and initialized with the new software version (this takes approximately 45 minutes per server), and database replication, where the subscriber's database is synchronized with that of the publisher node. The time taken to complete this database replication phase depends on the RTT between the publisher and subscriber nodes and the number of subscribers in the cluster. The database replication process has a minimal impact of the subscriber's call processing capability and typically can be run as a background process during normal SME cluster operation. Avoid making changes to the SME cluster configuration during the database replication phase because this increases the time it takes to complete the replication.

For SME clusters deployed with extended RTTs, before upgrading the cluster, run the following Admin level CLI command on the publisher node:

```
utils dbreplication setprocess 40
```

This command improves replication setup performance and reduces database replication times.

Figure 10-8 Unified CM Session Management Edition Clustering over the WAN with Extended Round Trip Times



Unified CM Versions

Using the latest Cisco Unified Communications System release and SIP trunks across all Unified CM leaf clusters and the SME cluster allows your Unified Communications deployment to benefit from common cross-cluster features such as Codec Preference Lists, Intercluster Lookup Service (ILS), Global Dial Plan Replication (GDPR), and Enhanced Locations Call Admission Control (ELCAC). If you do not wish to upgrade to the latest Unified Communications version on all clusters, the lowest recommended version is Cisco Unified CM 8.5 using SIP trunks, because this version includes features that improve and simplify call routing through Unified CM and Session Management Edition clusters.

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

For SIP trunk interoperability issues, Lua scripting can be used to modify inbound and outbound SIP messages and SDP content.

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. As a general rule, always enable the **Run on All Unified CM Nodes** feature if it is available. For more information on load balancing for trunk calls, refer to the chapter on Cisco Unified CM Trunks, page 6-1.

Design Guidance and Assistance

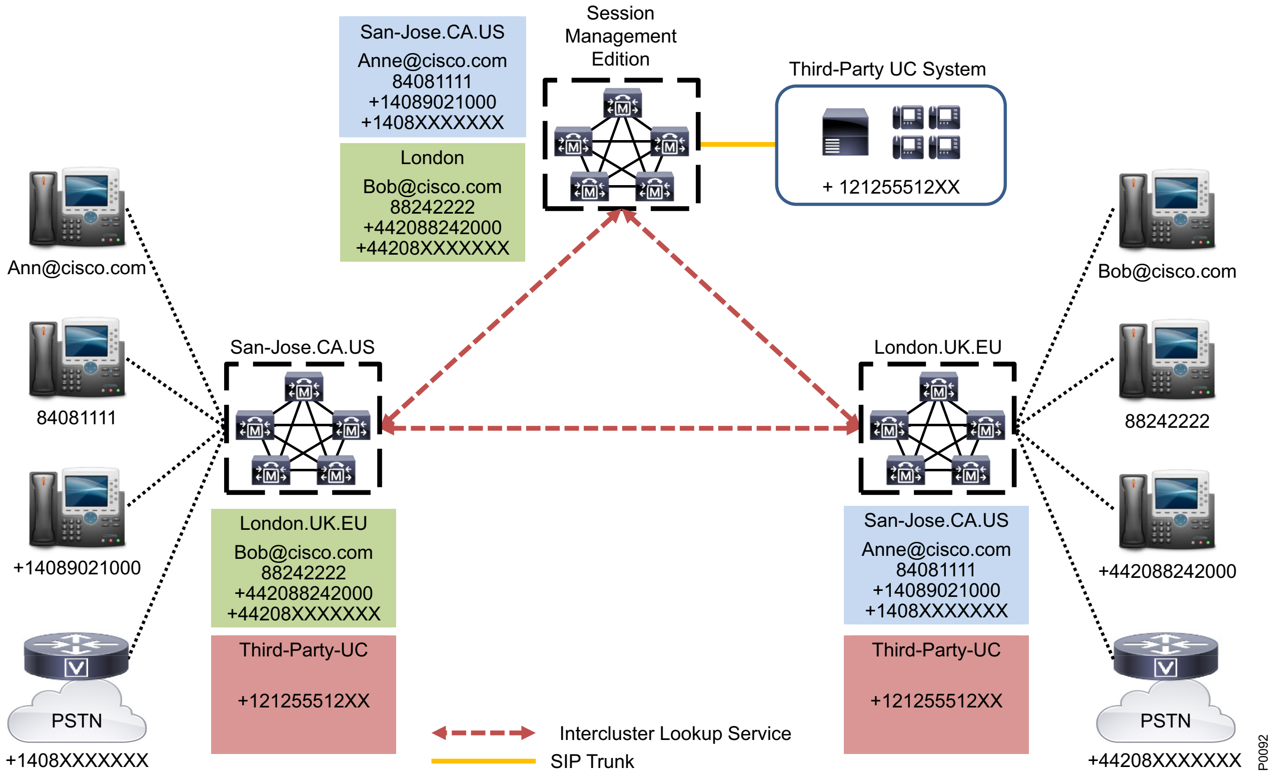
For detailed information on trunk configuration for Unified CM Session Management Edition designs and deployments, refer to the chapter on Cisco Unified CM Trunks, page 6-1.

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

Intercluster Lookup Service (ILS) and Global Dial Plan Replication (GDPR)

Global Dial Plan Replication (GDPR) uses the Intercluster Lookup Service (ILS) to share dial plan information between participating ILS-enabled clusters. GDPR allows each cluster to distribute information about its associated URIs, +E.164 numbers, enterprise numbers, +E.164 patterns, enterprise patterns, and PSTN failover numbers. Each participating cluster shares a common Global Dial Plan catalogue, which contains every number and URI advertised with GDPR and a corresponding route string that identifies in which cluster (or end Unified Communications system) the number or URI resides. (See Figure 10-9.)

Figure 10-9 ILS and GDPR Number, Pattern, and URI Distribution



With GDPR, each cluster advertises its dial plan information (numbers and URIs) with a location attribute, known as a *route string*. When a call is placed to a number or URI, Unified CM checks to see whether the number or URI is associated to a device within the cluster. If it is not, Unified CM searches its GDPR catalogue for the number or URI. If a match is found in the Global Dial Plan catalogue, GDPR returns the route string that corresponds to the cluster where the number or URI resides. Unified CM uses the returned route string as a candidate to match to an existing SIP route pattern and corresponding SIP trunk. (See Figure 10-10 and Figure 10-11.)

Figure 10-10 ILS and GDPR Number, Pattern, and URI Lookup

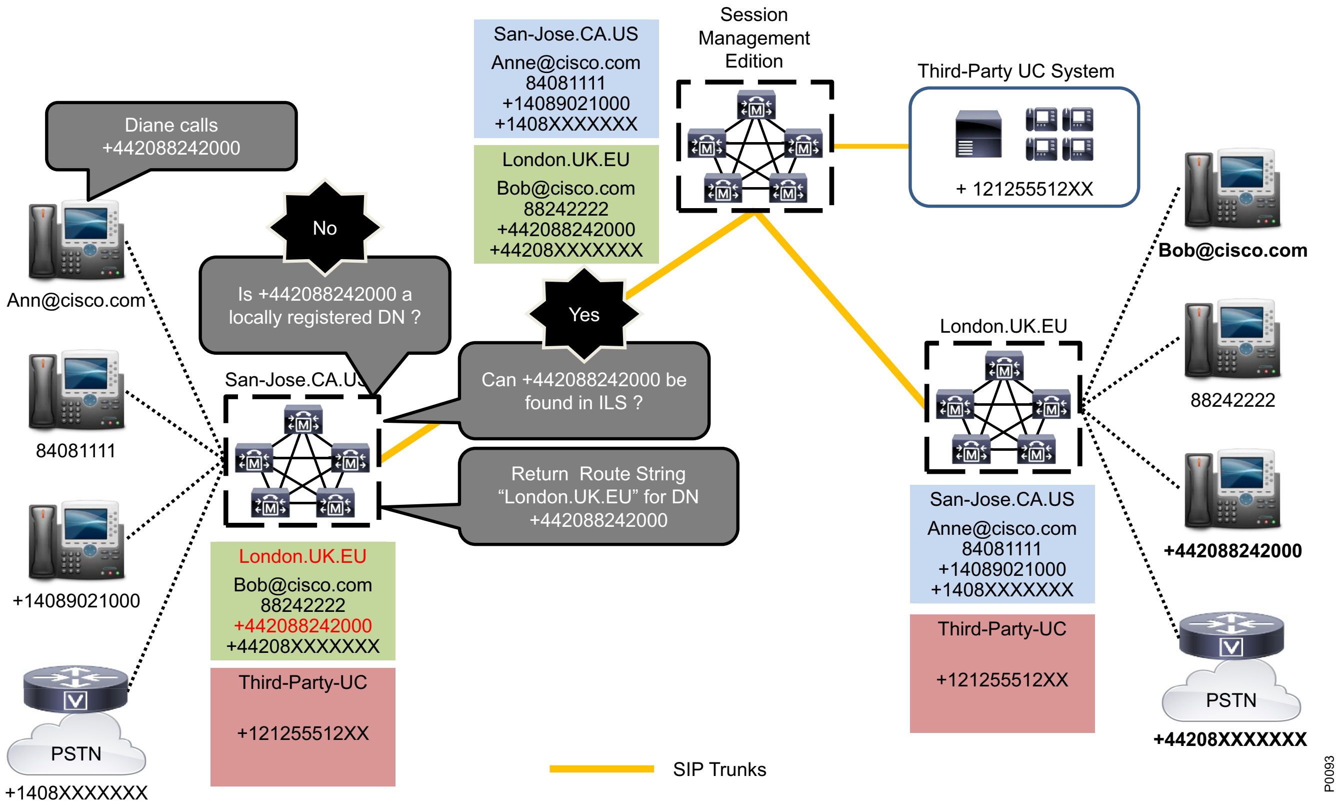
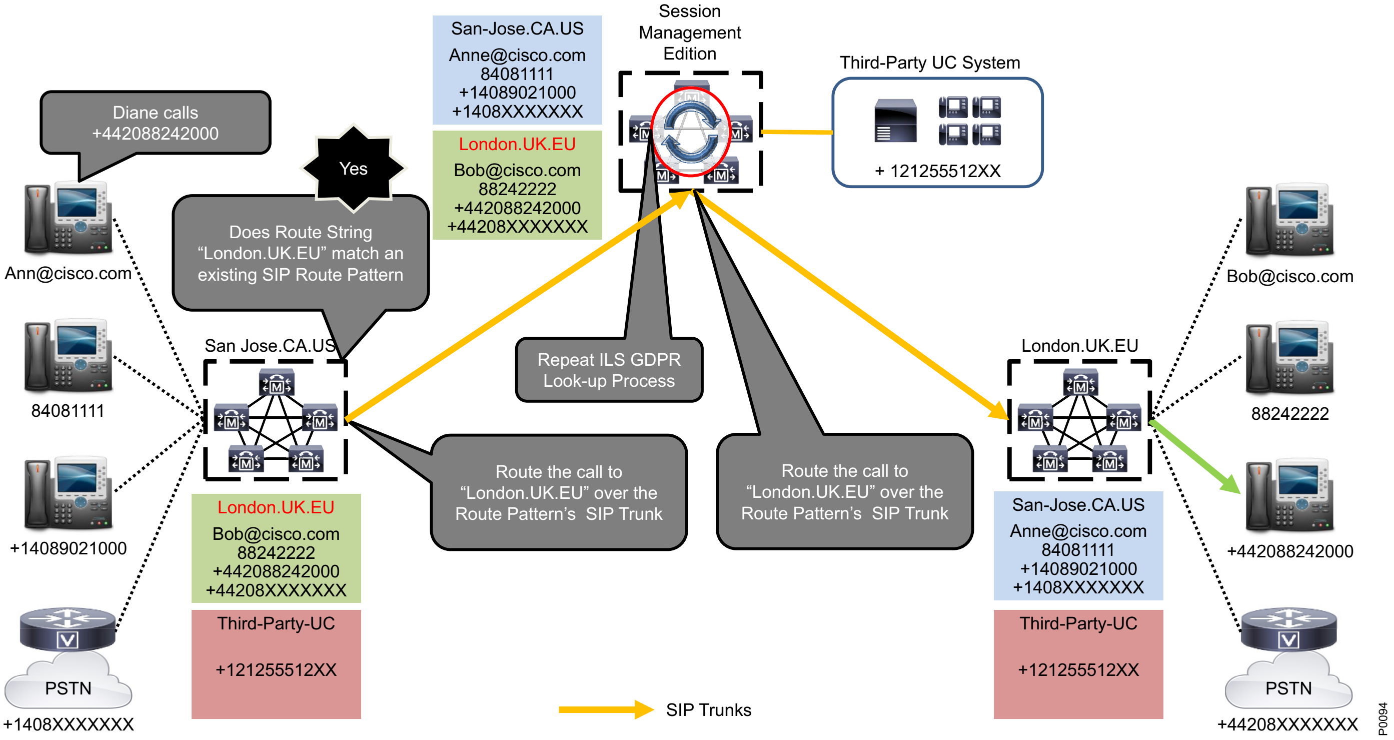


Figure 10-11 ILS and GDPR Call Routing



Benefits of ILS and GDPR

Using GDPR is significantly different from using a standard dial plan with numeric route patterns. Instead of requiring a route pattern for each unique number range within the Unified Communications network, GDPR distributes the numbers, number patterns, and URIs, and only a single SIP route pattern is required for each cluster within the Unified Communications network. Numbers and URIs associated with third-party unified communications systems (and Unified CM clusters that do not support ILS and GDPR) can be imported as catalogues into GDPR and distributed through ILS with a route string that corresponds to each unified communications system. Because both individual numbers and route patterns corresponding to groups of numbers can be advertised with GDPR, this abstraction of numbers and number ranges away from numeric route patterns allows GDPR to simply and easily support highly fragmented dial plans with many number ranges. Each cluster using ILS and GDPR can block and purge individual numbers and number ranges advertised from other participating clusters.

Each GDPR number type (+E.164 number, enterprise number, +E.164 pattern, or enterprise pattern) is placed into a specific partition when learned through ILS, allowing per-user or per-device class of service to be applied based on number type partitions and calling search spaces.

Cisco Unified Border Element also supports number and URI call routing using dial peers that match on a GDPR route string value, which is sent to Cisco Unified Border Element during call setup over a Unified CM SIP trunk. GDPR route string matching with Cisco IOS dial peers is supported with Cisco IOS releases 15.3(3)M, 15.4(1)T (ISR), 15.3(3)S (ASR), and later releases.

Deployments for the Collaboration Edge

The border between an enterprise Unified Communications network and the outside world is often referred to as the *collaboration edge*. Access to an enterprise network from the outside world can take a number of forms. For example, users can be teleworkers working from home, mobile workers with Wi-Fi internet access to the enterprise, or users making calls to and from the IP PSTN or calls to and from other businesses over the internet. The Unified Communications equipment needed at the collaboration edge largely depends upon the type of enterprise access required, which can be classified broadly into three categories:

- VPN access
- VPN-less mobile, teleworker, and business-to-business (B2B) access
- IP PSTN access

These three deployment options for the collaboration edge are discussed in the sections that follow.

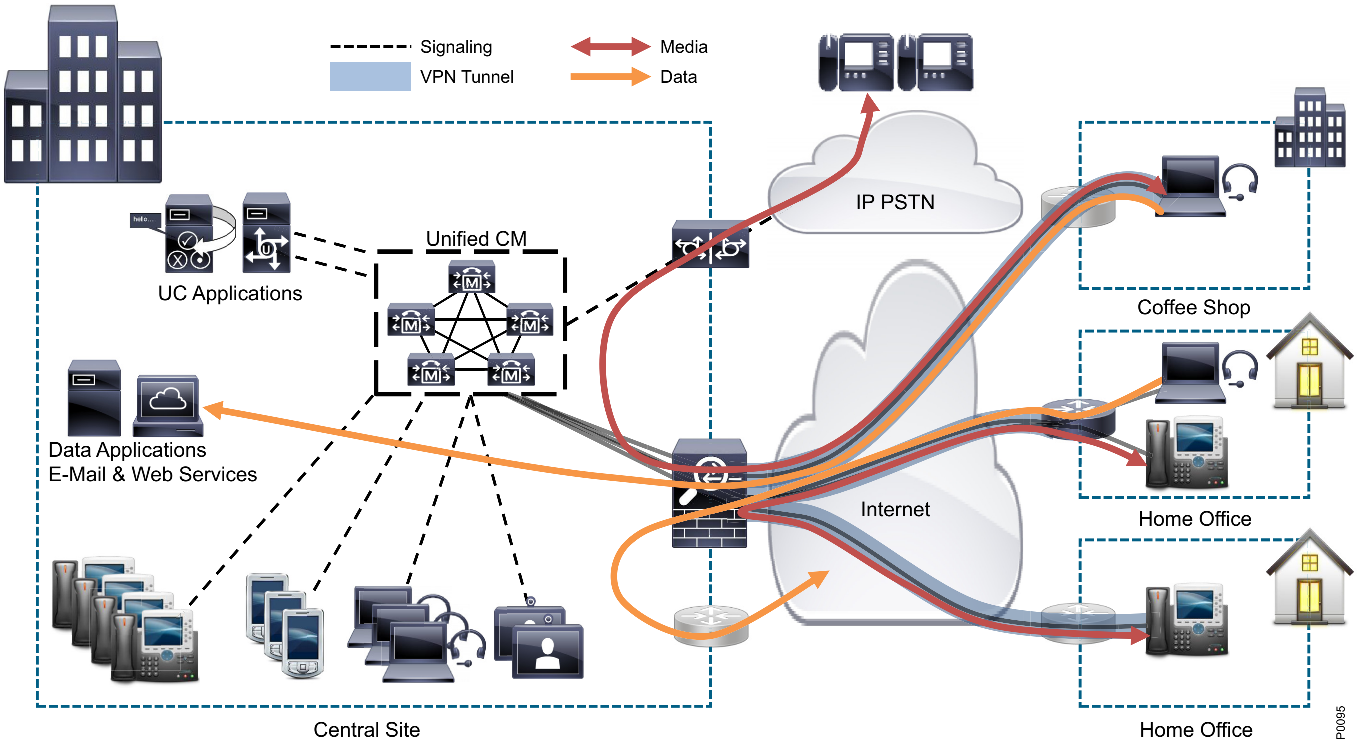
VPN Enterprise Access Deployments

VPN access to the enterprise network is probably the most common form of enterprise access today, and it can be provided in several ways:

- Mobile devices such as laptops, tablets, and smartphones can deploy the Cisco AnyConnect VPN client to access Unified Communications services (for example, Unified CM, Cisco IM and Presence, Cisco Unity, and others) as well as business application services (for example, the corporate email system and internal websites) within the enterprise network. With this VPN connection established, Unified Communications soft clients such as Cisco Jabber and Cisco IP Communicator can register with Unified CM and make voice, video, and encrypted calls between enterprise devices.
- Home office workers with one or more enterprise devices can deploy a Cisco Virtual Office (CVO) Integrated Services Router (ISR) to extend the enterprise network to their homes over a VPN. The CVO VPN connection provides connected devices with access to Unified Communications services (for example, Unified CM, Cisco IM and Presence, Cisco Unity, and others) as well as business application services (for example, the corporate email system and internal websites) within the enterprise network. With the CVO VPN connection established, Unified Communications soft clients and IP phones can register with Unified CM and make voice, video, and encrypted calls between enterprise devices.
- The Cisco VPN client for Cisco Unified IP Phones provides enterprise access for a subset of Cisco Unified IP Phone models. For more information on the devices supporting the Cisco VPN client for Cisco Unified IP Phones, see the chapter on Collaboration Endpoints, page 8-1. The phone's VPN client creates a tunnel (for the phone only), allowing to it register with Unified CM and to make voice, video, and encrypted calls between enterprise devices. A computer connected to the phone's PC port is responsible for authenticating and establishing its own tunnel to the enterprise with VPN client software.

VPN access gives the user access to all Unified Communications and business applications within the enterprise by creating a secure encrypted tunnel from the device to the VPN head-end. All traffic, including traffic destined for the internet and media for calls between VPN users, must always traverse the enterprise network rather than be established directly from the device over the internet to its destination. (See Figure 10-12.)

Figure 10-12 Collaboration Edge VPN Access



All of the above devices use their VPN client to connect to the enterprise network via a VPN head-end platform such as a Cisco Adaptive Security Appliance (ASA 5500) or a Cisco VPN aggregation router.

For more information on VPN access solutions, refer to the Teleworking and BYOD solutions guides available at

http://www.cisco.com/en/US/netsol/ns982/networking_solutions_program_home.html#~slng

VPN-less Enterprise Access and B2B Deployments

Instead of using a VPN tunnel, VPN-less clients establish a secure and encrypted signaling path to an enterprise edge traversal platform such as Cisco Expressway or Cisco Unified Border Element. VPN-less clients register with Unified CM within the enterprise, and the secured channel to the edge traversal platform allows the client to establish an encrypted media path over the internet for calls to other enterprise devices. Inside the enterprise signaling is typically unencrypted, whereas media can optionally remain encrypted.

Unlike VPN clients, VPN-less clients provide enterprise access to Unified Communications applications only; business applications within the enterprise (such as corporate email and internal websites) are not accessible, and connections to the internet are made directly from the device rather than through the enterprise. Cisco VPN-less client access can be deployed using Cisco Expressway and/or Cisco Unified Border Element as the edge traversal platform.

In addition to providing enterprise access for remote and/ or mobile employees, Cisco Expressway and Cisco Unified Border Element can also be deployed for business-to-business (B2B) communications. For all VPN-less deployments, Cisco Expressway is designed to provide voice, video, and IM and Presence communications, whereas Cisco Unified Border Element supports voice calls only.

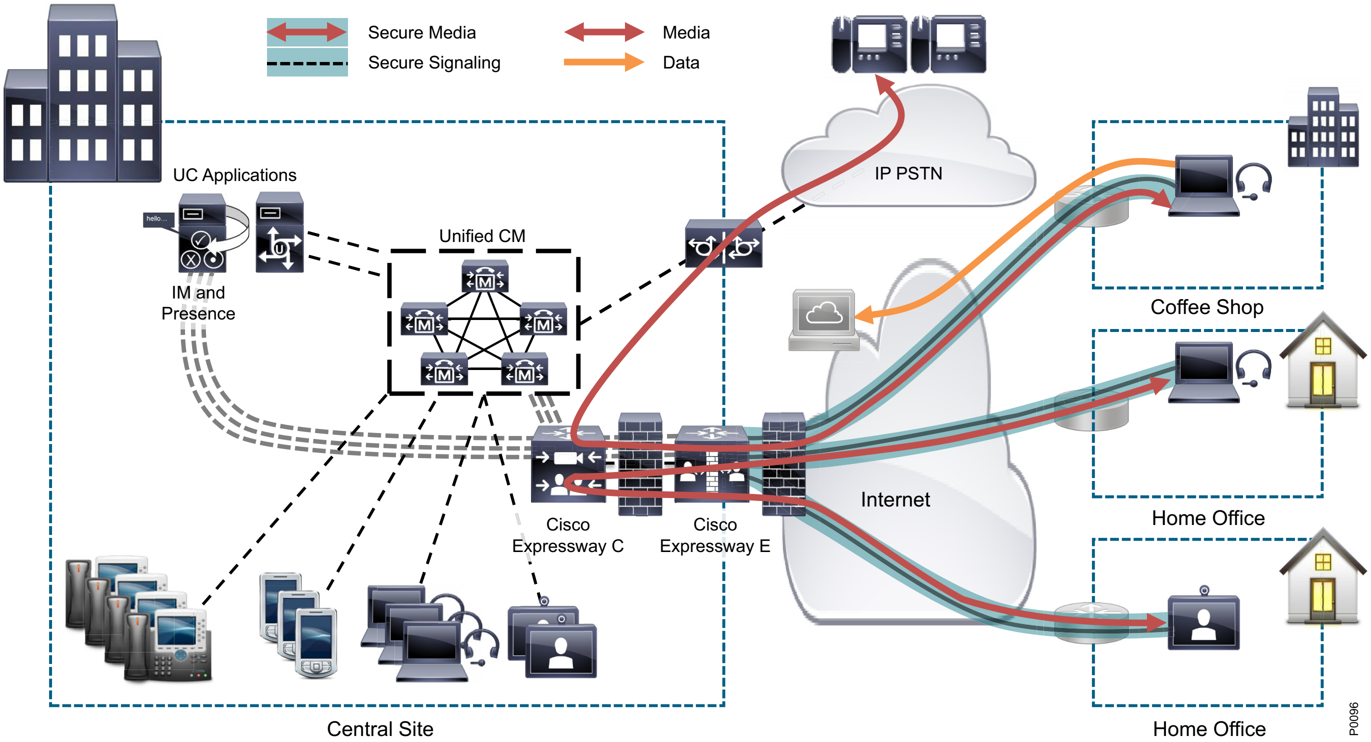
These VPN-less edge traversal options, the endpoints, and features that they support are discussed in the following sections.

VPN-Less Access with Cisco Expressway

This deployment type uses Cisco Expressway C and Expressway E. Cisco Expressway E can be placed either in a DMZ or in the public internet, and it communicates by means of Cisco Expressway C to the Unified CM cluster in the enterprise network. (See Figure 10-13.) Cisco Expressway supports VPN-less access primarily for Cisco Jabber clients and TelePresence endpoints. Voice, video, encrypted calls, and IM and Presence are supported between enterprise endpoints. Media and signaling for calls between remote VPN-less devices traverse Cisco Expressway C and Expressway E. For specific information on the range of endpoints supported with Cisco Expressway VPN-less enterprise access, see the chapter on Collaboration Endpoints, page 8-1. For more information on Cisco Expressway VPN-less client access, refer to the documentation available at the following locations:

- http://www.cisco.com/en/US/netsol/ns1246/index.html
- http://www.cisco.com/en/US/products/ps13435/index.html

Figure 10-13 Collaboration Edge VPN-Less Access with Cisco Expressway

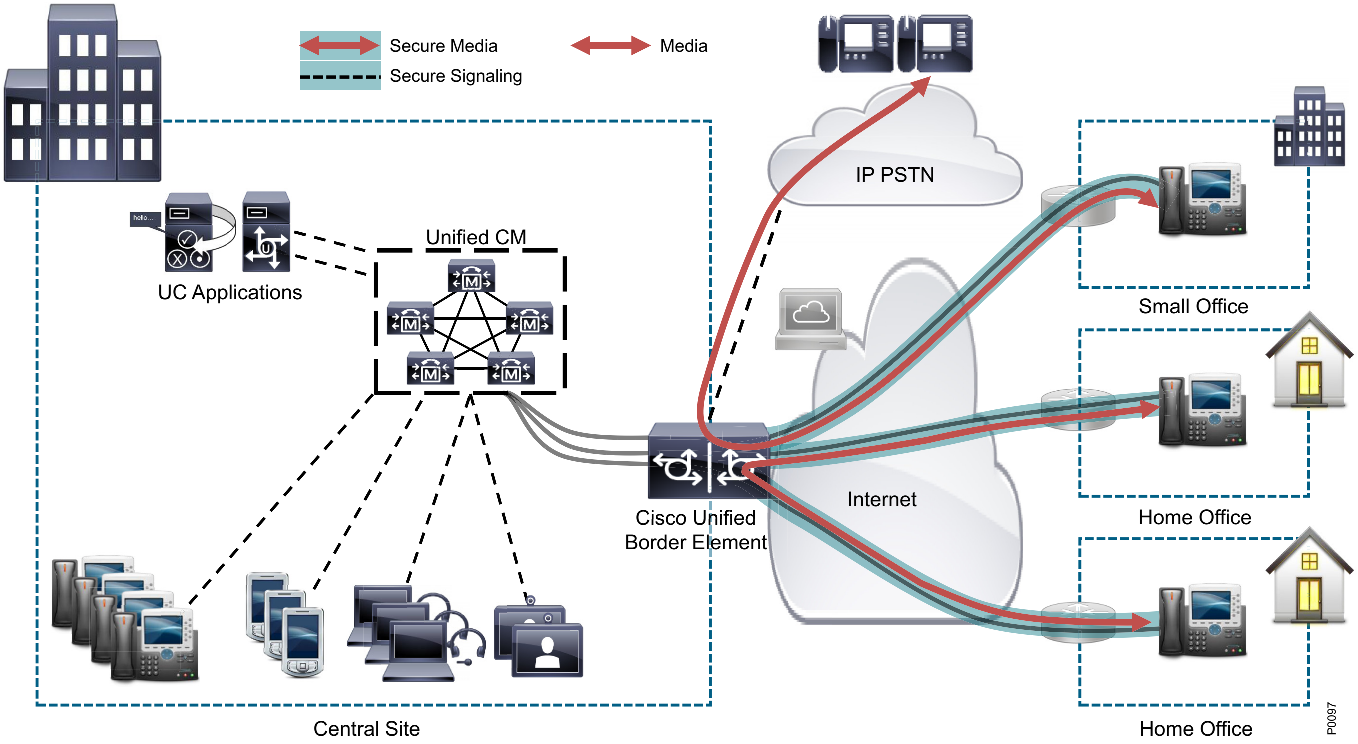


VPN-Less Access with Cisco Unified Border Element

This deployment type uses Cisco Unified Border Element as the edge traversal platform. Cisco Unified Border Element can be placed either in a DMZ or in the public internet, and it communicates directly to the Unified CM cluster in the enterprise network. (See Figure 10-14.) Cisco Unified Border Element supports VPN-less access for Unified CM IP phones. Only voice calls are supported between enterprise endpoints through the Cisco Unified Border Element traversal platform. Today, calls between remote IP phones always traverse Cisco Unified Border Element. For specific information on the range of endpoints supported with Cisco Unified Border Element VPN-less enterprise access see the chapter on Collaboration Endpoints, page 8-1. For more information on Cisco Unified Border Element VPN-less client access, refer to the documentation available at the following locations:

- http://www.cisco.com/en/US/products/sw/voicesw/ps5640/index.html
- http://www.cisco.com/en/US/products/sw/voicesw/ps5640/products_data_sheets_list.html

Figure 10-14 Collaboration Edge VPN-Less Access with Cisco Unified Border Element



IP PSTN Deployments

IP PSTN deployments are increasing in popularity and are gradually replacing existing TDM-based PSTN access. SIP is commonly used as the IP PSTN access protocol, and today many service providers offer a voice-only service to the IP PSTN through a session border controller such a Cisco Unified Border Element. Session border controllers are SIP Back-to-Back User Agents (B2BUAs) and are typically used in flow-through mode, where both the voice media and SIP signalling for each call flow through Cisco Unified Border Element. (See Figure 10-15.) As a B2BUA in flow-through mode, Cisco Unified Border Element can implement sophisticated QoS marking and call admission control policies while also providing support for transcoding, encryption, media forking for call recording applications, and scripting that allows SIP messages and SDP content to be modified for interoperability. For more info on Cisco Unified Border Element features and functions, refer to the latest version of the Cisco Unified Border Element Enterprise Edition data sheet, available at

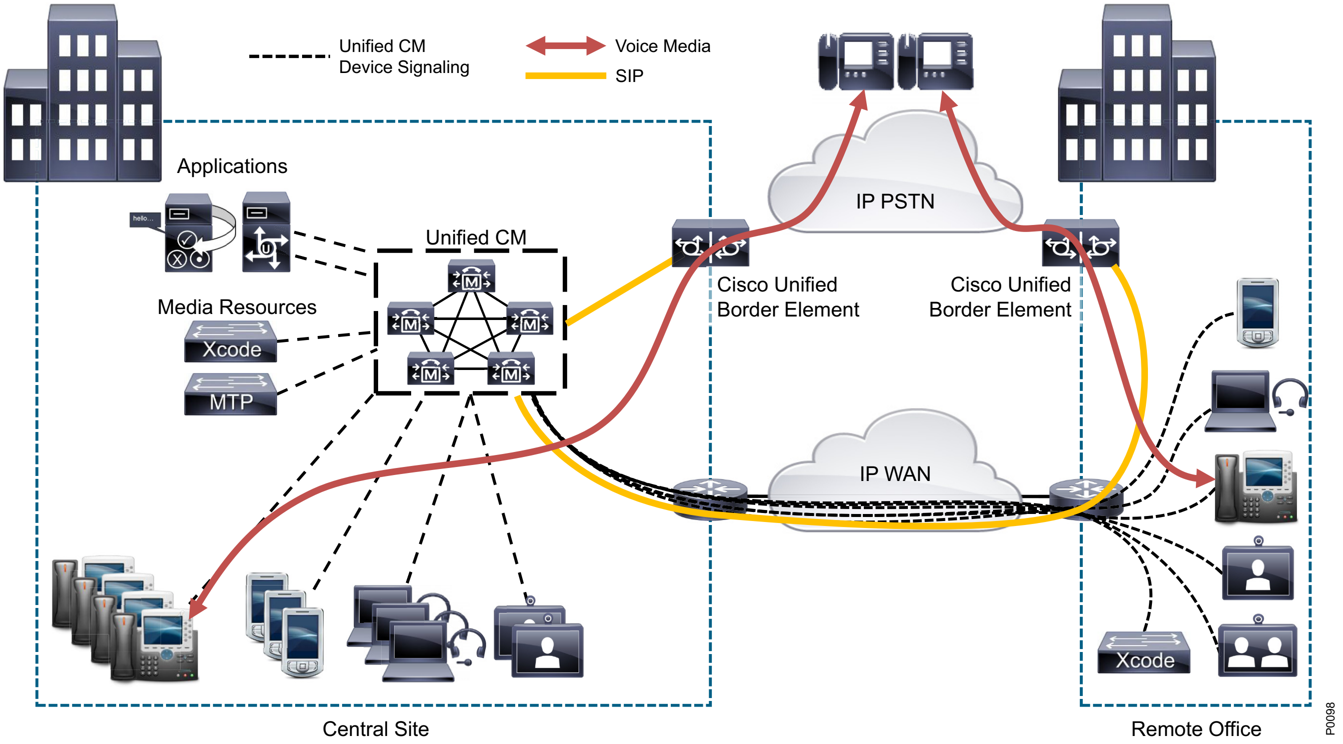
http://www.cisco.com/en/US/products/sw/voicesw/ps5640/products_data_sheets_list.html

Cisco Unified Border Element is supported on wide range of Cisco routing platforms, from the Cisco 800 Series Integrated Services Routers (ISR) to the Cisco 1000 Series Aggregation Service Routers (ASR). Depending on the hardware platform, Cisco Unified Border Element can provide session scalability from 4 to 16,000 concurrent voice calls. Cisco Unified Border Element also provides redundancy on the following platforms:

- The Cisco ISR-G2 platforms, which can provide box-to-box redundancy with media preservation for stable active calls (Cisco Unified Border Element Release 8.5 in Cisco IOS Release 15.1.2T or later).
- The Cisco ASR platforms, which can provide box-to-box or in-box redundancy with media and signaling preservation (stateful failover) for stable active calls.

Note Access to the IP PSTN and access to the enterprise for VPN-less clients can be deployed on the same Cisco Unified Border Element platform.

Figure 10-15 Collaboration Edge IP PSTN Access



Geographic Deployment Options for IP PSTN

SIP trunks may be connected to IP PSTN service providers in several different ways, depending on the desired architecture. The two most common architectures for this connectivity are centralized trunks and distributed trunks.

Centralized trunks connect to the service provider (SP) through one logical connection (although there may be more than one physical connection for redundancy) with session border controllers (SBCs) such as the Cisco Unified Border Element. All IP PSTN calls to and from the enterprise use this set of trunks, and for most calls, media and signaling traverse the enterprise WAN to connect devices in the enterprise to those in the PSTN.

Distributed trunks connect to the service provider through several logical connections. Each branch of an enterprise may have its own local trunk to the service provider. With distributed trunks, media from the branch no longer needs to traverse the enterprise WAN but can flow directly to the service provider through a local SBC.

Each connectivity model has its own advantages and disadvantages. Centralized trunks are generally easier to deploy in terms of both physical equipment and configuration complexity. Distributed trunks have the advantage of local hand-off of media and better number portability from local providers. Alternatively, a hybrid connectivity model that combines some centralized and distributed IP PSTN access can capture the advantages of both forms of IP PSTN deployment.

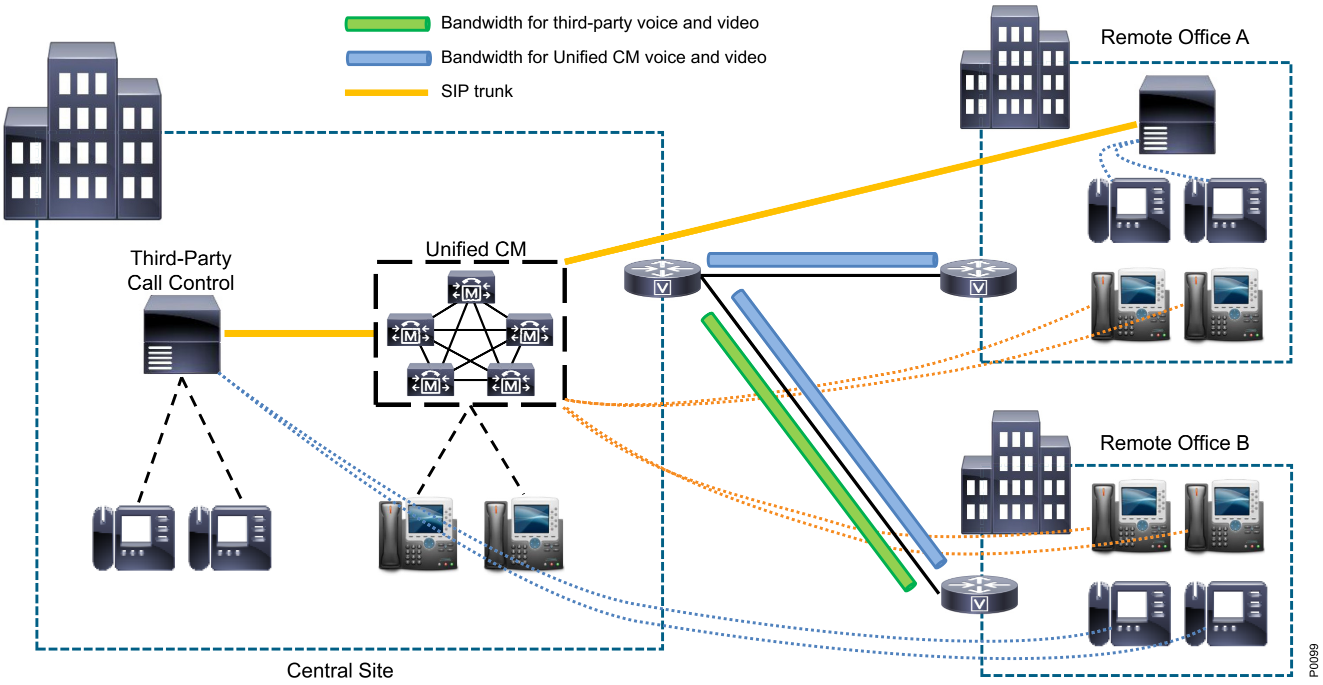
Design Considerations for Dual Call Control Deployments

In general, deployments where endpoints are registered to Cisco Unified CM and a third-party call control platform introduce a degree of complexity into any Collaboration Solution design, particularly with respect to dial plan and call admission control. In these dual call control deployments, both Unified CM and the third-party call control platform use independent mechanisms for call admission control and have independent dial plans, and the degree of complexity introduced is determined by the deployment model used. (See Figure 10-16.)

In campus deployments with dual call control, call admission control is not required and the dial plan in each call control system is relatively straight forward: If an endpoint cannot be found on one system, the call is forwarded to the other system. Standard dial plan configuration (calling search spaces and partitions) can be used to prevent routing loops between the systems.

In multisite centralized call processing deployments, a trade-off is typically made between call admission control complexity and dial plan complexity. If the Unified CM cluster and the third-party call control platform are deployed at the central site only, the dial plan is relatively straightforward, but the WAN bandwidth that each system uses for call admission control must be separately considered and provisioned for in the WAN. If additional third-party call control systems are deployed at remote sites where third-party endpoints reside, call admission control complexity can be avoided at the expense of a more fragmented dial plan. These trade-offs are discussed in more detail in the following sections.

Figure 10-16 Dual Call Control Deployments with Centralized and Distributed Third-Party Systems



Call Admission Control Considerations in Dual Call Control Deployments

Call admission control provides mechanisms for preventing the oversubscription of network bandwidth by limiting the number of calls that are allowed on the network at a given time based on overall call capacity of the call processing components and on network bandwidth.

In dual call control deployments, both Unified CM and the third-party call control platform use independent mechanisms for call admission control.

In multisite centralized call processing deployments, where the Unified CM cluster and the third-party call control platform are deployed at the central site only, consideration needs to be made in the WAN for the bandwidth that each platform uses for call admission control. At remote sites where third-party endpoints are registered to a locally deployed third-party call control system, these call admission control considerations can be avoided

Multisite Centralized Unified CM Deployments with Distributed Third-Party Call Control

Dual call control deployment models that use a third-party call control system at every site where third-party endpoints reside, can be tightly integrated with Unified CM call admission control. This may be achieved by using a SIP trunk to Unified CM from the third-party call control system at each site and configuring this SIP trunk in the same Unified CM location as the Unified CM endpoints residing at that site. Note, however, that while this approach resolves the call admission control issues for Unified CM and third-party call control, it does so at the expense of provisioning a large number of third-party call control systems, which in turn fragments the dial plan in the network.

Multisite Centralized Unified CM Deployments with Centralized Third-Party Call Control

Like Unified CM, a third-party call control system may be centralized to serve third-party endpoints residing at multiple sites over the WAN. With this type of deployment, the third-party call control system provides call admission control for calls between third-party endpoints at different sites; and likewise, the centralized Unified CM cluster provides call admission control for calls between Unified CM endpoints at different sites. Because the Unified CM cluster and the third-party call control system use independent call admission control mechanisms, at sites where both Unified CM and third-party endpoints reside in a centralized call control deployment, separate amounts of bandwidth must provisioned in the WAN for Unified CM call admission control and third-party call control. When calls are made between Unified CM endpoints and third-party endpoints in the same site, call admission control bandwidth will be decremented by both the third-party call control system and the Unified CM cluster even though the media path for the call might not traverse the WAN. Although not ideal from a call admission control perspective, this centralized call processing design is more cost effective in terms of hardware and it reduces dial plan fragmentation (because endpoints are registered to either the centralized Unified CM cluster or the centralized third-party call control system only).

A pragmatic approach should be taken for each dual call control deployment. From a strategic perspective, deploying a third-party call control system at every branch today might not make good commercial sense. However, if accurate call admission control is the design priority, then this deployment model might be appropriate. Likewise, for deployments with a centralized third-party call control and a centralized Unified CM cluster, the issue of independent call admission control domains can be addressed by over-provisioning bandwidth in the WAN.

Dial Plan Considerations in Dual Call Control Deployments

With only two call control systems in the deployment, if numbers are used to address endpoints and if the dialed number matches the local pattern of the local Unified CM cluster or third-party call control system, then the call will be sent to the locally attached endpoint. If the number matches the pattern for the other (remote) call control, then the call must be sent to the non-local endpoint through the interconnecting SIP trunk. In case of an overlapping dial plan, both the Unified CM cluster and the third-party call control system are able to send the call that does not match any internally registered endpoint, to the other call control cluster.

As the number of call control systems within a deployment increases, dial plan fragmentation also increases. This issue can be further exacerbated if endpoints from two or more call controls exist within the same site and do not have separate or easily summarized number ranges. In this case a default route cannot be used, but either of two options can be deployed to route calls in a Unified Communications system with multiple call control systems and a highly fragmented dial plan:

- Use an explicit route pattern and corresponding trunk for each of the unique number ranges associated with each call control.
- Within the Unified CM (and SME, if used) deployment, use the Intercluster Lookup Service (ILS) and Global Dial Plan Replication (GDPR) to share information about the number ranges supported by each Unified CM cluster and each third-party unified communications system. For third-party systems and their associated devices, import each unique number range into GDPR and associate each imported number range with a route string (a label that identifies the call control system). When a Unified CM user dials a number, Unified CM checks to see if the number is registered to its cluster. If the number is not registered to the Unified CM cluster, Unified CM searches ILS for the called number and its corresponding route string. The route string identifies the call control cluster where the number resides, which is used to match a SIP route pattern that then forwards the call over a SIP trunk toward its destination.

If alphanumeric URIs are used to address and call endpoints registered to Unified CM and the third-party call control system, then call routing can be implemented in either of the following ways, depending on the deployment:

- For deployments where only a single third-party call control system exists with a single SIP trunk to a Unified CM cluster, a default SIP route can be configured on Unified CM and the third-party call control system, so that calls to endpoints that are not found on one call control are sent to the other call control.
- If multiple third-party call control systems are deployed, use the Intercluster Lookup Service (ILS) and Global Dial Plan Replication (GDPR) to share information about the URIs supported by each Unified CM cluster and each third-party unified communications system. For URI-based call routing when a Unified CM user dials a URI, Unified CM checks to see if the URI is registered to its cluster. If it is not, Unified CM searches the ILS for the called URI and its corresponding route-string. The route string identifies the call control cluster where the URI resides, and it is used to match a SIP route pattern, which then forwards the call over a SIP trunk toward the destination URI. For URI-based endpoints registered to a third-party call control system, the list of URIs registered to the third-party call control system must be imported manually into ILS along with the corresponding route string for the third-party call control system.

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 9-1.

Clustering over the WAN can support two types of deployments:

- Local Failover Deployment Model, page 10-45

  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 10-52

  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 10-43, 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 10-44. 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 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 on 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 intra-cluster communications 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.

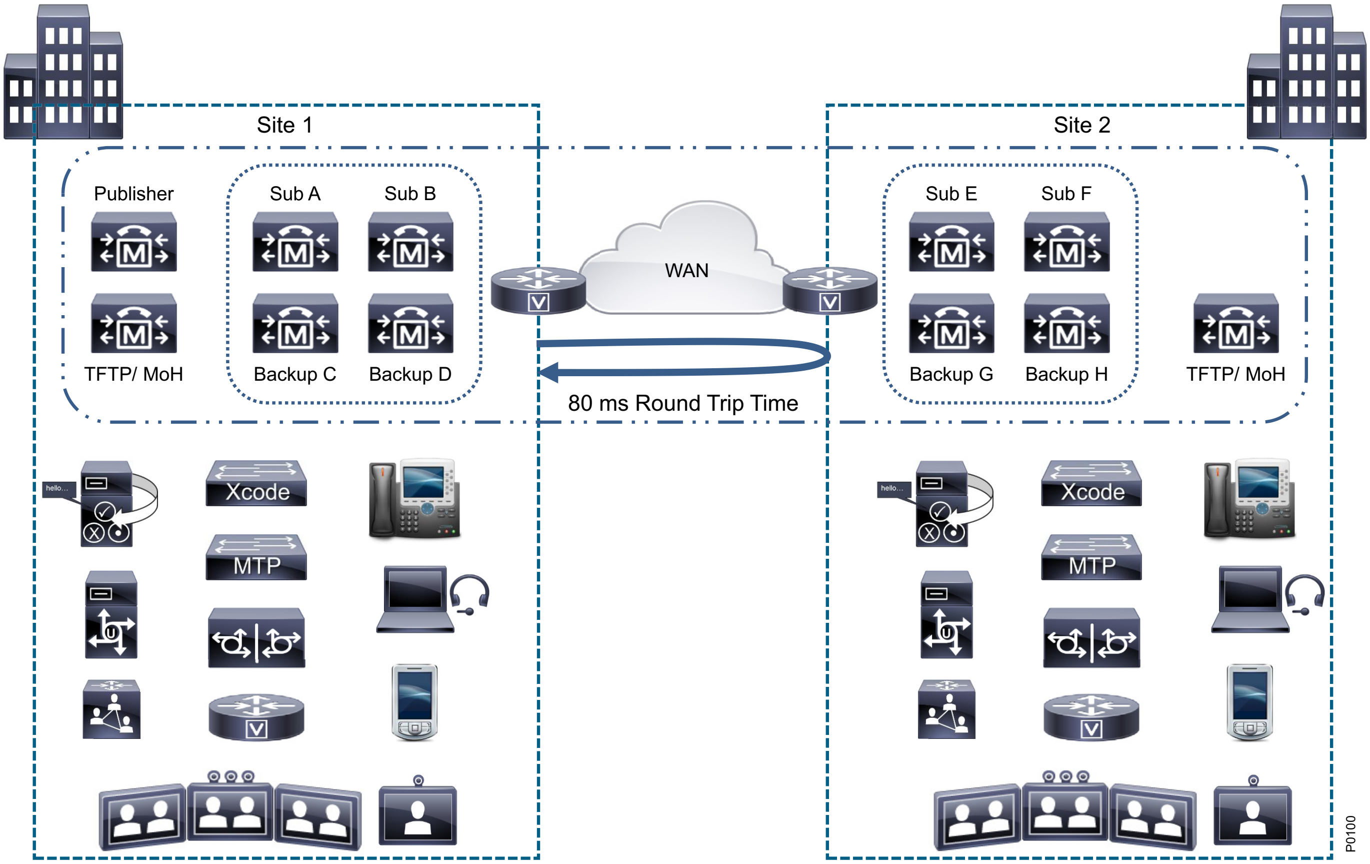
In summary, perform the following tasks to troubleshoot intra-cluster communication problems:

- 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 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 dependent on the quantity and type of servers deployed. The maximum total number of IP phones for all sites is 40,000. (See Figure 10-17.)

Figure 10-17 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 (transcoders, conferencing resources, annunciator, 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 10-43.
- 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) between each site and every other site that is clustered over the WAN. For example, if three sites are clustered over the WAN, each site would require 2 * 1.544 Mbps of WAN bandwidth for call control traffic. This minimum bandwidth requirement for call control traffic accounts for up to for 10,000 busy hour call attempts (BHCA) from one site to another site and applies only 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).

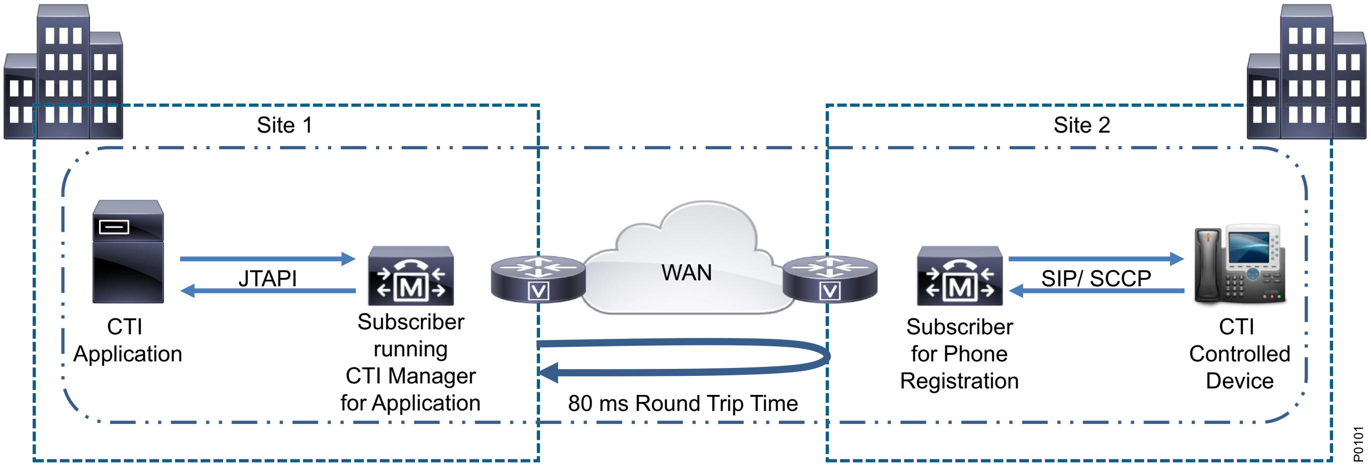
- In addition to the bandwidth required for Intra-Cluster Communication Signaling (ICCS) traffic, a minimum of 1.544 Mbps (T1) bandwidth is required for database and other inter-server traffic between the publisher and every subscriber node within the cluster.

- For customers who also want to deploy CTI Manager over the WAN (see Figure 10-18), 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 10-18 CTI over the WAN

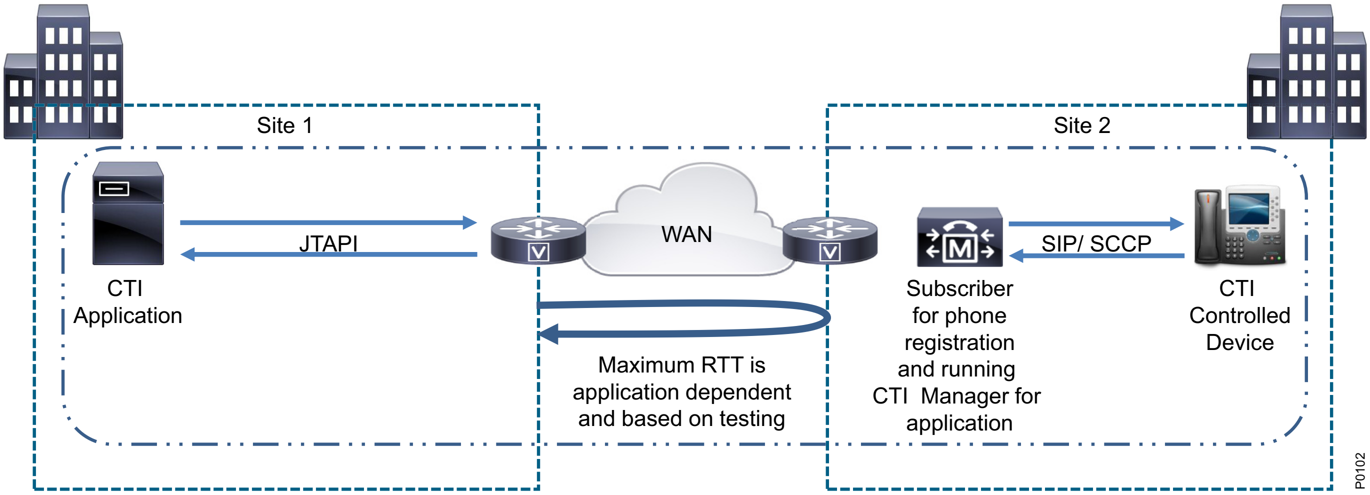


- For deployments where the J/TAPI application is remote from the Unified CM subscriber (see Figure 10-19), the following formula can be used to calculate the Quick Buffer Encoding (QBE) J/TAPI bandwidth for a typical J/TAPI application with Unified CM 8.6(2) and later releases:

  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 10-19 J/TAPI over the WAN


Site 1
Site 2
JTAPI
CTI Application
WAN
SIP/ SCCP
Subscriber for phone registration and running CTI Manager for application
CTI Controlled Device
Maximum RTT is application dependent and based on testing
P0102

Example 10-1 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 10-2 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 media traffic, depending on the voice and video codecs used for the calls. For details see Bandwidth Provisioning, page 3-45.

- 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. 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 with clustering over the WAN are connected to the central sites via the same WAN path that is used for clustering over the WAN traffic, 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. Publisher to subscriber bandwidth lower than the required 1.544 Mbps for each subscriber node can also cause the software upgrade process to take longer to complete. If a faster upgrade time is desired, additional bandwidth above the required 1.544 Mbps per remote subscriber can be provisioned during the upgrade period.

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 9-1. 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 TFTP 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. For more information on Unity Connection and clustering over then WAN, see Distributed Messaging with Clustering Over the WAN, page 19-17.

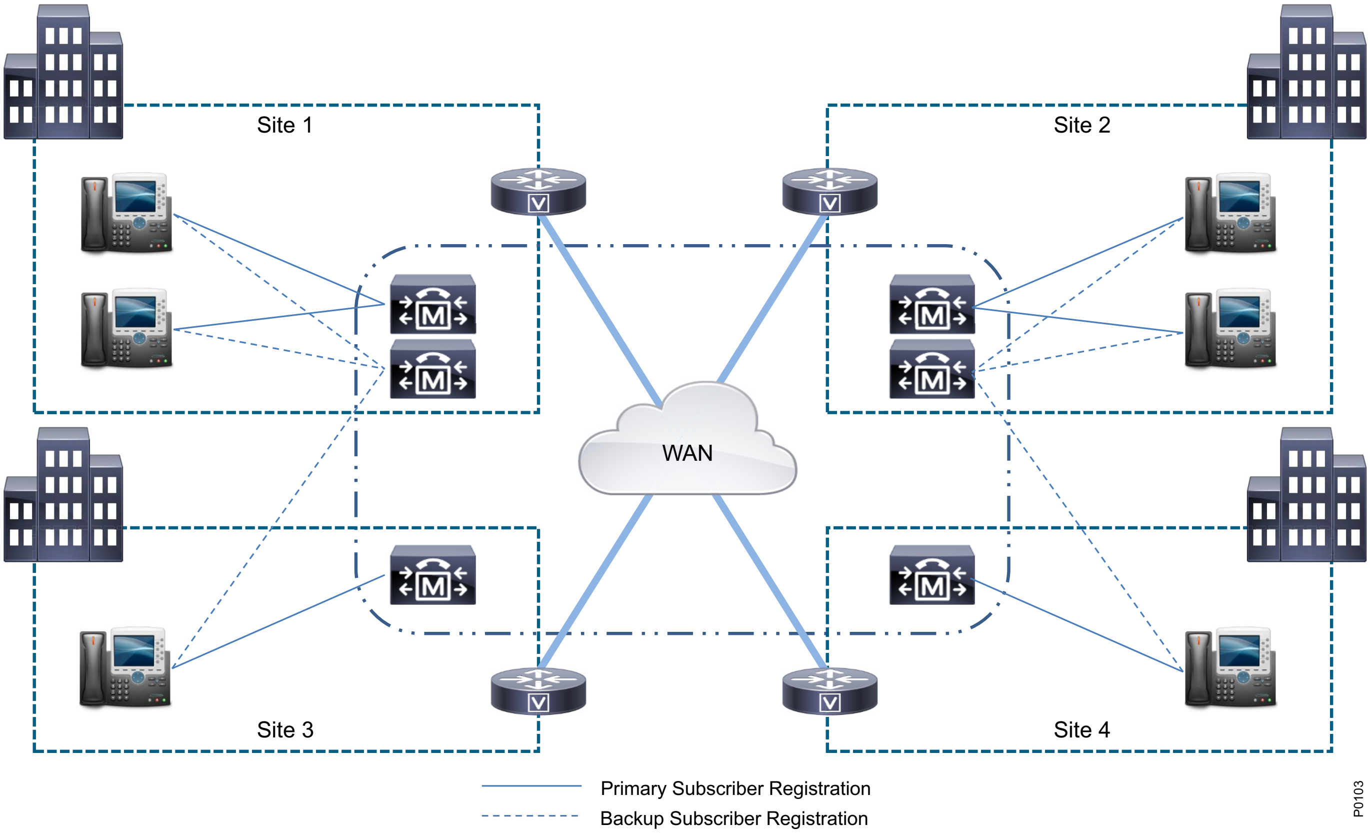
Music on Hold and Media Resources for Local Failover

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 9-1. Backup subscribers are located across the WAN at one or more of the other sites. (See Figure 10-20.)

Figure 10-20 Clustering over the WAN, Remote Failover Model with Four Sites


Site 1
Site 2
WAN
Site 3
Site 4
Primary Subscriber Registration
Backup Subscriber Registration
P0103

When implementing the remote failover model, observe all guidelines for the local failover model (see Local Failover Deployment Model, page 10-45), 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 or more Business Edition 6000 server nodes are deployed at each site to provide geographic redundancy for the Unified CM call processing application. The Business Edition 6000 server nodes may be UCS C200 or C220 Rack-Mount Servers.

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 (MTPs, conference bridges, music on hold, and so forth), 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 10-43.
- 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 inter-server traffic between the publisher and every subscriber node within the Business Edition 6000 cluster.

More than two 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 10-52). 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 on a medium-density server or 2,500 on a high-density server. A deployment of UCS C200 or C220 Rack-Mount Servers in a cluster exceeding these limits 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 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, Cisco Unified Contact Center Express and Cisco Video Communication Server) 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 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 20-20.
- For Cisco Unity Connection, see Cisco Unity Connection Redundancy and Clustering Over the WAN, page 19-20.
- For Cisco Unified Contact Center Express, see Clustering Over the IP WAN, page 24-11.

Deploying Unified Communications on Virtualized Servers

Most Cisco Collaboration Systems applications run only 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. They are tested and documented for specific guaranteed performance, capacity, and application co-residency scenarios running "full-load" Cisco Collaboration Systems virtual machines.
- Specification-based hardware that provides more hardware flexibility and that, for example, adds support for other Cisco UCS and third-party servers that are 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

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) runs on another level above the hypervisor. Hypervisors are one of the foundation elements in cloud computing and virtualization technologies, and they consolidate applications onto fewer servers.

Most of the Cisco Collaboration Systems applications are supported only with virtualization. This means that deploying the VMware vSphere ESXi hypervisor is required for those applications and that they cannot be installed directly on the server (bare metal).

VMware vCenter is a tool that helps to manage your virtual environment. With Tested Reference Configurations, VMware vCenter is not mandatory; however, it is strongly recommended when deploying a large number of hosts. With specification-based hardware, VMware vCenter is required.

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 10-56
- Cisco UCS C-Series Rack-Mount Servers, page 10-59

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

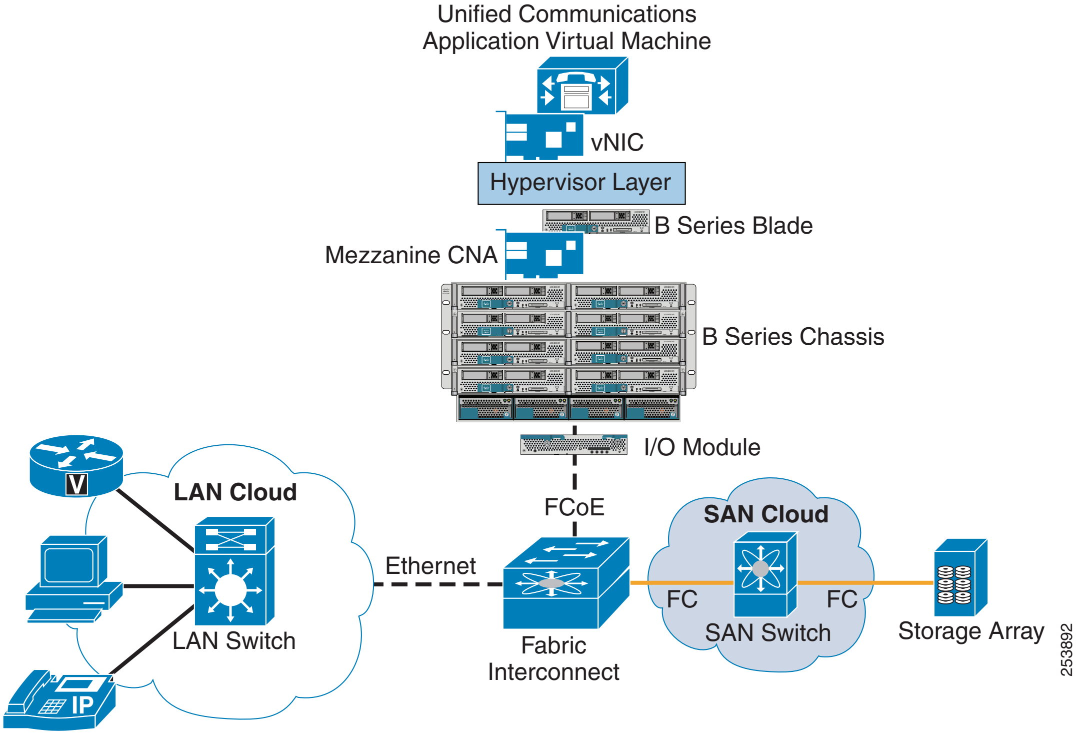
Cisco UCS E-Series server modules are blade servers designed to be deployed in Cisco Integrated Services Routers Generation 2 (ISR G2). Some Cisco Collaboration applications are supported on Cisco UCS E-Series, but the support might be limited (specification-based hardware support instead of TRC, for instance).

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 10-21.) 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 10-21 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 I/O Modules

Cisco UCS 2100 and 2200 Series I/O Modules (or Fabric Extender) 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.

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 Management Edition (SME)
- Cisco Unity Connection
- Cisco IM and Presence
- Cisco Unified Contact Center Express
- Cisco Unified Contact Center Enterprise
- Cisco Expressway

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. Each Unified Communications application should be allotted dedicated processing and memory resources to ensure that the resources are not oversubscribed.

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 local storage to the hypervisor 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 (TRCs) are also available with Cisco UCS C-Series Rack Mount Servers such as the Cisco UCS C220 and C240.

Most Cisco Collaboration Systems 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
- Cisco Expressway

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

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

Unlike with UCS B-Series, the Tested Reference Configurations based on UCS C-Series support storage for the hypervisor and the applications virtual machines locally on the directly attached storage drives, not on an FC SAN storage array. It is possible to use an external storage array with a C-Series server, but the server would then be considered as specifications-based hardware and not as a TRC.

For more details, refer to the documentation available at

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

Impact of Virtual Servers on Deployment Models

Deploying Cisco Unified Communications applications on virtualized servers supports the same deployment models as when physical servers were used. There are a few additional considerations with virtualization, however. 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. For example, the Unified CM VMware virtual application has no access to the host USB and serial ports. Therefore, Unified CM no longer supports the Cisco Messaging Interface (CMI) service for Simplified

Message Desk Interface (SMDI) integrations, fixed MoH audio source integration for live MoH audio feeds using the audio cards (MOH-USB-AUDIO=), or flash drives to these servers. The following alternative options are available:

- For MoH live audio source feed, consider using Cisco IOS-based gateway multicast MoH for live audio source connectivity.
- For saving system install logs, use virtual floppy softmedia.

There is no alternative option for the Cisco Messaging Interface (CMI) service for Simplified Message Desk Interface (SMDI) integrations.

Call Routing and Dial Plan Distribution Using Call Control Discovery (CCD) for the Service Advertisement Framework (SAF)

The Cisco Service Advertisement Framework (SAF) is a Cisco IOS service routing protocol that can be used to share call routing and dial plan information automatically between call processing platforms. SAF allows non-Cisco call processing platforms (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 with Call Control Discovery (CCD)

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).

Note SAF CCD supports the distribution of internal enterprise DN ranges only, unlike GDPR which supports the distribution of internal enterprise DN ranges, external (PSTN) DN ranges, and URIs.

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.

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 SAF CCD in Unified Communications networks, refer to the SAF sections in the *Unified Communications Deployment Models* chapter of the *Cisco Unified Communications System 9.0 SRND*, available at

http://www.cisco.com/en/US/docs/voice_ip_comm/cucm/srnd/9x/models.html

For more information on SAF itself, refer to the *Service Advertisement Framework (SAF)* section in the *Network Infrastructure* chapter of the *Cisco Collaboration 9.x Solution Reference Network Designs (SRND)*, available at

http://www.cisco.com/en/US/docs/voice_ip_comm/cucm/srnd/collab09/netstruc.html

SAF CCD Deployment Considerations

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 100,000 learned DN patterns per Unified CM cluster (Default value = 20,000 learned patterns)
- 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 autonomous system (AS) and consisting of Cisco Unified CM and SAF CCD running on a Cisco IOS platform, SAF CCD system-wide scalability is limited to 6,000 learned DN patterns.