CHAPTER 2

Unified Communications Deployment Models

Last revised on: July 15, 2009

This chapter describes the deployment models for Cisco Unified Communications Manager (Unified CM) 6.*x*. For design guidance with earlier releases of Cisco Unified CM, refer to the Unified Communications Solution Reference Network Design (SRND) documentation available at

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

Each Cisco Unified Communications solution is based on a Unified CM deployment model, and the type of deployment model is based on one or more of the following factors:

- Number of call processing agent clusters
- Number of IP phones
- Locations of the call processing agent cluster(s) and IP phones

The following sections describe the various types of deployment models:

What's New in This Chapter

Table 2-1 lists the topics that are new in this chapter or that have changed significantly from previous releases of this document.

Table 2-1 New or Changed Information Since the Previous Release of This Document

New or Revised Topic	Described in:
Cisco Unified Communications Manager Express in SRST mode can now support Cisco Unity Express.	Best Practices for Unified CME in SRST Mode, page 2-11
Propagation delay limits have been redefined.	WAN Considerations, page 2-19

Single Site

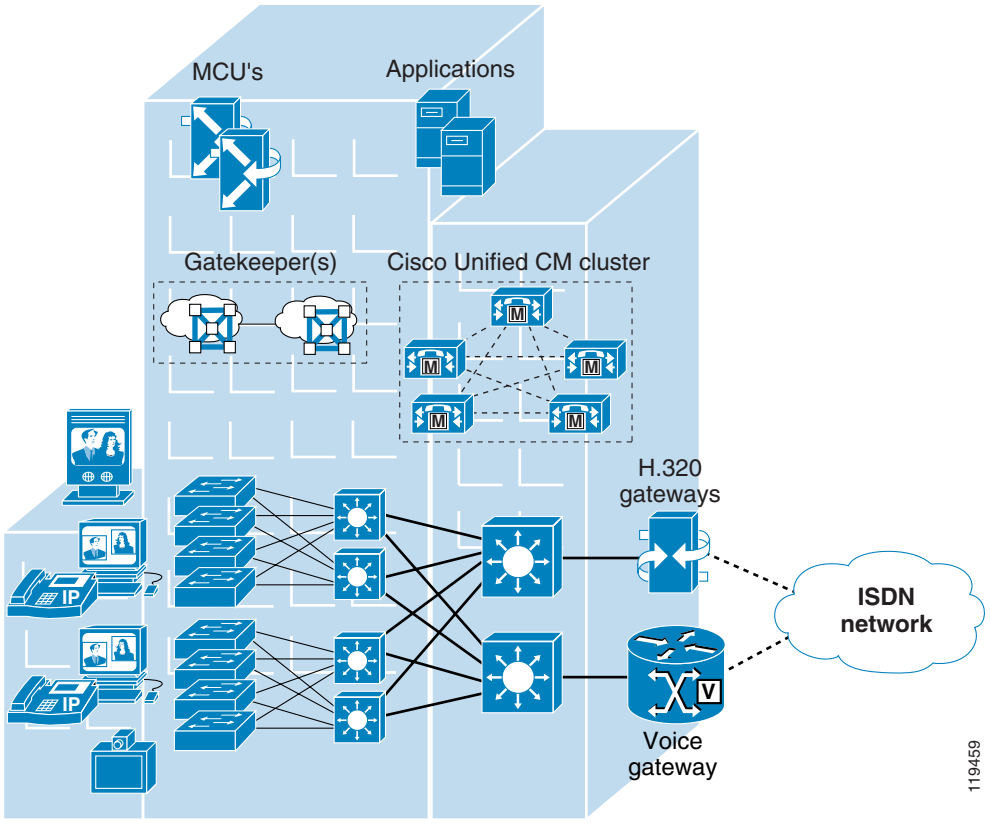
The single-site model for Cisco Unified Communications consists of a call processing agent cluster located at a single site, or campus, with *no* telephony services provided over an IP WAN. An enterprise would typically deploy the single-site model over a LAN or metropolitan area network (MAN), which carries the voice traffic within the site. In this model, calls beyond the LAN or MAN use the public switched telephone network (PSTN).

The single-site model has the following design characteristics:

- Single Cisco Unified CM cluster.
- Maximum of 30,000 configured and registered Skinny Client Control Protocol (SCCP) or Session Initiation Protocol (SIP) IP phones or SCCP video endpoints per cluster.
- Maximum of 1100 H.323 devices (gateways, MCUs, trunks, and clients) or MGCP gateways per Unified CM cluster.
- PSTN for all calls outside the site.
- Digital signal processor (DSP) resources for conferencing, transcoding, and media termination point (MTP).
- Voicemail, unified messaging, Cisco Unified Presence, audio and video components.
- Capability to integrate with legacy private branch exchange (PBX) and voicemail systems.
- H.323 clients, MCUs, and H.323/H.320 gateways that require a gatekeeper to place calls must register with a Cisco IOS Gatekeeper (Cisco IOS Release 12.3(8)T or greater). Unified CM then uses an H.323 trunk to integrate with the gatekeeper and provide call routing and bandwidth management services for the H.323 devices registered to it. Multiple Cisco IOS Gatekeepers may be used to provide redundancy.
- MCU resources are required for multipoint video conferencing. Depending on conferencing requirements, these resources may be either SCCP or H.323, or both.
- H.323/H.320 video gateways are needed to communicate with H.320 videoconferencing devices on the public ISDN network.
- High-bandwidth audio (for example, G.711, G.722, or Cisco Wideband Audio) between devices within the site.
- High-bandwidth video (for example, 384 kbps or greater) between devices within the site. The Cisco Unified Video Advantage Wideband Codec, operating at 7 Mbps, is also supported.

Figure 2-1 illustrates the model for a Cisco Unified Communications network within a single campus or site.

Figure 2-1 Single-Site Deployment



Best Practices for the Single-Site Model

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

- Provide a highly available, fault-tolerant infrastructure based on a common infrastructure philosophy. A sound infrastructure is essential for easier migration to Cisco Unified Communications, integration with applications such as video streaming and video conferencing, and expansion of your Cisco Unified Communications deployment across the WAN or to multiple Unified CM clusters.
- Know the calling patterns for your enterprise. Use the single-site 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).
- Use Media Gateway Control Protocol (MGCP) gateways for the PSTN if you do *not* require H.323 functionality. This practice simplifies the dial plan configuration. H.323 might be required to support specific functionality such as support for Signaling System 7 (SS7) or Non-Facility Associated Signaling (NFAS).

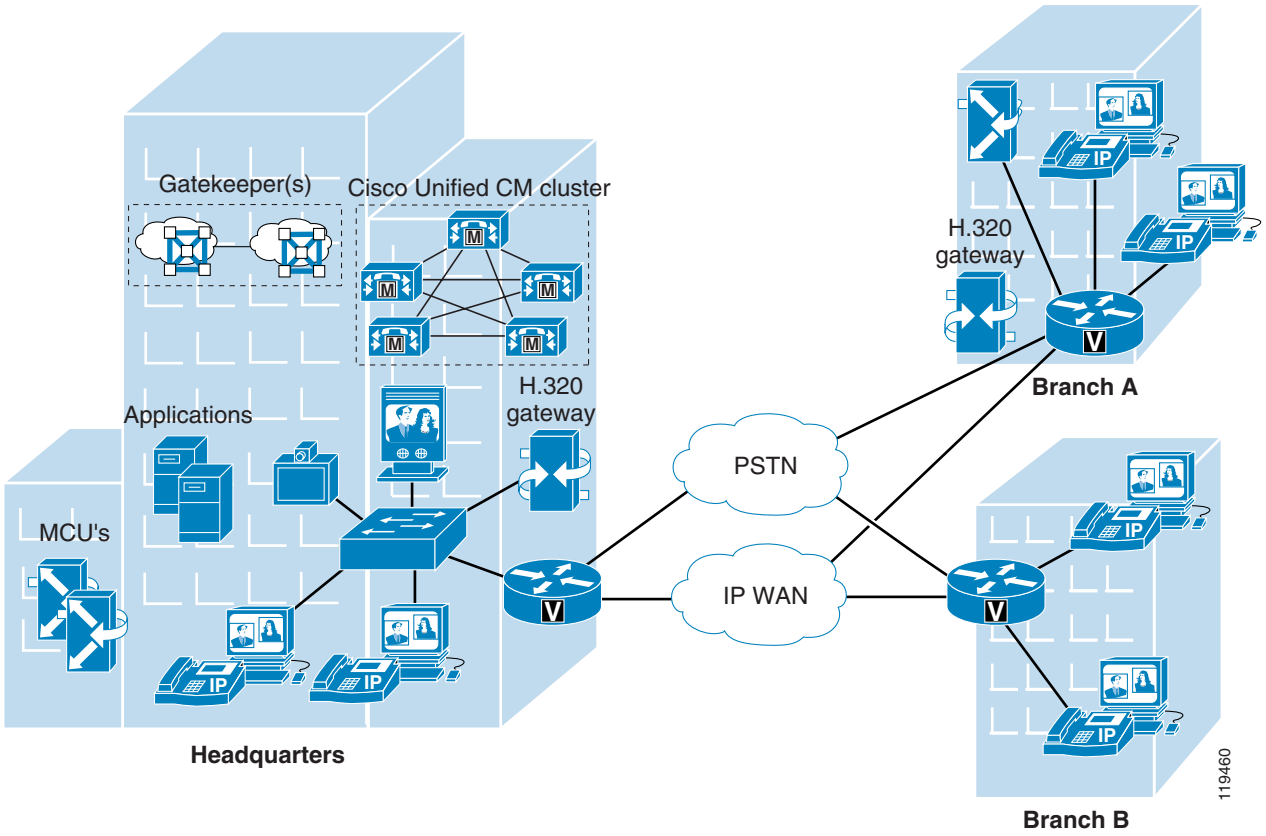
- Implement the recommended network infrastructure for high availability, connectivity options for phones (in-line power), Quality of Service (QoS) mechanisms, and security. (See Network Infrastructure, page 3-1.)
- Follow the provisioning recommendations listed in the chapter on Call Processing, page 8-1.

Multiple Sites with Centralized Call Processing

The model for a multisite deployment with centralized call processing consists of a single call processing agent cluster that provides services for many remote sites and uses the IP WAN to transport Cisco Unified Communications traffic between the sites. The IP WAN also carries call control signaling between the central site and the remote sites. Figure 2-2 illustrates a typical centralized call processing deployment, with a Unified CM cluster as the call processing agent at the central site and an IP WAN with QoS enabled to connect all the sites. The remote sites rely on the centralized Unified CM cluster to handle their call processing. Applications such as voicemail, presence servers, interactive voice response (IVR) systems, and so forth, are typically centralized as well to reduce the overall costs of administration and maintenance.

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 2-2 Multisite Deployment with Centralized Call Processing



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

- Single Unified CM cluster.
- Maximum of 30,000 configured and registered Skinny Client Control Protocol (SCCP) or Session Initiation Protocol (SIP) IP phones or SCCP video endpoints per cluster.
- Maximum of 1000 locations or branch sites per Unified CM cluster.
- Maximum of 1100 H.323 devices (gateways, MCUs, trunks, and clients) or 1100 MGCP gateways per Unified CM cluster.
- PSTN for all external calls.
- Digital signal processor (DSP) resources for conferencing, transcoding, and media termination point (MTP).
- Voicemail, unified messaging, Cisco Unified Presence, audio and video components.
- Capability to integrate with legacy private branch exchange (PBX) and voicemail systems.
- H.323 clients, MCUs, and H.323/H.320 gateways that require a gatekeeper to place calls must register with a Cisco IOS Gatekeeper (Cisco IOS Release 12.3(8)T or greater). Unified CM then uses an H.323 trunk to integrate with the gatekeeper and provide call routing and bandwidth management services for the H.323 devices registered to it. Multiple Cisco IOS Gatekeepers may be used to provide redundancy.
- MCU resources are required for multipoint video conferencing. Depending on conferencing requirements, these resources may be either SCCP or H.323, or both, and may all be located at the central site or may be distributed to the remote sites if local conferencing resources are required.
- H.323/H.320 video gateways are needed to communicate with H.320 videoconferencing devices on the public ISDN network. These gateways may all be located at the central site or may be distributed to the remote sites if local ISDN access is required.
- High-bandwidth audio (for example, G.711, G.722, or Cisco Wideband Audio) between devices in the same site, and low-bandwidth audio (for example, G.729 or G.728) between devices in different sites.
- High-bandwidth video (for example, 384 kbps or greater) between devices in the same site, and low-bandwidth video (for example, 128 kbps) between devices at different sites. The Cisco Unified Video Advantage Wideband Codec, operating at 7 Mbps, is recommended only for calls between devices at the same site.
- Minimum of 768 kbps or greater WAN link speeds. Video is *not* recommended on WAN connections that operate at speeds lower than 768 kbps.
- Unified CM locations (static or RSVP-enabled) provide call admission control, and automated alternate routing (AAR) is also supported for video calls.
- Survivable Remote Site Telephony (SRST) versions 4.0 and higher support video. However, versions of SRST prior to 4.0 do *not* support video, and SCCP video endpoints located at remote sites become audio-only devices if the WAN connection fails.
- Cisco Unified Communications Manager Express (Unified CME) versions 4.0 and higher may be used for remote site survivability instead of an SRST router. Unified CME also provides more features than the SRST router during WAN outage.
- Cisco Unified Communications Manager Express (Unified CME) can be integrated with the Cisco Unity server in the branch office or remote site. The Cisco Unity server is registered to the Unified CM at the central site in normal mode and can fall back to Unified CME in SRST mode when Unified CM is not reachable, or during a WAN outage, to provide the users at the branch offices with access to their voicemail with MWI.

Connectivity options for the IP WAN include:

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

Routers that reside at the WAN edges require quality of service (QoS) mechanisms, such as priority queuing and traffic shaping, to protect the voice traffic from the data traffic across the WAN, where bandwidth is typically scarce. In addition, a call admission control scheme is needed to avoid oversubscribing the WAN links with voice traffic and deteriorating the quality of established calls. For centralized call processing deployments, the *locations* (static or RSVP-enabled) construct within Unified CM provides call admission control. (Refer to the chapter on Call Admission Control, page 9-1, for more information on locations.)

A variety of Cisco gateways can provide the remote sites with PSTN access. When the IP WAN is down, or if all the available bandwidth on the IP WAN has been consumed, users at the remote sites can dial the PSTN access code and place their calls through the PSTN. The Cisco Unified Survivable Remote Site Telephony (SRST) feature, available for both SCCP and SIP phones, provides call processing at the branch offices for Cisco Unified IP Phones if they lose their connection to the remote primary, secondary, or tertiary Unified CM or if the WAN connection is down. Cisco Unified SRST functionality is available on Cisco IOS gateways running the SRST feature or on Cisco Unified CME Release 4.0 and higher running in SRST mode. Unified CME running in SRST mode provides more features for the phones than SRST on a Cisco IOS gateway.

Best Practices for the Centralized Call Processing Model

Follow these guidelines and best practices when implementing a multisite deployment with centralized call processing:

- Minimize delay between Unified CM and remote locations to reduce voice cut-through delays (also known as clipping).
- Use the locations (static or RSVP-enabled) mechanism in Unified CM to provide call admission control into and out of remote branches. See the chapter on Call Admission Control, page 9-1, 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 720 phones, while Unified CME running in SRST mode supports 240 phones. (For the latest SRST or Unified CME platform and code specifications, refer to the SRST and Unified CME documentation available at http://www.cisco.com.) Generally speaking, however, the choice of whether to adopt a centralized call processing or distributed call processing approach for a given site depends on a number of factors such as:
  - IP WAN bandwidth or delay limitations
  - Criticality of the voice network
  - Feature set needs
  - Scalability

  - Ease of management
  - Cost

  If a distributed call processing model is deemed more suitable for the customer's business needs, the choices include installing a Unified CM cluster at each site or running Unified CME at the remote sites.

- At the remote sites, use the following features to ensure call processing survivability in the event of a WAN failure:
  - For SCCP phones, use SRST on a Cisco IOS gateway or Unified CME running in SRST mode.
  - For SIP phones, use SIP SRST.
  - For MGCP phones, use MGCP Gateway Fallback.

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

Remote Site Survivability

When deploying Cisco Unified Communications across a WAN with the centralized call processing model, you should take additional steps to ensure that data and voice services at the remote sites are highly available. Table 2-2 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 2-2 Strategies for High Availability at the Remote Sites

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

The first two solutions listed in Table 2-2 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 2-2 use an ISDN backup link to provide survivability during WAN failures. The two deployment options for ISDN backup are:

- Data-only ISDN backup

  With this option, ISDN is used for data survivability only, while SRST or Unified CME is used for voice survivability. Note that you should configure an access control list on the branch router to prevent Skinny Client Control Protocol (SCCP) or Session Initiation Protocol (SIP) traffic from entering the ISDN interface, so that signaling from the IP phones does not reach the Unified CM at the central site.

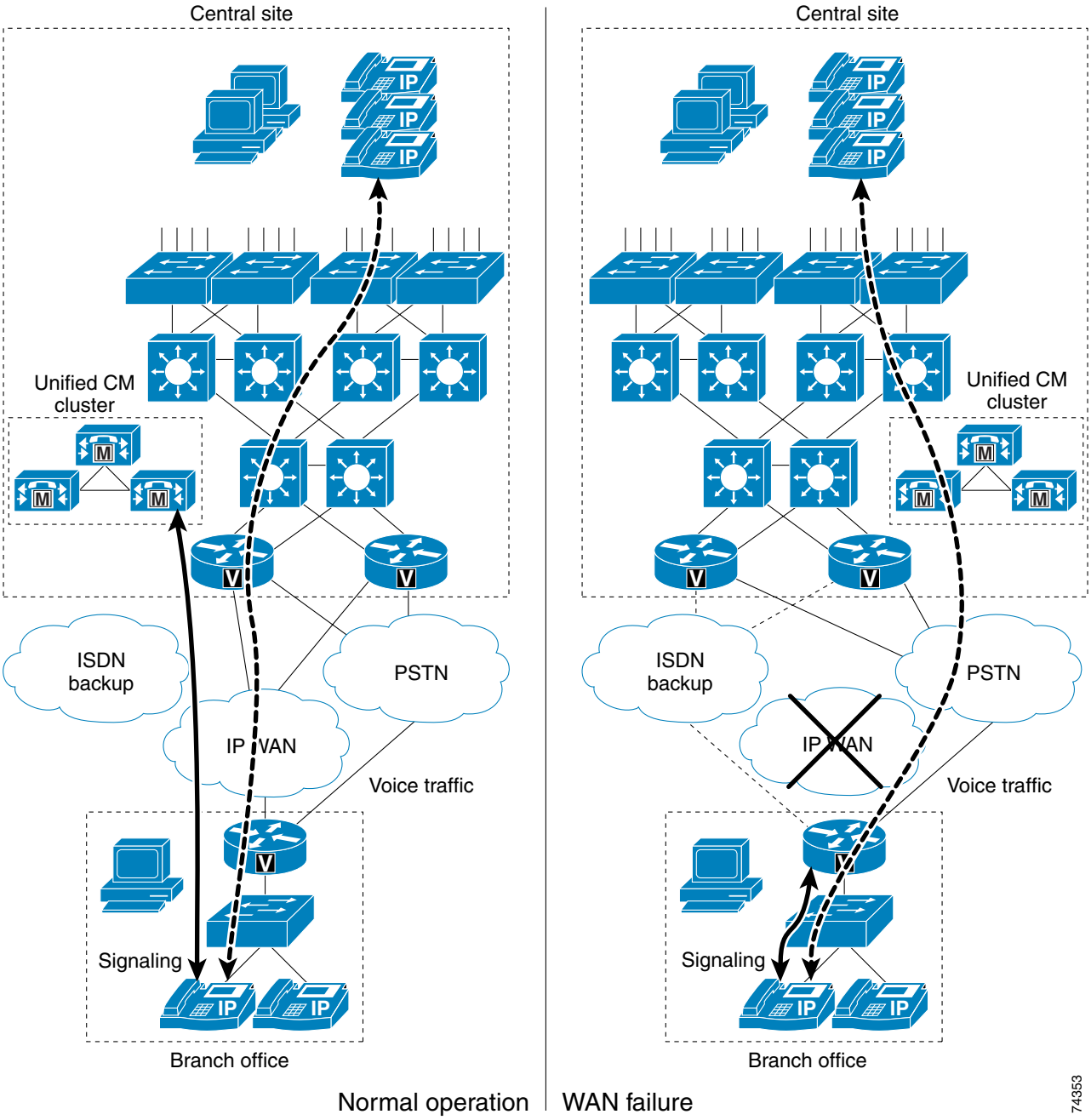
- Data and voice ISDN backup

  With this option, ISDN is used for both data and voice survivability. In this case, SRST or Unified CME 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 2-2, Survivable Remote Site Telephony (SRST) or Unified CME in SRST mode, provides high availability for voice services only, by providing a subset of the call processing capabilities within the remote office router and enhancing the IP phones with the ability to “re-home” to the call processing functions in the local router if a WAN failure is detected. Figure 2-3 illustrates a typical call scenario with SRST or Unified CME in SRST mode.

Figure 2-3 Survivable Remote Site Telephony (SRST) or Unified CME in SRST Mode



Under normal operations shown in the left part of Figure 2-3, the branch office connects to the central site via an IP WAN, which carries data traffic, voice traffic, and call signaling. The IP phones at the branch office exchange call signaling information with the Unified CM cluster at the central site and place their calls across the IP WAN. The branch router or gateway forwards both types of traffic (call signaling and voice) transparently and has no knowledge of the IP phones.

If the WAN link to the branch office fails, or if some other event causes loss of connectivity to the Unified CM cluster, the branch IP phones re-register with the branch router in SRST mode. The branch router, SRST, or Unified CME running in SRST mode, queries the IP phones for their configuration and uses this information to build its own configuration automatically. The branch IP phones can then make and receive calls either internally or through the PSTN. The phone displays the message "Unified CM fallback mode," and some advanced Unified CM features are unavailable and are grayed out on the phone display.

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

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

Unified CME in SRST Mode

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

- Paging
- Conferencing
- Hunt groups
- Basic automatic call distribution (B-ACD)
- Call park, call pickup, call pickup groups
- Overlay-DN, softkey templates
- Cisco IP Communicator 2.0
- Cisco Unified Video Advantage 2.0
- Integration with Cisco Unity with MWI support at remote sites, with distributed Microsoft Exchange or IBM Lotus Domino server

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

Best Practices for Unified CME in SRST Mode

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

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

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

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

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

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

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

Best Practices for SRST Router

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

- For supporting a maximum of 720 phones on a single SRST router.
- For up to 1000 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.

Voice Over the PSTN as a Variant of Centralized Call Processing

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

You can implement VoPSTN in one of the following ways:

- Using the automated alternate routing (AAR) feature. (For more information on AAR, see the section on Automated Alternate Routing, page 10-28.)
- Using a combination of dial plan constructs in both Unified CM and the PSTN gateways.

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

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

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

- Centralized voicemail requires:
  - A telephony network provider that supports redirected dialed number identification service (RDNIS) end-to-end for all locations that are part of the deployment. RDNIS is required so that calls redirected to voicemail carry the redirecting DN, to ensure proper voicemail box selection.
  - If the voicemail system is accessed through an MGCP gateway, the voicemail pilot number must be a fully qualified E.164 number.
- The Extension Mobility feature is limited to IP phones contained within a single branch site.
- All on-net (intra-cluster) calls will be delivered to the destination phone with the same call treatment as an off-net (PSTN) call. This includes the quantity of digits delivered in the call directories such as Missed Calls and Received Calls.
- Each inter-branch call generates two independent call detail records (CDRs): one for the call leg from the calling phone to the PSTN, and the other for the call leg from the PSTN to the called phone.
- There is no way to distinguish the ring type for on-net and off-net calls.
- All destination phones require a fully qualified Direct Inward Dial (DID) PSTN number that can be called directly. Non-DID DNs cannot be reached directly from a different branch site.
- With VoPSTN, music on hold (MoH) is limited to cases where the holding party is co-located with the MoH resource. If MoH servers are deployed at the central site, then only calls placed on hold by devices at the central site will receive the hold music.
- Transfers to a destination outside the branch site will result in the hairpinning of the call through the branch's gateway. Traffic engineering of the branch's gateway resources must be adjusted accordingly.

- Call forwarding of any call coming into the branch's gateway to a destination outside the branch site will result in hairpinning of the call through the gateway, thus using two trunk ports. This behavior applies to:
  - Calls forwarded to a voicemail system located outside the branch
  - Calls forwarded to an on-net abbreviated dialing destination located in a different branch

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

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

- VoPSTN Using AAR, page 2-13
- VoPSTN Using Dial Plan, page 2-14

VoPSTN Using AAR

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

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

The AAR implementation method for VoPSTN offers the following benefits:

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

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

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

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

VoPSTN Using Dial Plan

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

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

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

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

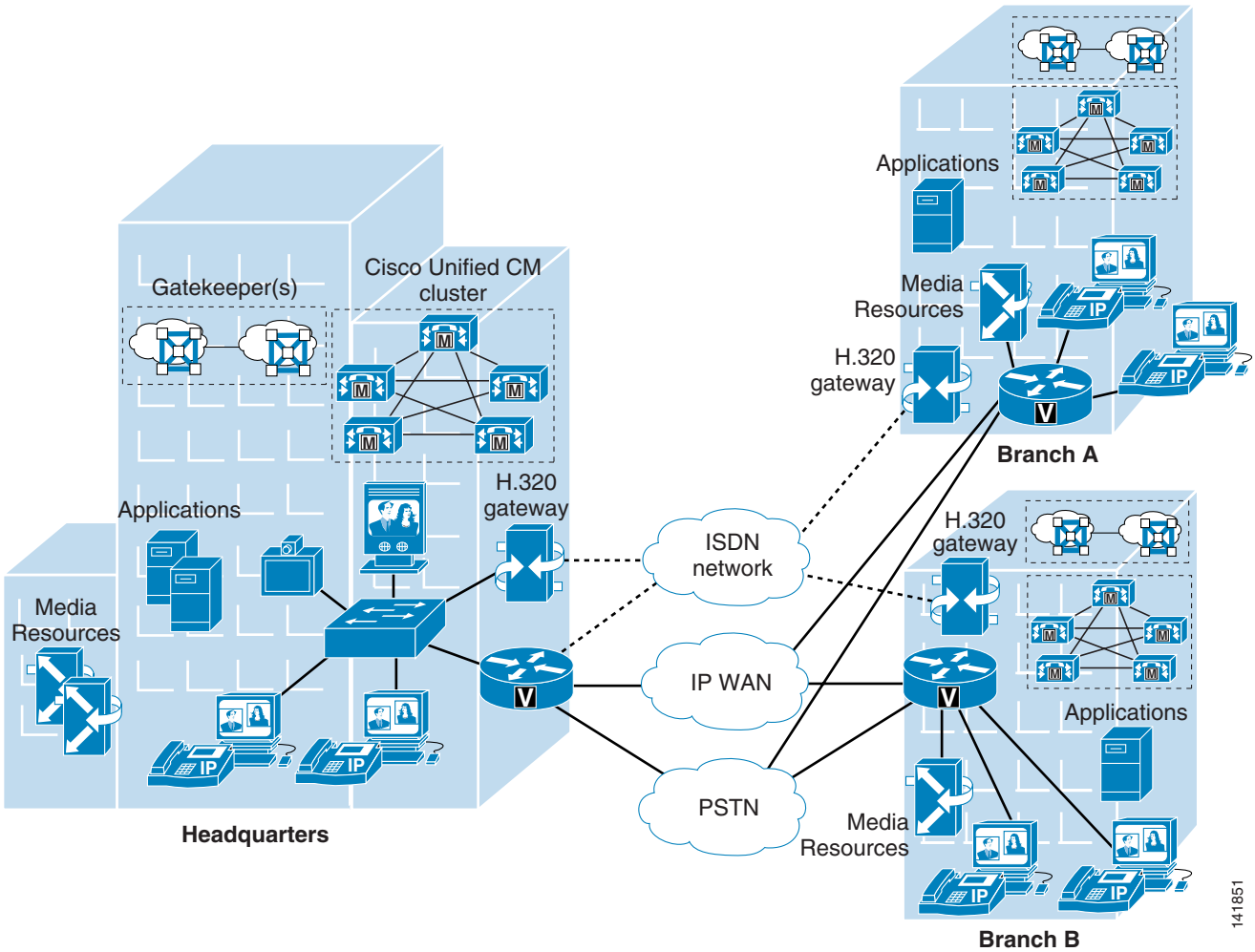
In addition to the general considerations listed for VoPSTN, the following design guidelines apply to the dial plan implementation method:

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

Multiple Sites 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 2-4 illustrates a typical distributed call processing deployment.

Figure 2-4 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 Unified Communications Manager Express (Unified CME)
  - Other IP PBX
- A centralized call processing site and all of its associated remote sites
- A legacy PBX with Voice over IP (VoIP) gateway

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

- Maximum of 30,000 configured and registered Skinny Client Control Protocol (SCCP) or Session Initiation Protocol (SIP) IP phones or SCCP video endpoints per cluster.
- Maximum of 1100 MGCP gateways or H.323 devices (gateways, MCUs, trunks, and clients) per Unified CM cluster.
- PSTN for all external calls.
- Digital signal processor (DSP) resources for conferencing, transcoding, and media termination point (MTP).
- Voicemail, unified messaging, and Cisco Unified Presence components.
- Capability to integrate with legacy private branch exchange (PBX) and voicemail systems.
- H.323 clients, MCUs, and H.323/H.320 gateways that require a gatekeeper to place calls must register with a Cisco IOS Gatekeeper (Cisco IOS Release 12.3(8)T or greater). Unified CM then uses an H.323 trunk to integrate with the gatekeeper and provide call routing and bandwidth management services for the H.323 devices registered to it. Multiple Cisco IOS Gatekeepers may be used to provide redundancy. Cisco IOS Gatekeepers may also be used to provide call routing and bandwidth management between the distributed Unified CM clusters. In most situations, Cisco recommends that each Unified CM cluster have its own set of endpoint gatekeepers and that a separate set of gatekeepers be used to manage the intercluster calls. It is possible in some circumstances to use the same set of gatekeepers for both functions, depending on the size of the network, complexity of the dial plan, and so forth. (For details, see Gatekeepers, page 17-21.)
- MCU resources are required in each cluster for multipoint video conferencing. Depending on conferencing requirements, these resources may be either SCCP or H.323, or both, and may all be located at the regional sites or may be distributed to the remote sites of each cluster if local conferencing resources are required.
- H.323/H.320 video gateways are needed to communicate with H.320 videoconferencing devices on the public ISDN network. These gateways may all be located at the regional sites or may be distributed to the remote sites of each cluster if local ISDN access is required.
- High-bandwidth audio (for example, G.711, G.722, or Cisco Wideband Audio) between devices in the same site, but low-bandwidth audio (for example, G.729 or G.728) between devices in different sites.
- High-bandwidth video (for example, 384 kbps or greater) between devices in the same site, but low-bandwidth video (for example, 128 kbps) between devices at different sites. The Cisco Unified Video Advantage Wideband Codec, operating at 7 Mbps, is recommended only for calls between devices at the same site. Note that the Cisco VT Camera Wideband Video Codec is not supported over intercluster trunks.
- Minimum of 768 kbps or greater WAN link speeds. Video is *not* recommended on WAN connections that operate at speeds lower than 768 kbps.

- Call admission control is provided by Unified CM locations for calls between sites controlled by the same Unified CM cluster, and by the Cisco IOS Gatekeeper for calls between Unified CM clusters (that is, intercluster trunks). Automated alternate routing (AAR) is also supported for both intra-cluster and inter-cluster video calls.

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 Single Site, page 2-2.)

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 Single Site, page 2-2, and Multiple Sites with Centralized Call Processing, page 2-4.)

Gatekeeper or Session Initiation Protocol (SIP) proxy servers are among the key elements in multisite deployments with distributed call processing. They each provide dial plan resolution, with the gatekeeper also providing call admission control. A gatekeeper is an H.323 device that provides call admission control and E.164 dial plan resolution.

The following best practices apply to the use of a gatekeeper:

- Use a Cisco IOS gatekeeper to provide call admission control into and out of each site.
- To provide high availability of the gatekeeper, use Hot Standby Router Protocol (HSRP) gatekeeper pairs, gatekeeper clustering, and alternate gatekeeper support. In addition, use multiple gatekeepers to provide redundancy within the network. (See Gatekeeper Design Considerations, page 8-17.)
- Size the platforms appropriately to ensure that performance and capacity requirements can be met.
- Use only one type of codec on the WAN because the H.323 specification does not allow for Layer 2, IP, User Data Protocol (UDP), or Real-time Transport Protocol (RTP) header overhead in the bandwidth request. (Header overhead is allowed only in the payload or encoded voice part of the packet.) Using one type of codec on the WAN simplifies capacity planning by eliminating the need to over-provision the IP WAN to allow for the worst-case scenario.
- Gatekeeper networks can scale to hundreds of sites, and the design is limited only by the WAN topology.

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

- For gatekeeper call admission control, see Call Admission Control, page 9-1.
- For gatekeeper scalability and redundancy, see Call Processing, page 8-1.
- For gatekeeper dial plan resolution, see Dial Plan, page 10-1.

SIP devices provide resolution of E.164 numbers as well as SIP uniform resource identifiers (URIs) to enable endpoints to place calls to each other. Unified CM supports the use of E.164 numbers only.

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

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

Call Processing Agents for the Distributed Call Processing Model

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

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

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

Table 2-3 lists recommended call processing agents.

Table 2-3 Recommended Call Processing Agents

Call Processing Agent	Recommended Size	Comments
Cisco Unified Communications Manager Express (Unified CME)	Up to 240 phones	• For small remote sites • Capacity depends on Cisco IOS platform
Cisco Unified Communications Manager Business Edition (Unified CMBE)	Up to 575 phones	• For small sites • Supports centralized call processing
Cisco Unified Communications Manager (Unified CM)	50 to 30,000 phones	• Small to large sites, depending on the size of the Unified CM cluster • Supports centralized or distributed call processing
Legacy PBX with VoIP gateway	Depends on PBX	• Number of IP WAN calls and functionality depend on the PBX-to-VoIP gateway protocol and the gateway platform

Clustering Over the IP WAN

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

Clustering over the WAN can support two types of deployments:

- Local Failover Deployment Model, page 2-23

  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 2-28

  Remote failover allows you to deploy primary and backup call processing servers split across the WAN. Using this type of deployment, you may have up to eight 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 up to eight 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 2-20, 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 6.0 servers should not exceed 20 msec, or 40 msec round-trip time (RTT). Beginning with Cisco Unified CM Release 6.1, the maximum one-way delay between two Unified CM servers can be up to 40 msec, or 80 msec round-trip time. Measuring the delay is covered in Delay Testing, page 2-21.

- 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. Because only eight servers may have the Cisco CallManager Service enabled in a cluster, there may be up to seven connections on each server. 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 port usage document at http://www.cisco.com/en/US/docs/voice_ip_comm/cucm/port/6_1/61plrev1.pdf.

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 40 msec for Unified CM 6.0, or 80 msec for Unified CM 6.1 and later releases. 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 ToS bit (IP Precedence) set to 3:

```
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]: 3
Set DF bit in IP header? [no]:
Validate reply data? [no]:
Data pattern [0xABCD]:
Loose, Strict, Record, Timestamp, Verbose[none]:
Sweep range of sizes [n]:
Type escape sequence to abort.
Sending 5, 100-byte ICMP Echos to 10.10.10.10, timeout is 2 seconds:
!!!!!
Success rate is 100 percent (5/5), round-trip min/avg/max = 1/2/4 ms
```

Error Rate

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

Troubleshooting

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

- IP phones, gateways, or other devices on a remote Unified CM server within the cluster might temporarily be unreachable.
- Calls might be disconnected or might fail during call setup.
- Users might experience longer than expected delays before hearing dial tone.
- Busy hour call completions (BHCC) might be low.
- The ICCS (SDL session) might be reset or disconnected.

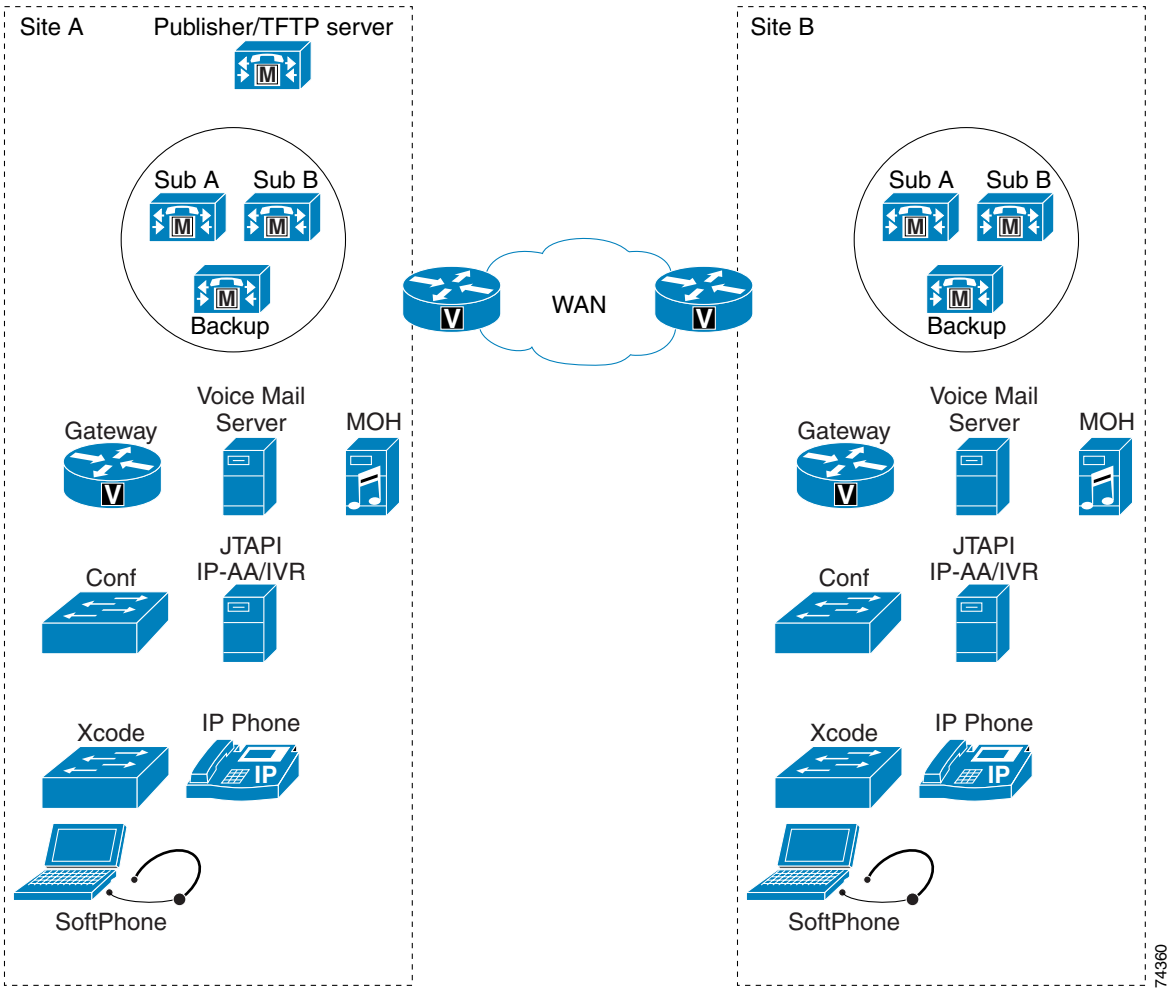
In summary, perform the following tasks to troubleshoot ICCS communication problems:

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

Local Failover Deployment Model

The local failover deployment model provides the most resilience for clustering over the WAN. Each of the sites in this model contains at least one primary Unified CM subscriber and one backup subscriber. This configuration can support up to four sites. The maximum number of phones and other devices will be dependant on the quantity and type of servers deployed. The maximum total number of IP phones for all sites is 30,000. (See Figure 2-5.)

Figure 2-5 Example of Local Failover Model



Observe the following guidelines when implementing the local failover model:

- Configure each site to contain at least one primary Unified CM subscriber and one backup subscriber.
- Configure Unified CM *groups* and *device pools* to allow devices within the site to register with only the servers at that site under all conditions.
- Cisco highly recommends that you replicate key services (TFTP, DNS, DHCP, LDAP, and IP Phone Services), all media resources (conference bridges and music on hold), and gateways at each site to provide the highest level of resiliency. You could also extend this practice to include a voicemail system at each site.

- Under a WAN failure condition, sites without access to the publisher database will lose some functionality. For example, system administration at the remote site will not be able to add, modify, or delete any part of the configuration. However, users can continue to access the user-facing features listed in the section on Unified CM Publisher, page 2-21.
- 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 40 msec for Unified CM 6.0 or 80 msec for Unified CM 6.1 and later releases.

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.

- Prior to Release 6.1 of Unified CM, 900 kbps of bandwidth was required for Intra-Cluster Communications Signaling (ICCS) traffic for 10,000 BHCA between sites that are clustered over the WAN. An additional 644 kbps of bandwidth was required for database and other intra-cluster communication for every remote server. Thus, a minimum of T1 bandwidth was required in deployments with clustering over the WAN to account for ICCS and additional database traffic between Unified CM servers.
- For Unified CM 6.1 and later releases, a minimum of 1.544 Mbps (T1) bandwidth is required for Intra-Cluster Communication Signaling (ICCS) for 10,000 busy hour call attempts (BHCA) between sites that are clustered over the WAN. This is a minimum bandwidth requirement for call control traffic, and it applies to deployments where directory numbers are not shared between sites that are clustered over the WAN. The following equation may be used as a guideline to calculate the bandwidth for more than 10,000 BHCA between non-shared directory numbers at a specific delay:

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

  This call control traffic is classified as priority traffic. Priority ICCS traffic is marked with IP Precedence 3 (DSCP 24 or PHB CS3).
- In addition to the bandwidth required for Intra-Cluster Communication Signaling (ICCS) traffic, a minimum of 1.544 Mbps (T1) bandwidth is required for database and other inter-server traffic for every subscriber server remote to the publisher in Unified CM 6.1 and later releases.

Example 2-1 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 msec 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 msec

  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 msec

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

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

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

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

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

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

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

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

- In prior versions of Unified CM, subscriber servers in the cluster use the publisher's database for read/write access, and they use their local database for read access only when the publisher's database cannot be reached. With Unified CM 6.*x*, subscriber servers in the cluster read their local database. Database modifications can occur in both the local database as well as the publisher database, depending on the type of changes. Informix Dynamic Server (IDS) database replication is used to synchronize the databases on the various servers in the cluster. Therefore, when recovering from failure conditions such as the loss of WAN connectivity for an extended period of time, the Unified CM databases must be synchronized with any changes that might have been made during the outage. This process happens automatically when database connectivity is restored to the publisher and other servers in the cluster. This process can take longer over low bandwidth and/or higher delay links. In rare scenarios, manual reset or repair of the database replication between servers in the cluster might be required. This is performed by using the commands such as **utils dbreplication repair all** and/or **utils dbreplication reset all** at the command line interface (CLI). Repair or reset of database replication using the CLI on remote subscribers over the WAN causes all Unified CM databases in the cluster to be re-synchronized, in which case additional bandwidth above 1.544 Mbps might be required. Lower bandwidths can take longer for database replication repair or reset to complete.

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

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

Unified CM Provisioning for Local Failover

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

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

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

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

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

Gateways for Local Failover

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

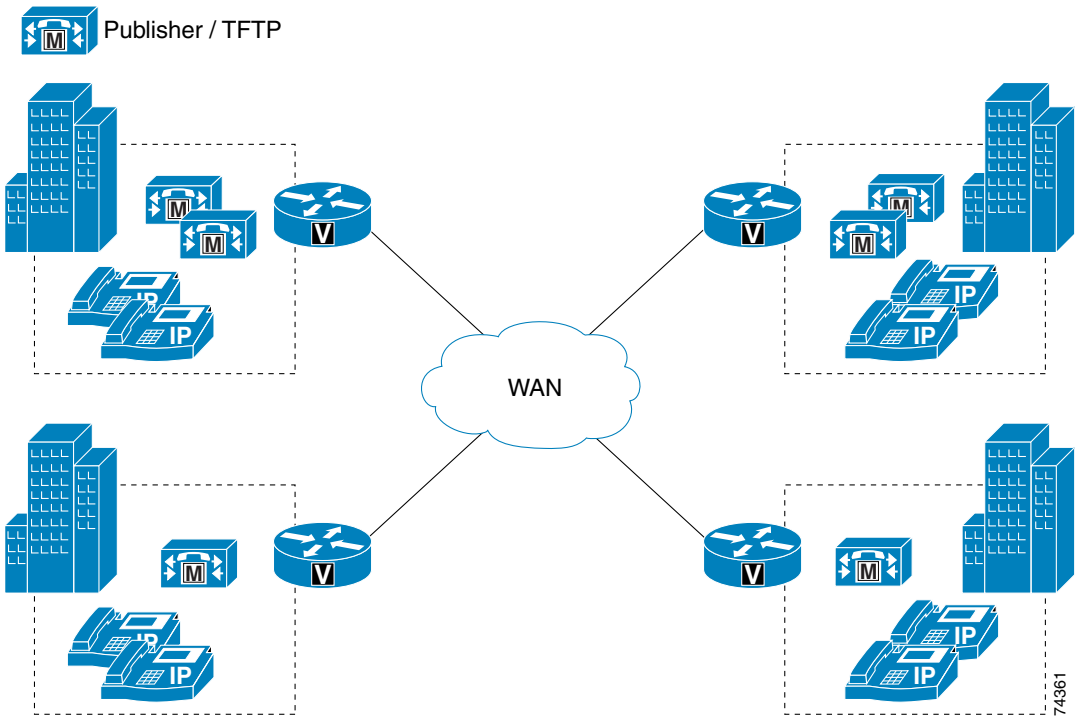
Music on Hold and Media Resources for Local Failover

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

Remote Failover Deployment Model

The remote failover deployment model provides flexibility for the placement of backup servers. Each of the sites contains at least one primary Unified CM subscriber and may or may not have a backup subscriber. This model allows for a deployment of up to eight sites, with IP phones and other devices normally registered to a local subscriber when using 1:1 redundancy and the 50/50 load balancing option described in the chapter on Call Processing, page 8-1. Backup subscribers are located across the WAN at one or more of the other sites. (See Figure 2-6.)

Figure 2-6 Remote Failover Model with Four Sites



When implementing the remote failover model, observe all guidelines for the local failover model (see Local Failover Deployment Model, page 2-23), 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.

- 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 3-50.

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.

Design Considerations for Section 508 Conformance

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

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

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

    Plug a TTY/TDD with an RJ-11 analog line option directly into a Cisco FXS port. Any FXS port will work, such as the one on the Cisco VG248, Catalyst 6000, Cisco ATA 188 module, or any other Cisco voice gateway with an FXS port. Cisco recommends this method of connection.
  - Acoustic coupling

    Place the IP phone handset into a coupling device on the TTY/TDD. Acoustic coupling is less reliable than an RJ-11 connection because the coupling device is generally more susceptible to transmission errors caused by ambient room noise and other factors.
- If stutter dial tone is required, use an analog phone in conjunction with an FXS port on the Cisco VG248 or ATA 188. In addition, most Cisco IP Phones support stutter dial tone, which is sometimes referred to as audible message waiting indication (AMWI). For a list of the specific Cisco IP Phone models that support this functionality, see the Endpoint Features Summary, page 21-39.