



CHAPTER 8

Call Processing

Last revised on: June 4, 2010

This chapter provides guidance for designing scalable and resilient call processing systems with Cisco Unified Communications Manager (Unified CM). This chapter also discusses how to choose the appropriate hardware and deployment scenario for Unified CM based on specific requirements that include the following:

- Scale — The number of users, locations, gateways, applications, and so forth
- Performance — The call rate
- Resilience — The amount of redundancy

This chapter focuses on the following topics:

- [Unified CM Cluster Guidelines, page 8-2](#)

This section discusses the minimum hardware requirements for Unified CM. This section also explains the various feature services that can be enabled on the Unified CM server and the purpose of each.

- [Unified CM Platform Capacity Planning, page 8-15](#)

This section provides guidelines for using the Cisco Unified CM Capacity Tool, which must be used when planning an IP Telephony deployment. The Cisco Unified CM Capacity Tool provides guidance on resources used on a Unified CM server, based on certain deployment requirements.

- [Computer Telephony Integration \(CTI\), page 8-18](#)

This section explains the Cisco Computer Telephony Integration (CTI) architecture and discusses CTI components and interfaces, CTI functionality, and CTI provisioning and capacity planning.

- [Gatekeeper Design Considerations, page 8-26](#)

This section explains how gatekeepers can be used in a Cisco Unified Communications deployment. Cisco Gatekeeper may also be paired with other standby gatekeepers or may be clustered for higher performance and resilience. Gatekeepers may also be used for call routing and call admission control.

- [Interoperability of Unified CM and Unified CM Express, page 8-36](#)

This section explains the H.323 and SIP integration between Cisco Unified CM and Cisco Unified Communications Manager Express (Unified CME) in a distributed call processing deployment.

What's New in This Chapter

Table 8-1 lists the topics that are new in this chapter or that have changed significantly from previous releases of this document.

Table 8-1 ***New or Changed Information Since the Previous Release of This Document***

New or Revised Topic	Described in:
Computer Telephony Integration (CTI)	Computer Telephony Integration (CTI), page 8-18
CTI and clustering over the WAN	CTI Applications and Clustering Over the WAN, page 8-20
CTI capacity limits	Unified CM Capacity Planning, page 8-22
NIC teaming	NIC Teaming for Network Fault Tolerance, page 8-4
Server capacity planning	Unified CM Platform Capacity Planning, page 8-15
The definitions of standard and high-availability servers have been revised.	Table 8-2
Trace file capacities	Unified CM Platform Capacity Planning, page 8-15
Video calls between multiple Unified CMEs in a distributed call processing environment	General Design Considerations for Unified CM and Unified CME Interoperability via SIP, page 8-40

Unified CM Cluster Guidelines

The Unified CM architecture enables a group of physical servers to work together as a single IP PBX system. This grouping of servers is known as a *cluster*. A cluster of Unified CM servers may be distributed across an IP network, within design limitations, allowing for spatial redundancy and, hence, resilience to be designed into the IP Communications system.

This section describes the various functions performed by the servers that form a Unified CM cluster, and it provides guidelines for deploying the servers in ways that achieve the desired scale, performance, and resilience.

Hardware Platforms

Unified CM clusters utilize various types of servers, depending on the scale, performance, and redundancy required. They range from non-redundant, single-processor servers to highly redundant, multi-processor units.

Table 8-2 lists the general types of servers you can use in a cluster, along with their main characteristics.

Table 8-2 *Types of Cisco Unified CM Servers*

Server Type	Cisco Server Model	Characteristics
Standard server (not high availability)	MCS 7815, MCS 7816, or equivalent	<ul style="list-style-type: none"> • Single processor • Single power supply • Non-RAID SATA hard disk
Standard server with RAID	MCS 7825 or equivalent	<ul style="list-style-type: none"> • Single processor • Single power supply • SATA controller with RAID 0/1 support
Standard server with RAID, for Cisco Unified Communications Manager Business Edition (Unified CMBE)	MCS 7828 ¹	<ul style="list-style-type: none"> • Single processor • Single power supply • SATA controller with RAID 0/1 support
High-availability server	MCS 7835, MCS 7845, or equivalent	<ul style="list-style-type: none"> • Multiple processors • Multiple power supplies • Multiple Serial Attached SCSI (SAS) drives with RAID 1

1. The Cisco MCS 7828 supports only Unified CMBE.

Cisco Unified CM is supported on specific Cisco MCS 7815, 7816, 7825, 7835, and 7845 servers or on customer-provided HP and IBM servers that have been verified by Cisco to meet the following minimum requirements:

- Processor speed must be 2.0 GHz or greater
- Physical memory size must be 2 GB or greater
- Physical hard disk size must be 72 GB or larger

For a complete list of currently supported hardware configurations, refer to the documentation available at

<http://www.cisco.com/go/swonly>

Servers should be deployed in an environment that provides high availability, not only for the IP network but also for power and cooling. Servers should be powered from an uninterruptible power supply (UPS) if building power does not have the required availability. Servers with dual power supplies could also be plugged into two different power sources to avoid the failure of one power circuit causing the server to fail.

Connectivity to the IP network should also ensure maximum performance and availability. The Unified CM servers should be connected to the Ethernet at 100 Mbps full-duplex. If 100 Mbps is not available on smaller deployments, then use 10 Mbps full-duplex. Many servers also include the capability of using Gigabit Ethernet, which is also an option. Ensure that servers are connected to the network using full-duplex, which can be achieved with 10 Mbps and 100 Mbps by hard-coding the switch port and the server NIC. For 1000 Mbps, Cisco recommends using Auto/Auto for speed and duplex configuration on both the NIC and the switch port. The default is Auto/Auto, and this setting is also the default following an upgrade from a previous Unified CM release.

**Note**

A mismatch will occur if either the server port or the Ethernet switch port is left in Auto mode and the other port is configured manually. The best practice is to configure both the server port and the Ethernet switch port manually, with the exception of Gigabit Ethernet ports which should be set to Auto/Auto.

NIC Teaming for Network Fault Tolerance

The NIC teaming feature allows a server to be connected to the Ethernet via two NICs and, therefore, two cables. NIC teaming prevents network downtime by transferring the workload from the failed port to the working port. NIC teaming cannot be used for load balancing or increasing the interface speed.

Hewlett-Packard (HP) and IBM server platforms with dual Ethernet network interface cards can support NIC teaming for Network Fault Tolerance.

**Note**

The Cisco MCS 7815 platform (or HP or IBM equivalent) has only a single network interface port and therefore cannot perform NIC teaming.

General Clustering Guidelines

The following guidelines apply to all Unified CM clusters:

**Note**

A cluster may contain a mix of server platforms, but all servers in the cluster must run the same Unified CM software release.

- Under normal circumstances, place all members of the cluster within the same LAN or MAN. Cisco does not recommend placing all members of a cluster on the same VLAN or switch.
- For redundancy, the members of the cluster should be deployed in the following manner to minimize the impact of any failures in the infrastructure or building:
 - Different access switches connected to the same distribution or core switch
 - Different access switches attached to different distribution or core switches
 - Different buildings within the same LAN or MAN
- If the cluster spans an IP WAN, follow the guidelines for clustering over an IP WAN as specified in the section on [Clustering Over the IP WAN, page 2-21](#).

Unified CM Cluster Services

Within a Unified CM cluster, there are servers that provide unique services. Each of these services can coexist with others on the same physical server. For example, in a small system it is possible to have a single server be a database publisher, backup subscriber, music on hold (MoH) server, TFTP server, CTI Manager, and Conference Bridge. As the scale and performance requirements of the cluster increase, many of these services should be moved to a single, dedicated physical server.

A cluster may contain as many as 20 servers, of which a maximum of eight may run the Cisco CallManager Service that provides call processing. The other servers may be configured as a dedicated database publisher, dedicated Trivial File Transfer Protocol (TFTP) server, or music on hold (MoH) servers. Media streaming applications (conference bridge or media termination point) may also be enabled on a separate server that registers with the cluster.

**Note**

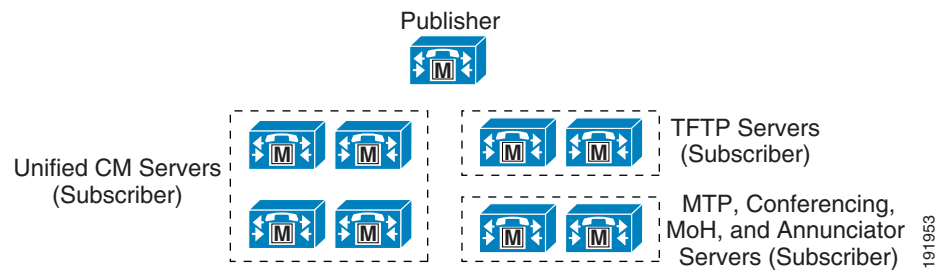
While Cisco recommends using the same server model for all servers in a cluster, mixing server models within a cluster is supported provided that all of the individual hardware versions are supported and that all servers are running the same version of Unified CM. However, differences in capacity between various server models within a cluster must be considered because the overall cluster capacity might ultimately be dictated by the capacity of the smallest server within the cluster. Mixing servers from different vendors within a cluster is also supported and does not have any adverse capacity implications, provided that all servers in the cluster are the same model type. For information on call processing capacity, see the section on [Unified CM Platform Capacity Planning, page 8-15](#).

When deploying a cluster with Cisco MCS 7816 or equivalent servers, there is a maximum limit of two servers in a cluster: one as the publisher, TFTP server, and backup call processing server, and the other as the primary call processing server. A maximum of 500 phones is supported in this configuration on a Cisco MCS 7816 or equivalent server. When deploying a two-server cluster with higher-capacity servers, Cisco recommends that you do not exceed 1250 users in the cluster. Above 1250 users, a dedicated publisher and separate servers for primary and secondary call processing services is recommended, thus increasing the number of servers in a cluster.

It is also possible to deploy a single-server cluster with an MCS 7825 or greater servers. With an MCS 7825 or equivalent server, the limit is 500 users; with a higher-availability server, the single-server cluster should not exceed 1000 users. In a single-server configuration, there is no redundancy unless Survivable Remote Site Telephony (SRST) is also deployed to provide service during periods when the Unified CM is not available. Cisco does not recommend a single-server deployment for production environments. The load balancing option is not available when the publisher is a backup call processing subscriber.

[Figure 8-1](#) illustrates a typical Unified CM cluster.

Figure 8-1 Typical Unified CM Cluster



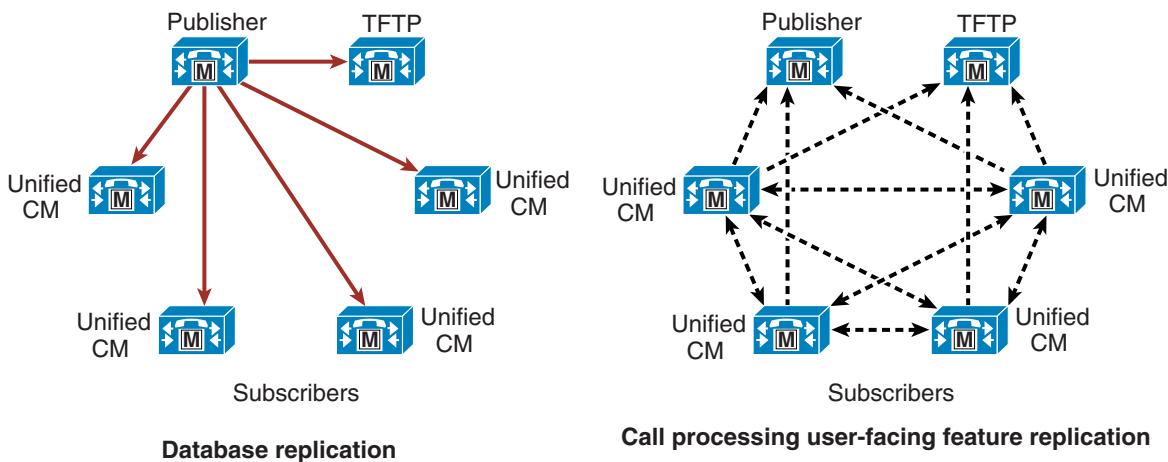
Intracluster Communications

There are two primary kinds of intracluster communications, or communications within a Unified CM cluster (see [Figure 8-2](#) and [Figure 8-3](#).) The first is a mechanism for distributing the database that contains all the device configuration information (see “Database replication” in [Figure 8-2](#)). The configuration database is stored on a publisher server, and a read-only copy is replicated to the subscriber members of the cluster. Most of the database changes are made on the publisher and are then communicated to the subscriber databases, thus ensuring that the configuration is consistent across the members of the cluster and facilitating spatial redundancy of the database.

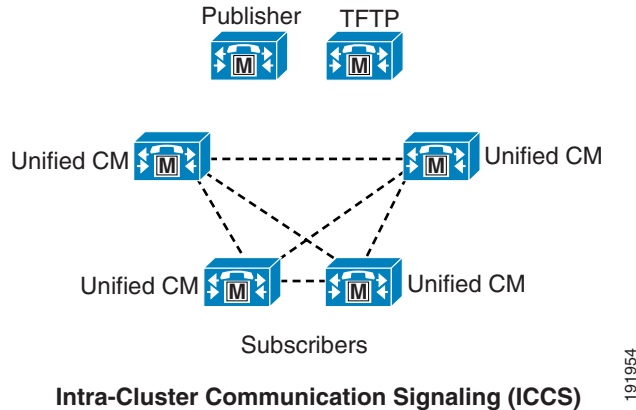
Database modifications for user-facing call processing features are made on the subscriber servers to which the IP phones are registered. The subscriber servers then replicate these database modifications to all the other servers in the cluster, thus providing redundancy for the user-facing features. (See “Call processing user-facing feature replication” in [Figure 8-2](#).) These features include:

- Call Forward All (CFA)
- Message waiting indicator (MWI)
- Privacy Enable/Disable
- Extension Mobility login/logout
- Hunt Group login/logout
- Device Mobility
- Certificate Authority Proxy Function (CAPF) status for end users and applications users
- Credential hacking and authentication

Figure 8-2 Replication of the Database and User-Facing Features

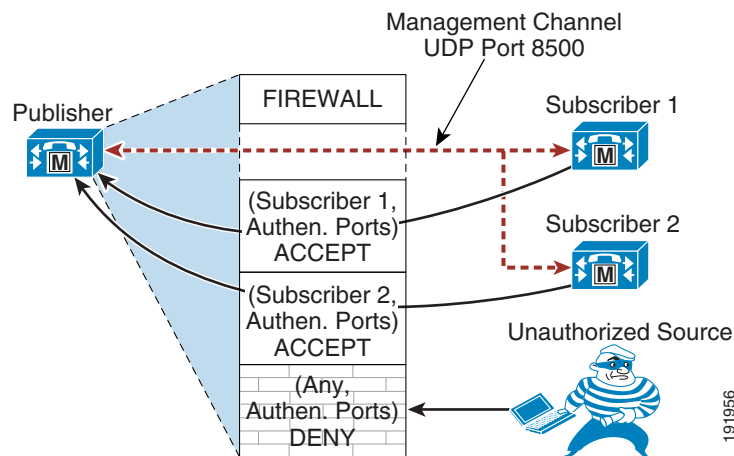


The second type of intracluster communication, called Intra-Cluster Communication Signaling (ICCS), involves the propagation and replication of run-time data such as registration of devices, locations bandwidth, and shared media resources (see [Figure 8-3](#)). This information is shared across all members of a cluster running the Cisco CallManager Service, and it ensures the optimum routing of calls between members of the cluster and associated gateways.

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

Intracuster Security

Each server in a Unified CM cluster runs an internal dynamic firewall. The application ports on Unified CM are protected using source IP filtering. The dynamic firewall opens these applications ports only to authenticated or trusted servers. (See [Figure 8-4.](#))

Figure 8-4 Intracuster Security

This security mechanism is applicable only between servers in a single Unified CM cluster. Unified CM subscribers are authenticated in a cluster before they can access the publisher's database. The Intra-cluster communication and database replication take place only between authenticated servers. During the installation process, a subscriber is authenticated to the publisher using a pre-shared key authentication mechanism. The authentication process involves the following steps:

1. Install the publisher server using a security password.
2. Configure the subscriber server on the publisher by using Unified CM Administration.
3. Install the subscriber server using the same security password used during publisher server installation.

4. After the subscriber is installed, the server attempts to establish connection to the publisher on a management channel using UDP 8500. The subscriber sends all the credentials to the publisher, such as hostname, IP address, and so forth. The credentials are authenticated using the security password used during the install process.
5. The publisher verifies the subscriber's credentials using its own security password.
6. The publisher adds the subscriber as a trusted source to its dynamic firewall table if the information is valid. The subscriber is allowed access to the database.
7. The subscriber gets a list of other subscriber servers from the publisher. All the subscribers establish a management channel with each other, thus creating a mesh topology.

Publisher

The publisher is a required server in all clusters, and there can currently be only one per cluster. This server is the first to be installed and provides the database services to all other members in the cluster. The publisher server is the only server that has full read and write access to the configuration database. On larger systems with more than 1250 users, Cisco recommends a dedicated publisher to prevent administrative operations from affecting the telephony services. A dedicated publisher does not have any call processing services or TFTP services running on the server. Other servers run the TFTP and Unified CM services.

Subscriber servers in the cluster attempt to use the local database when initializing. This reduces the Cisco CallManager Service initialization time. In prior versions of Unified CM, subscriber servers in the cluster attempted to use the publisher's database when initializing. If the publisher was not available, they would use the local read-only copy from their hard drives.

The choice of hardware platform for the publisher is based on the scale and performance of the cluster. Cisco recommends that the publisher have the same performance capability as the call processing subscribers. Ideally the publisher should also be a high-availability server to minimize the impact of a hardware failure.

Call Processing Subscriber

When installing the Unified CM software, you can define two types of servers, publisher and subscriber. These terms are used to define the database relationship during installation. Once the software is installed, only the database and network services are enabled. All subscribers will subscribe to the publisher to obtain a copy of the database information.

A call processing subscriber is a server that has the Cisco CallManager Service enabled. A single server license is required to enable this service on a subscriber. The Cisco CallManager Service cannot be enabled on a server if the publisher is not available because the publisher acts as a licensing server and distributes the licenses needed to activate the Cisco CallManager Service. Once this service is enabled, the server is able to perform call processing functions. Devices such as phones, gateways, and media resources can register and make calls only to servers with this service enabled. Unified CM supports up to eight servers in a cluster with the Cisco CallManager Service enabled.

Depending on the redundancy scheme chosen (see [Call Processing Redundancy, page 8-9](#)), the call processing subscriber will be either a primary (active) subscriber or a backup (standby) subscriber. In the load-balancing option, the subscriber can be both a primary and backup subscriber. When planning the design of a cluster, you should generally dedicate the call processing subscribers to this function. In larger-scale or higher-performance clusters, the call processing service should not be enabled on the publisher and TFTP server. Call processing subscribers normally operate in either dedicated pairs or

shared pairs, depending on the redundancy scheme adopted. One-to-one redundancy uses dedicated pairs, while two-to-one redundancy uses two pairs of servers that share one server from each pair (the backup server).

The choice of hardware platform depends on the scale, performance, redundancy, and cost of the servers. Scale and performance are covered in the section on [Unified CM Platform Capacity Planning, page 8-15](#), and redundancy is covered in the section on [Call Processing Redundancy, page 8-9](#).

Call Processing Redundancy

You can choose from the following Unified CM redundancy configurations:

- Two to one (2:1) — For every two primary subscribers, there is one shared secondary or backup subscriber.
- One to one (1:1) — For every primary subscriber, there is a secondary or backup subscriber.

The 1:1 redundancy scheme allows upgrades with only the failover periods impacting the cluster. The failover mechanism has been enhanced so that you can achieve failover rates for Skinny Client Control Protocol (SCCP) IP phones of approximately 125 registrations per second. The failover mechanism for Session Initiation Protocol (SIP) phones is approximately 40 registrations per second.

A cluster can be upgraded without impacting the services. Two different versions (releases) of Unified CM may be on the same server, one in the active partition and the other in the inactive partition. All services and devices use the Unified CM version in the active partition for all Unified CM functionality. During the upgrade process, the cluster operations continue using its current release of Unified CM in the active partition, while the upgrade version gets installed in the inactive partition. Once the upgrade process is completed, the servers can be rebooted to switch the inactive partition to the active partition, thus running the new version of Unified CM.

Upgrading Unified CM

The 1:1 redundancy scheme enables you to upgrade the cluster using the following method:

-
- | | |
|---------------|--|
| Step 1 | Install the new version of Unified CM on the publisher. Do not reboot. |
| Step 2 | Install the new version of Unified CM on all subscribers simultaneously. Do not reboot. |
| Step 3 | Reboot only the publisher. Switch to the new version of Unified CM and allow some time for the database to initialize. |
| Step 4 | Reboot the TFTP server(s) one at a time. Switch to the new version of Unified CM and wait for the configuration files to be rebuilt before upgrading any further servers in the cluster. |
| Step 5 | Reboot the dedicated music on hold (MoH) server(s) one at a time. Switch to the new version of Unified CM. |
| Step 6 | Reboot the backup subscriber(s) one at a time. Switch to the new version of Unified CM. This step might impact some users if 50/50 load balancing is implemented. |
| Step 7 | Fail-over the devices from the primary subscribers to their backups. |
| Step 8 | Reboot the primary subscriber(s) one at a time. Switch to the new version of Unified CM. |
-

With this upgrade method, there is no period (except for the failover period) when devices are registered to subscriber servers that are running different versions of the Unified CM software.

The 2:1 redundancy scheme allows for fewer servers in a cluster, but it can potentially result in an outage during upgrades.

**Note**

You must use 1:1 redundancy when more than 7,500 IP phones are registered on the two primary subscribers because there cannot be more than 7,500 backup registrations on a single backup subscriber.

**Note**

Before you do an upgrade, Cisco recommends that you back up the Unified CM and Call Detail Record (CDR) database to an external network directory using the Disaster Recovery Framework. This practice will prevent any loss of data if the upgrade fails.

Call Processing Subscriber Redundancy

The following figures illustrate typical cluster configurations to provide call processing redundancy with Unified CM.

Figure 8-5 Basic Redundancy Schemes

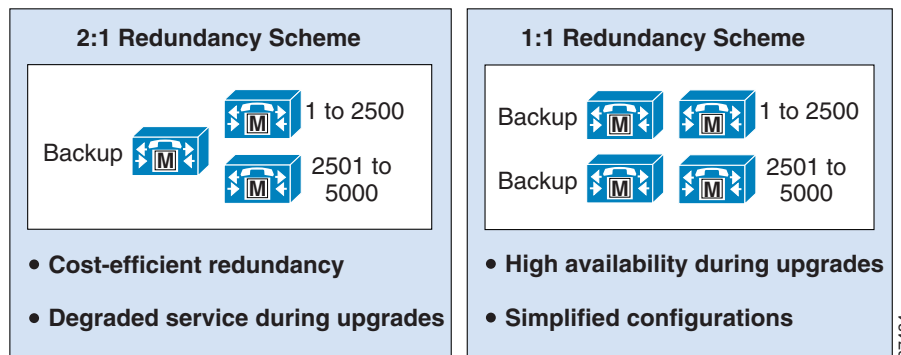
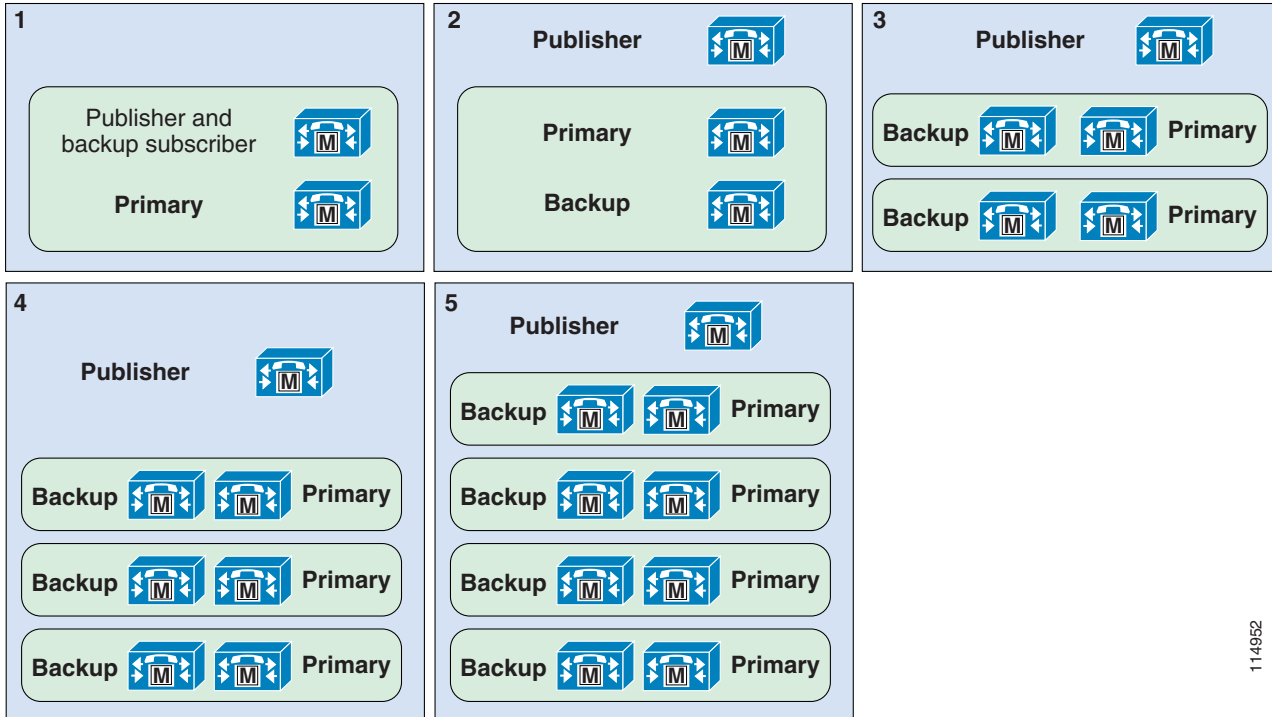


Figure 8-5 illustrates the two basic redundancy schemes available. In each case the backup server must be capable of handling the capacity of at least a single primary call processing server failure. In the 2:1 redundancy scheme, the backup might have to be capable of handling the failure of a single call processing server or potentially both primary call processing servers, depending on the requirements of a particular deployment. Sizing the capacity of the servers and choosing the hardware platforms is covered in the section on [Unified CM Platform Capacity Planning, page 8-15](#).

87424

Figure 8-6 1:1 Redundancy Configuration Options

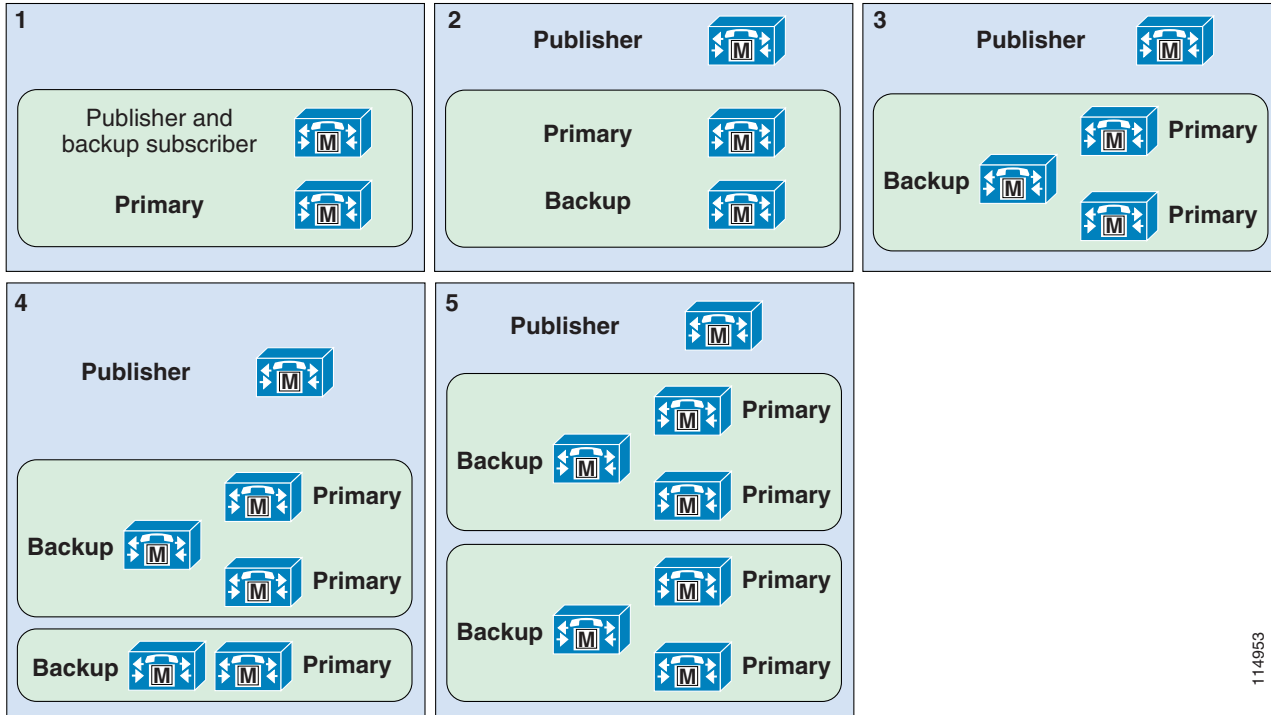


114952

In Figure 8-6 the five options shown all indicate 1:1 redundancy. Option 1 is used for clusters supporting less than 1250 users. Options 2 through 5 illustrate increasingly scalable clusters. The exact scale will depend on the hardware platforms chosen or required.

Note that the illustrations show only publisher and call processing subscribers.

Figure 8-7 2:1 Redundancy Configuration Options



114953

**Note**

It is possible to define up to three call processing subscribers per Unified CM group. Adding a tertiary subscriber for additional backup extends the above redundancy schemes to 2:1:1 or 1:1:1 redundancy. However, with the exception of using tertiary subscriber servers in deployments with clustering over the WAN (see [Remote Failover Deployment Model, page 2-30](#)), tertiary subscriber redundancy is not recommended for endpoint devices located in remote sites because failover to SRST will be further delayed if the endpoint must check for connectivity to a tertiary subscriber.

Load Balancing

Normally a backup server has no devices registered to it unless its primary is unavailable. This model allows for:

- Easier troubleshooting — Because all call processing takes place on primary servers, obtaining traces and alert notifications becomes easier.
- Less configuration — Because all the devices are registered to the primary server, the need to define additional Unified CM redundancy groups or device pools for the various devices can be reduced by 50%.

The 1:1 redundancy scheme enables you to balance distribution of the devices over the primary and backup server pairs. With load balancing, you can move up to half of the device load from the primary to the secondary subscriber by using the Unified CM redundancy groups and device pool settings. This model allows for:

- Load sharing — The call processing load is distributed on multiple servers, which can provide faster response time.

- Faster failover and failback — Because all devices (such as IP phones, CTI ports, gateways, trunks, voicemail ports, and so forth) are distributed across all active subscribers, only some of the devices fail over to the secondary subscriber if the primary subscriber fails. In this way, you can reduce by 50% the impact of any server becoming unavailable.

To plan for 50/50 load balancing, calculate the capacity of a cluster without load balancing, and then distribute the load across the primary and backup subscribers based on devices and call volume. To allow for failure of the primary or the backup server, the total load on the primary and secondary subscribers should not exceed that of a single subscriber server.

TFTP Server

The TFTP server performs two main functions:

- The serving of files for services such as MoH, configuration files for devices such as phones and gateways, binary files for the upgrade of phones as well as some gateways, and various security files.
- Generation of configuration and security files. Most files generated by the Cisco TFTP service are signed and in some cases encrypted before being available for download.

The TFTP service can be enabled on any server in the cluster. However, in a cluster with more than 1250 users, other services might be impacted by configuration changes that can cause the TFTP service to regenerate configuration files. Therefore, Cisco recommends that you dedicate a specific server to the TFTP service in a cluster with more than 1250 users, with Extension Mobility, or with other features that cause configuration changes.

The TFTP server is used by phones and MGCP gateways to obtain configuration information. There is no restriction on the number of servers that can have TFTP service enabled, however Cisco recommends deploying 2 TFTP servers for a large cluster, thus providing redundancy for TFTP service. More than 2 TFTP servers can be deployed in a cluster, but this can result in an extended period for rebuilding of all TFTP files on all TFTP servers. When configuring the TFTP options using DHCP or statically, you can normally define an IP address array (more than one IP address) for a TFTP server. Therefore, you can assign half of the devices to use TFTP server A as the primary and TFTP server B as the backup, and the other half to use TFTP server B as the primary and TFTP server A as the backup. To improve performance on dedicated TFTP servers, you can set service parameters to increase the number of simultaneous TFTP sessions allowed on the server.

When upgrading a Unified CM cluster, Cisco highly recommends that you upgrade the TFTP servers after the publisher and before any other server, also allowing additional time following the upgrade for the TFTP server to rebuild all the configuration files. Either use the typical Cisco TFTP - BuildDuration time or use the real-time monitoring tool to monitor the Cisco TFTP - DeviceBuildCount until it stops incrementing. This upgrade order ensures that any new binaries and configuration changes are available before the upgrade of other services in the cluster. If you are manually adding a specific binary or firmware load for a phone or gateway, be sure to copy the file to each TFTP server in the cluster.

Cisco recommends that you use the same hardware platform for the TFTP servers as used for the call processing subscribers.

CTI Manager

CTI Manager is required in a cluster for applications that use TAPI or JTAPI Computer Telephony Integration (CTI). The CTI Manager acts as a broker between the CTI application and the Cisco CallManager Service. It provides authentication of the application and enables control or monitoring of authorized devices. The CTI application communicates with a primary CTI Manager and, in the event of a failure, will switch to a backup CTI Manager. The CTI Manager should be enabled only on call

processing subscribers, thus allowing for a maximum of eight CTI Managers in a cluster. Cisco recommends that you load-balance CTI applications across the various CTI Managers in the cluster to provide maximum resilience, performance, and redundancy.

Generally, it is good practice to associate devices that will be controlled or monitored by an application with the same server pair used for the CTI Manager. For example, an interactive voice response (IVR) application requires four CTI ports. They would be provisioned as follows, assuming the use of 1:1 redundancy and 50/50 load balancing:

- Two CTI Ports would have a Unified CM redundancy group of server A as the primary and server B as the backup (or secondary). The other two ports would have a Unified CM redundancy group of server B as the primary and server A as the backup.
- The IVR application would be configured to use the CTI Manager on server A as the primary and server B as the backup.

The above example allows for redundancy in case of failure of the CTI Manager on server A and also allows for the IVR call load to be spread across two servers. This approach also minimizes the impact of a Unified CM server failure.

For more details on CTI Manager, see [Computer Telephony Integration \(CTI\)](#), page 8-18.

IP Voice Media Streaming Application

Media resources such as conferencing and music on hold may be provided by IP Voice Media Streaming Application services running on the same physical server as the Cisco CallManager Service.

Media resources include:

- Music on Hold (MoH) — Provides multicast or unicast music to devices that are placed on hold or temporary hold, transferred, or added to a conference. (See [Music on Hold](#), page 7-1.)
- Annunciator service — Provides announcements in place of tones to indicate incorrectly dialed numbers or call routing unavailability. (See [Annunciator](#), page 6-25.)
- Conference bridges — Provide software-based conferencing for ad-hoc and meet-me conferences. (See [Audio Conferencing](#), page 6-10.)
- Media termination point (MTP) services — Provide features for H.323 clients, H.323 trunks, and Session Initiation Protocol (SIP) trunks. (See [Media Termination Point \(MTP\)](#), page 6-17.)

Because of the additional processing and network requirements for media, it is essential to follow all guidelines for running media resources within a cluster. Generally, Cisco recommends non-dedicated servers for multicast MoH and the annunciator, but dedicated media resource subscribers as shown in [Figure 8-1](#) are recommended for unicast MoH as well as large-scale software-based conferencing and MTP unless those services are within the design guidelines detailed in the chapters on [Media Resources](#), page 6-1, and [Music on Hold](#), page 7-1.

Voice Activity Detection

Cisco also recommends that you leave voice activity detection (VAD) disabled within the cluster. VAD is disabled by default in the Unified CM service parameters, and you should disable it on H.323 and SIP dial peers by using the **no vad** command.

Unified CM Applications

Various types of applications can be enabled on Unified CM, such as Cisco Unified CM Assistant, Extension Mobility, and WebDialer. For detailed design guidance on these applications, see the chapter on [Cisco Unified CM Applications, page 24-1](#).

Unified CM Platform Capacity Planning

Many types of devices can register with Unified CM; for example, IP phones, voicemail ports, CTI (TAPI or JTAPI) devices, gateways, and DSP resources such as transcoding and conferencing. Each of these devices requires resources from the server platform with which it is registered. The required resources can include memory, processor usage, and disk I/O. Each device then consumes additional server resources during transactions, which are normally in the form of calls. For example, a device that makes only 6 calls per hour consumes fewer resources than a device making 12 calls per hour.

The recommendations provided in this section are based on calculations made using the Unified CM Capacity Tool, with default trace levels and CDRs enabled. Higher levels of performance can be achieved by disabling, reducing, or reconfiguring other functions that are not directly related to processing calls. Increasing some of these functions can also have an impact on the call processing capabilities of the system. These functions include tracing, call detail recording, highly complex dial plans, and other services that are co-resident on the server. Highly complex dial plans can include multiple line appearances as well as large numbers of partitions, calling search spaces, route patterns, translations, route groups, hunt groups, pickup groups, route lists, extensive use of call forwarding, co-resident services, and other co-resident applications. All of these functions can consume additional resources within the Unified CM server.

To improve system performance, the following techniques can provide useful options:

- Install additional certified memory in the server, up to the maximum supported for the particular platform. Doubling the RAM in MCS 7825 and MCS 7835 or equivalent servers is recommended in large configurations for that server class. Verification using the Cisco Real Time Monitoring Tool (RTMT) will indicate if this memory upgrade is required. As the server approaches maximum utilization of physical memory, the operating system will start to swap to disk. This swapping is a sign that additional physical memory should be installed.
- A Unified CM cluster with a very large dial plan containing many gateways, route patterns, translation patterns, and partitions, can take an extended amount of time to initialize when the Cisco CallManager Service is first started. If the system does not initialize within the default time, you can modify the system initialization timer (a Unified CM service parameter) to allow additional time for the configuration to initialize. For details on the system initialization time, refer to the online help for Service Parameters in Unified CM Administration.

The following guidelines apply to Cisco Unified CM:

- Within a cluster, a maximum of 8 servers can be enabled with the Cisco CallManager Service. Other servers may be used for more dedicated functions such as TFTP, publisher, music on hold, and so forth.
- Each cluster can support a maximum of 30,000 unsecured SCCP or SIP phones.
- Each cluster can support a maximum of 27,000 secured SCCP or SIP phones.
- You can configure a maximum of 500 locations on a Unified CM cluster consisting of MCS 7825 or MCS 7835 servers.
- A cluster consisting of MCS 7825 or MCS 7835 servers can support a maximum of 600 H.323 devices (gateways, trunks, and clients), digital MGCP devices, and SIP trunks.

- You can configure a maximum of 2000 locations on a Unified CM cluster consisting of MCS 7845 servers. (See [Unified CM Support for Locations and Regions, page 8-16.](#))
- A cluster consisting of MCS 7845 servers can support a maximum of 2100 H.323 devices (gateways, trunks, and clients), digital MGCP devices, and SIP trunks. (See [Unified CM Support for Gateways and Trunks, page 8-17.](#))
- The maximum recommended trace setting for Unified CM is 2,000 files of 2 MB for both System Diagnostic Interface (SDI) and Signaling Description Layer (SDL) traces, for a total of 4,000 files. Each process has a setting for maximum number of files, and each process is allowed 2,000 files for SDL and 2,000 files SDI. Trace settings for all other components must be configured within the limit of 126 MB (for example, 63 files of 2 MB each). These are suggested upper limits. Unless specific troubleshooting under high call rates requires increasing the maximum file setting, the default settings are sufficient for collecting sufficient traces in most circumstances.

The maximum number of users Unified CM can support depends on the server platform, as indicated in [Table 8-3.](#)

Table 8-3 Maximum Number of Devices per Server Platform

Server Platform Characteristics	Maximum Users per Server ¹	High-Availability Server ²	High-Performance Server
Cisco MCS 7845 (All supported models)	7500	Yes	Yes
Cisco MCS 7835 (All supported models)	2500	Yes	No
Cisco MCS 7825 (All supported models)	1000	No	No
Cisco MCS 7815 or MCS 7816 (All supported models) ³	500 ⁴	No	No

1. A platform that is not a high-availability server can support a maximum of 500 IP Phones in a non-redundant installation.
2. A high-availability server supports redundancy for both the power supplies and the hard disks.
3. MCS 7815 and MCS 7816 servers support only 1+1 redundancy (maximum of 2 servers) and cannot be a member of a cluster containing other servers.
4. The MCS 7815 server supports a maximum of 300 users.

For the latest information on supported platforms, third-party platforms, and specific hardware configurations, refer to the online documentation at

<http://www.cisco.com/go/swonly>

Unified CM Support for Locations and Regions

Cisco Unified Communications Manager 7.1(2) and later releases support 2000 locations and 2000 regions with Cisco MCS-7845 servers. To deploy up to 2000 locations and regions, you must configure the following service parameters in the **Clusterwide Parameters > System > Location and Region** and **Clusterwide Parameters > System > RSVP** configuration menus:

- Intraregion Audio Codec Default
- Interregion Audio Codec Default
- Intraregion Video Call Bandwidth Default
- Interregion Video Call Bandwidth Default
- Default inter-location RSVP Policy

When adding regions, you should then select **Use System Default** for the Audio Codec and Video Call Bandwidth values. If you are using RSVP call admission control, you should also select **Use System Default** for the RSVP Setting parameter.

Changing these values for individual regions and locations from the default has an impact on server initialization and publisher upgrade times. Hence, with a total of 2000 regions and 2000 locations, you can modify up to 200 of them to use non-default values. With a total of 1000 or fewer regions and locations, you can modify up to 500 of them to use non-default values. [Table 8-4](#) summarizes these limits.

Table 8-4 Number of Allowed Non-Default Regions and Locations

Number of non-default regions and locations	Maximum number of regions	Maximum number of locations
0 to 200	2000	2000
200 to 500	1000	1000



Note

The audio codec value is used by both voice calls and fax calls. If you plan to use G.729 as the interregion codec value, use T.38 Fax Relay for fax calls. If you plan to use fax pass-through over the WAN, change the default Interregion Audio Codec value to G.711, or else add a region for fax machines to each location with a non-default codec value of G.711 (subject to the limits in [Table 8-4](#)).



Note

Irrespective of the MCS model you are using, your Cisco Partner or Cisco Systems Engineer should always use the Cisco Unified Communications Sizing Tool (<http://tools.cisco.com/cucst>) to validate all designs that incorporate a large number of remote sites, because there are many interdependent variables that can affect Unified CM cluster scalability (such as regions, locations, gateways, media resources, and so forth). Use the Sizing Tool to accurately determine the number of servers or clusters required to meet your design criteria.

Unified CM Support for Gateways and Trunks

Cisco Unified Communications Manager 7.1(2) and later releases support 2100 gateways and trunks (that is, the total number of H.323 gateways, H323 trunks, digital MGCP devices, and SIP trunks) with Cisco MCS-7845 servers.

As you increase the numbers of active gateways, trunks, and media resources within a cluster, it is important to distribute the registration of these devices evenly across all call processing servers to avoid overloading the CPU of one or more servers in the cluster.



Note

Irrespective of the MCS model you are using, your Cisco Partner or Cisco Systems Engineer should always use the Cisco Unified Communications Sizing Tool (<http://tools.cisco.com/cucst>) to validate all designs that incorporate a large number of gateways and trunks, because there are many interdependent variables that can affect Unified CM cluster scalability (such as regions, locations, gateways, media resources, and so forth). Use the Sizing Tool to accurately determine the number of servers or clusters required to meet your design criteria.

Capacity Calculations

Capacity planning tools are available to Cisco partners and employees to help calculate the capacity of the Unified Communications System for large configurations. Contact your Cisco partner or Cisco Systems Engineer (SE) for assistance with sizing of your system.

For Cisco partners and employees, the tools are available at the following locations:

- Cisco Unified Communications Sizing Tool is available at
<http://tools.cisco.com/cucst>
- Cisco Unified CM Capacity Tool is available at
<http://www.cisco.com/go/cucmct>

Refer to the documentation of both of these tools to determine which tool is appropriate for the design of your system.

Computer Telephony Integration (CTI)

Cisco Computer Telephony Integration (CTI) extends the rich feature set available on Cisco Unified CM to third-party applications. These Cisco CTI-enabled applications improve user productivity, enhance the communication experience, and deliver superior customer service. At the desktop, Cisco CTI enables third-party applications to make calls from within Microsoft Outlook, open windows or start applications based on incoming caller ID, and remotely track calls and contacts for billing purposes. Cisco CTI-enabled server applications can intelligently route contacts through an enterprise network, provide automated caller services such as auto-attendant and interactive voice response (IVR), as well as capture media for contact recording and analysis.

CTI applications generally fall into one of two major categories:

- First-party applications — Monitor, control, and media termination

First-party CTI applications are designed to register devices such as CTI ports and route points for call setup, tear-down, and media termination. Because these applications are directly in the media path, they can respond to media-layer events such as in-band DTMF. Interactive voice response and Cisco Attendant Console are examples of first-party CTI applications that monitor and control calls while also interacting with call media.

- Third-party application — Monitor and control

Third-party CTI applications can also monitor and control calls, but they do not directly control media termination.

- Monitoring applications

A CTI application that monitors the state of a Cisco IP device is called a monitoring application. A busy-lamp-field application that displays on-hook/off-hook status or uses that information to indicate a user's availability in the form of Presence are both examples of third-party CTI monitoring applications.

- Call control applications

Any application that uses Cisco CTI to remotely control a Cisco IP device using out-of-band signaling is a call control application. Cisco Unified Personal Communicator, when configured to remotely control a Cisco IP device, is a good example of a call control application.

- Monitor + call control applications

These are any CTI applications that monitor and control a Cisco IP device. Cisco Unified Contact Center Enterprise is a good example of a combined monitor and control application because it monitors the status of agents and also controls agent phones through the agent desktop.

**Note**

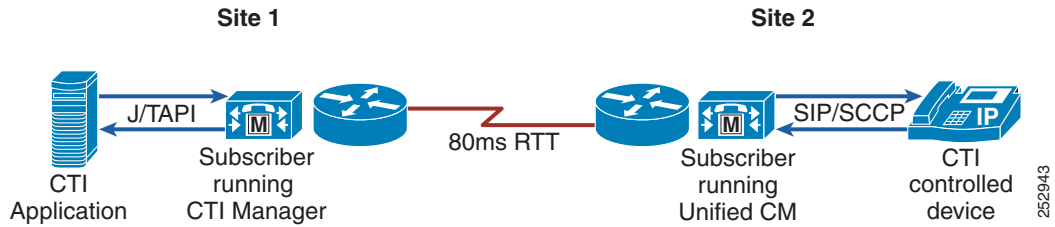
While the distinction between a monitor, call control, and monitor + control application is called out here, this granularity is not exposed to the application developer. All CTI applications using Cisco CTI are enabled for both monitoring and control.

CTI Architecture

Cisco CTI consists of the following components (see [Figure 8-8](#)), which interact to enable applications to take advantage of the telephony feature set available in Cisco Unified CM:

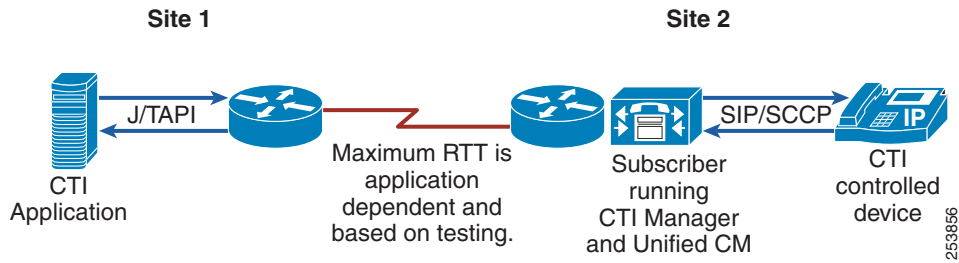
- CTI-enabled application — Cisco or third-party application written to provide specific telephony features and/or functionality.
- JTAPI and TAPI — Two standard interfaces supported by Cisco CTI. Developers can choose to write applications using their preferred method library.
- Unified JTAPI and Unified TSP Client — Converts external messages to internal Quick Buffer Encoding (QBE) messages used by Cisco Unified CM.
- Quick Buffer Encoding (QBE) — Unified CM internal communication messages.
- Provider — A logical representation of a connection between the application and CTI Manager, used to facilitate communication. The provider sends device and call events to the application while accepting control instructions that allow the application to control the device remotely.
- Specification and Description Language (SDL) — Unified CM internal communication messages.
- Publisher and subscriber — Cisco Unified Communications Manager (Unified CM) servers.
- CCM — The Cisco CallManager Service (ccm.exe), the telephony processing engine.
- CTI Manager (CTIM) — A service that runs on one or more Unified CM subscribers operating in primary/secondary mode and that authenticates and authorizes telephony applications to control and/or monitor Cisco IP devices.

Figure 8-9 CTI Over the WAN



- TAPI and JTAPI applications over the WAN (CTI application over the WAN; see [Figure 8-10](#))
- In this scenario, the CTI application is on one side of the WAN (Site 1), and its associated CTI Manager is on the other side (Site 2). In this scenario, it is up to the CTI application developer or provider to ascertain whether or not their application can accommodate the RTT as implemented. In some cases failover and failback times might be higher than if the application is co-located with its CTI Manager. In those cases, the application developer or provider should provide guidance as to the behavior of their application under these conditions.

Figure 8-10 JTAPI Over the WAN



Note

Support for TAPI and JTAPI over the WAN is application dependent. Because support for TAPI and JTAPI over the WAN is new in Cisco Unified CM 7.1(2), both customers and application developers or providers should ensure that their applications are compatible with any such deployment involving clustering over the WAN.

Unified CM Capacity Planning

Unified CM supports the following capacities for CTI.

CTI Connection Limits

Beginning with Cisco Unified CM 7.1(2), CTI capacity limits for connections have been increased based on the latest server classes, as follows:

- Cisco MCS 7825-H3/I3 supports 900 CTI connections per server when used as a dedicated subscriber, or 3,600 CTI connections per cluster. A Cisco MCS 7825-H3/I3 combined publisher/subscriber node supports 800 CTI connections.
- Cisco MCS 7835-H2/I2 supports 2,000 CTI connections per server or 8,000 per cluster.
- Cisco MCS 7845-H2/I2 supports 5,000 CTI connections per server or 20,000 per cluster.

Older server classes and older releases of Cisco Unified CM prior to version 7.1(2) support the following CTI capacity limits:

- Cisco MCS 7825 and MCS 7835 support 800 CTI connections per server or 3,200 per cluster.
- Cisco MCS 7845 supports 2,500 CTI connections per server or 10,000 per cluster.

Notes:

- Cisco CTI-enabled IP devices should always be distributed evenly across all nodes in the cluster.
- For JTAPI applications, a CTI connection is a single TCP/IP connection between each JTAPI application and a Unified CM server.
- For TAPI applications, a CTI connection is a single TCP/IP connection between the Cisco TSP residing on the TAPI application server and a Unified CM server. There could be multiple TAPI applications (on the same server) interfacing to a single TSP, in which case a single CTI connection would be utilized for all of those TAPI applications.
- Each CTI connection services one application or CTI "provider" session.
- The CTI Connection Active CTI performance monitor (perfmon) can be used to determine the total number of CTI connections on a specific Unified CM server.

CTI Associated Controlled Device Limits

Beginning with Cisco Unified CM 7.1(2), CTI capacity limits for associated controlled devices have been increased based on the latest server classes, as follows:

- Cisco MCS 7825-H3/I3 supports 900 CTI devices per server when used as a dedicated subscriber, or 3,600 CTI devices per cluster. A Cisco MCS 7825-H3/I3 combined publisher/subscriber node supports 800 CTI devices.
- Cisco MCS 7835-H2/I2 supports 2,000 CTI devices per server or 8,000 per cluster.
- Cisco MCS 7845-H2/I2 supports 5,000 CTI devices per server or 20,000 per cluster.

Older server classes and older releases of Cisco Unified CM prior to version 7.1(2) support the following CTI capacity limits:

- Cisco MCS 7825 and MCS 7835 support 800 CTI devices per server or 3,200 per cluster.
- Cisco MCS 7845 supports 2,500 CTI devices per server or 10,000 per cluster.

Notes:

- Cisco CTI-enabled IP devices should always be distributed evenly across all nodes in the cluster.
- The controlled device limits apply only to active applications; controlled devices associated to inactive (disabled) applications do not count against the limit.
- The controlled device limits assume one or two CTI controlled lines per device. Each additional CTI controlled line on the same device counts as a separate device for CTI capacity planning purposes. (For example, 400 devices with 2 CTI controlled lines per device counts the same as 400 CTI controlled devices, whereas 400 devices with 3 CTI controlled lines would count as 800 CTI devices. Shared lines on another device also count against the limit if the shared line is controlled by a CTI application.
- The controlled device limits also assume that each CTI controlled device can have up to a maximum of three (3) shared lines. Each additional shared line on the same device counts as a separate device for CTI capacity planning purposes.
- The controlled device limits further assume that each device is monitored and/or controlled by up to three (3) CTI applications.
- The controlled device limits assume basic calls only at 6 calls per hour per device, regardless of whether the device has one or two CTI controlled lines. More involved call scenarios (for example, transfer or conference) and/or higher call rates will affect the limits. (For proper sizing, use the Cisco Unified Communications Sizing Tool, available to Cisco employees and partners with proper login authentication at <http://tools.cisco.com/cucst>.)
- The greater the number of controlled devices associated with a CTI application, the longer it will take for application initialization and Unified CM failover/failback handling. This applies even if the application is not actively controlling the device.
- The Devices Open and Lines Open CTI Performance monitors (perfmon) can be used to determine the total number of devices and lines that are currently controlled by applications on a specific Unified CM server.

For Cisco Unified Communications Manager Business Edition, this server has a maximum of 500 CTI devices.

Provisioning

CTI Manager

CTI Manager must be enabled on at least one and possibly all call processing subscribers within the Unified CM cluster. The client-side interfaces (TAPI TSP or JTAPI client) allow for two IP addresses each, which then point to Unified CM servers running the CTIM service. For CTI application redundancy, Cisco recommends having the CTIM service activated on at least two Unified CM servers in a cluster, as shown in [Figure 8-11](#).

Redundancy, Failover, and Load Balancing

For CTI applications that require redundancy, the TAPI TSP or JTAPI client can be configured with two IP addresses, thereby allowing an alternate CTI Manager to be used in the event of a failure. It should be noted that this redundancy is not stateful in that no information is shared and/or made available between the two CTI Managers, and therefore the CTI application will have some degree of re-initialization to go through, depending on the exact nature of the failover.

When a CTI Manager fails-over, just the CTI application login process is repeated on the now-active CTI Manager. Whereas, if the Unified CM server itself fails, then the re-initialization process is longer due to the re-registration of all the devices from the failed Unified CM to the now-active Unified CM, followed by the CTI application login process.

For CTI applications that require load balancing or that could benefit from this configuration, the CTI application can simply connect to two CTI Managers simultaneously, as shown in [Figure 8-11](#).

Figure 8-11 Redundancy and Load Balancing

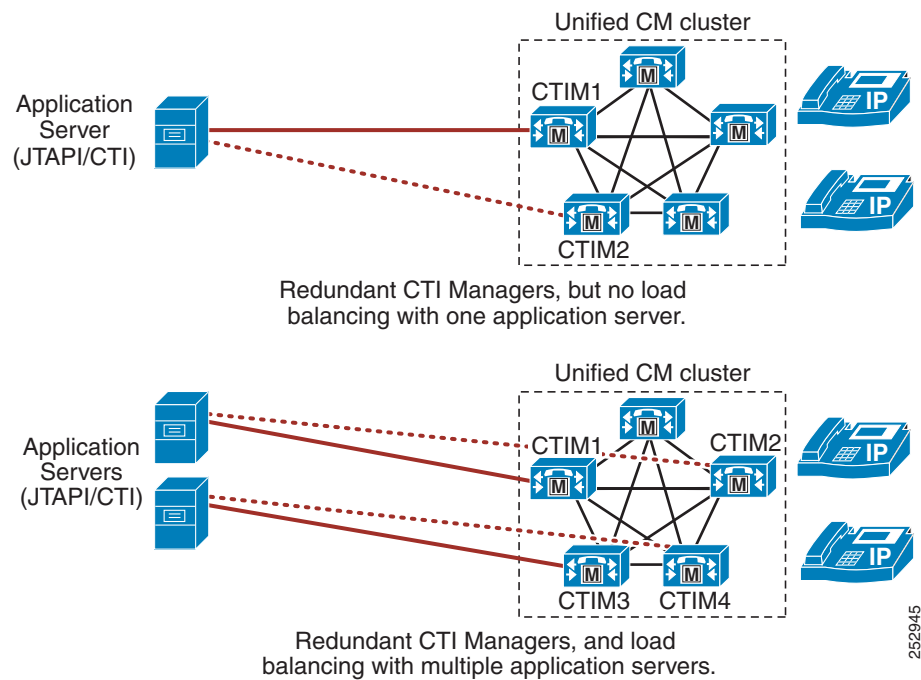
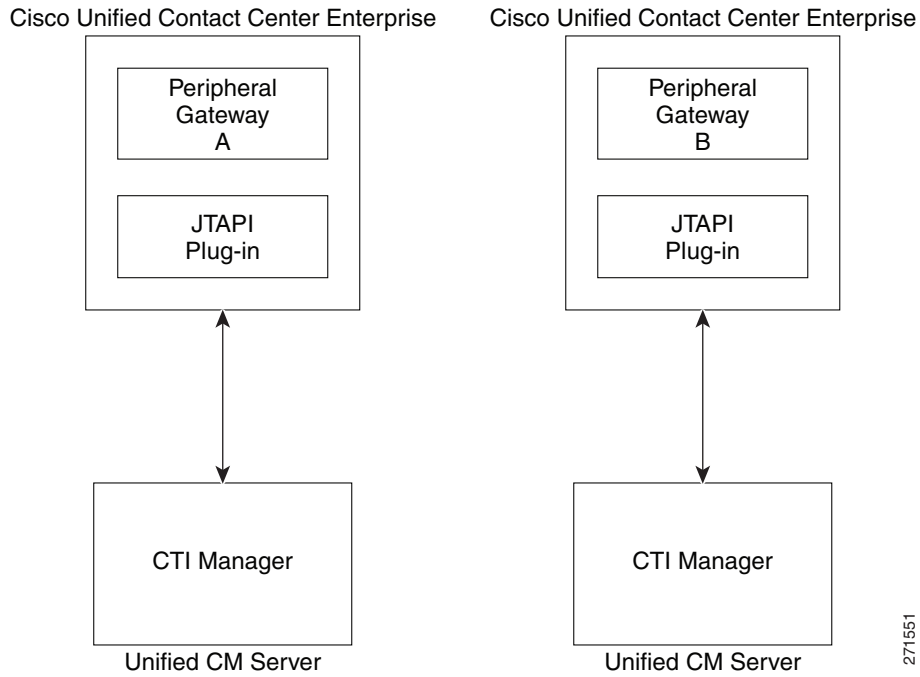


Figure 8-12 shows an example of this type of configuration for Cisco Unified Contact Center Enterprise (Unified CCE). This type of configuration has the following characteristics:

- Unified CCE uses two Peripheral Gateways (PGs) for redundancy.
- Each PG logs into a different CTI Manager.
- Only one PG is active at any one time.

Figure 8-12 CTI Redundancy with Cisco Unified Contact Center Enterprise

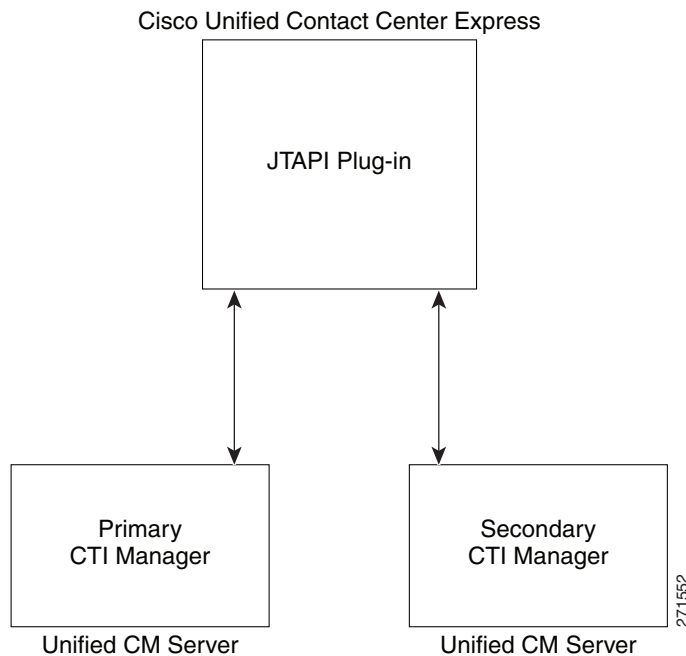


271551

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

- Unified CCX has two IP addresses configured, one for each CTI Manager.
- If connection to the primary CTI Manager is lost, Unified CCX fails-over to its secondary CTI Manager.

Figure 8-13 CTI Redundancy with Cisco Unified Contact Center Express



Implementation

For guidance and support on writing applications, application developers should consult the Cisco Developer Connection, located at

<http://developer.cisco.com/web/cdc/community>

Gatekeeper Design Considerations

A single Cisco IOS gatekeeper can provide call routing and call admission control for up to 100 Unified CM clusters in a distributed call processing environment. Multiple gatekeepers can be configured to support thousands of Unified CM clusters. You can also implement a hybrid Unified CM and toll-bypass network by using Cisco IOS gatekeepers to provide communication and call admission control between the H.323 gateways and Unified CM.

Gatekeeper call admission control is a policy-based scheme requiring static configuration of available resources. The gatekeeper is not aware of the network topology, so it is limited to hub-and-spoke topologies.

The Cisco 2600, 2800, 2900, 3600, 3700, 3800, 3900, and 7200 Series routers all support the gatekeeper feature. You can configure Cisco IOS gatekeepers in a number of different ways for redundancy, load balancing, and hierarchical call routing. This section considers the design requirements for building a gatekeeper network, but it does not deal with the call admission control or dial plan resolution aspects, which are covered in the chapters on [Call Admission Control, page 9-1](#), and [Dial Plan, page 10-1](#), respectively.

For additional information regarding gatekeepers, refer to the *Cisco IOS H.323 Configuration Guide*, available at

http://www.cisco.com/en/US/products/ps6441/products_installation_and_configuration_guides_list.html

Hardware Platform Selection

The choice of gatekeeper platform is based on the number of calls per second and the number of concurrent calls. A higher number of calls per second requires a more powerful CPU, such as a Cisco 3800, 3900, or 7200 Series Router. A higher number of concurrent calls requires more memory. For more information about gatekeeper platforms, refer to the *Cisco IOS H323 Gatekeepers Data Sheet*, available at

http://www.cisco.com/en/US/prod/collateral/voicesw/ps6788/vcallcon/ps4139/data_sheet_c78_561921.html

For additional information on gatekeeper platform selection, contact your Cisco partner or Cisco Systems Engineer (SE).

Gatekeeper Redundancy

With gatekeepers providing all call routing and admission control for intercluster communications, redundancy is required. Three methods of gatekeeper redundancy are available: Hot Standby Router Protocol (HSRP), gatekeeper clustering, and redundant gatekeeper trunks. The following sections describe these methods.



Note

Cisco recommends that you use gatekeeper clustering to provide gatekeeper redundancy whenever possible. Use HSRP for redundancy only if gatekeeper clustering is not available in your software feature set.

Hot Standby Router Protocol (HSRP)

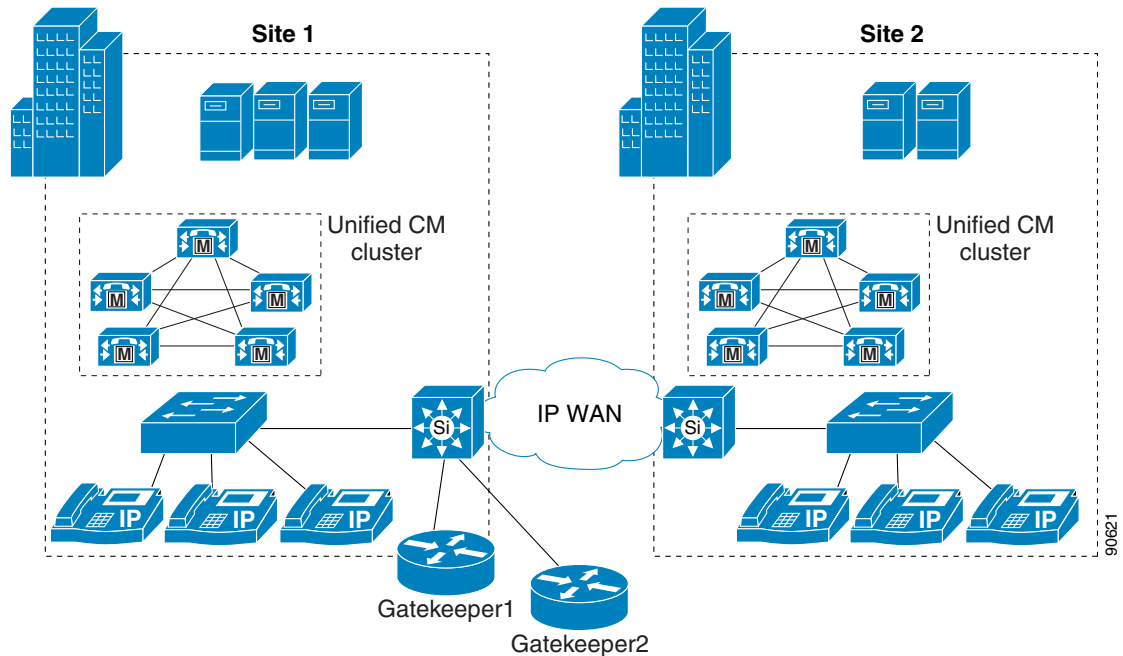
The following guidelines apply to HSRP:

- Only one gatekeeper is active at a time.
 - The standby gatekeeper does not process any calls unless the primary fails.
 - No load balancing features are available.
- All gatekeepers must reside in the same subnet or location.
- No previous state information is available after failover.
- After a failover, the standby gatekeeper is not aware of the calls that are already active, so over-subscription of the bandwidth is possible.

- Failover time can be substantial because the endpoints have to re-register with the HSRP standby gatekeeper before calls can be placed. The failover time depends on the settings of the registration timers.

Figure 8-14 show a network configuration using HSRP for gatekeeper redundancy.

Figure 8-14 Gatekeeper Redundancy Using HSRP



Example 8-1 shows the configuration for Gatekeeper 1 and Example 8-2 shows the configuration for Gatekeeper 2 in Figure 8-14. Both configurations are identical except for the HSRP configuration on the Ethernet interface.

Example 8-1 Configuration for Gatekeeper 1

```
interface Ethernet0/0
ip address 10.1.10.2 255.255.255.0
 standby ip 10.1.10.1
 standby priority 110

gatekeeper
zone local GK-Site1 customer.com 10.1.10.1
zone local GK-Site2 customer.com
zone prefix GK-Site1 408.....
zone prefix GK-Site2 212.....
bandwidth interzone default 160
gw-type-prefix 1#* default-technology
arq reject-unknown-prefix
no shutdown
```

Example 8-2 Configuration for Gatekeeper 2

```
interface Ethernet0/0
 ip address 10.1.10.3 255.255.255.0
 standby ip 10.1.10.1

gatekeeper
 zone local GK-Site1 customer.com 10.1.10.1
 zone local GK-Site2 customer.com
 zone prefix GK-Site1 408.....
 zone prefix GK-Site2 212.....
 bandwidth interzone default 160
 gw-type-prefix 1#* default-technology
 arq reject-unknown-prefix
 no shutdown
```

The following notes also apply to [Example 8-1](#) and [Example 8-2](#):

- Each router has **standby** commands configured for HSRP and to identify the virtual IP address shared by each. Gatekeeper 1 is configured as the primary with the command **standby priority 110**.
- Each Unified CM cluster has a local zone configured on each router to support Unified CM trunk registrations. Note that the IP address defined on the first zone should match the virtual IP address used by HSRP.
- A zone prefix is configured for each zone on both routers, allowing inter-zone and inter-cluster call routing.
- Bandwidth statements are configured on each router for both sites. Cisco recommends that you use the **bandwidth interzone** command because the **bandwidth total** command does not work in some configurations.
- The **gw-type-prefix 1# default-technology** command is configured on both routers, allowing all locally unresolved calls to be forwarded to a device registered with a technology prefix of 1# in the local zone. In this example, all Unified CM trunks have been configured to register with a 1# prefix.
- The **arq reject-unknown-prefix** command is configured on both routers to guard against potential call routing loops across redundant Unified CM trunks.

For additional and advanced HSRP information, refer to the online documentation at the following locations:

- <http://www.cisco.com/en/US/docs/internetworking/case/studies/cs009.html>
- http://www.cisco.com/en/US/tech/tk648/tk362/technologies_q_and_a_item09186a00800a9679.shtml
- http://www.cisco.com/en/US/tech/tk648/tk362/technologies_tech_note09186a0080094afd.shtml

Gatekeeper Clustering (Alternate Gatekeeper)

Gatekeeper clustering (alternate gatekeeper) enables the configuration of a "local" gatekeeper cluster, with each gatekeeper acting as primary for some Unified CM trunks and an alternate for others. Gatekeeper Update Protocol (GUP) is used to exchange state information between gatekeepers in a local cluster. GUP tracks and reports CPU utilization, memory usage, active calls, and number of registered endpoints for each gatekeeper in the cluster. Load balancing is supported by setting thresholds for any of the following parameters in the GUP messaging:

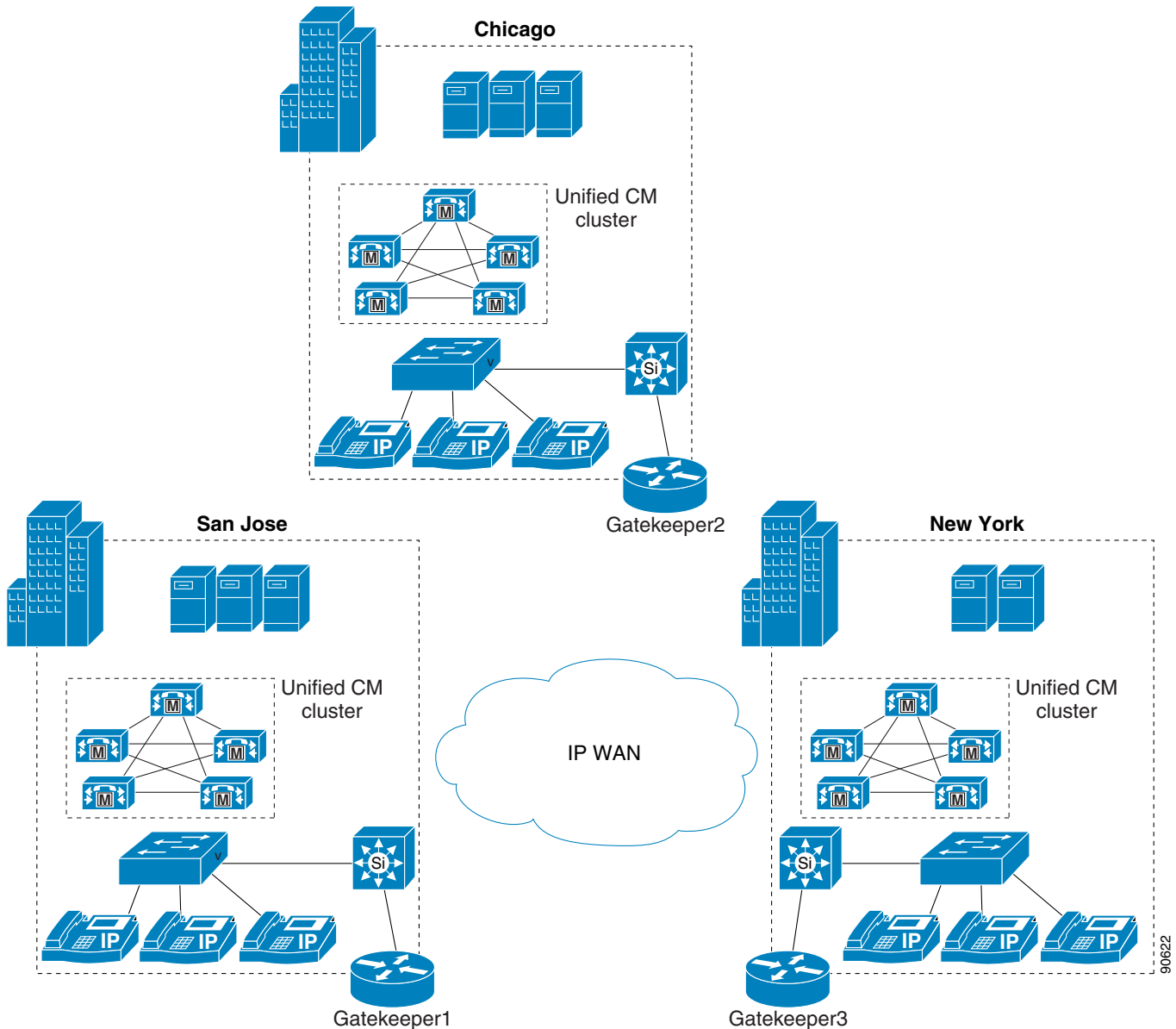
- CPU utilization
- Memory utilization
- Number of active calls
- Number of registered endpoints

With the support of gatekeeper clustering (alternate gatekeeper), stateful redundancy and load balancing are available. Gatekeeper clustering provides the following features:

- Local and remote clusters
- Up to five gatekeepers in a local cluster
- Gatekeepers in local clusters can be located in different subnets or locations
- No failover delay (Because the alternate gatekeeper is already aware of the endpoint, it does not have to go through the full registration process.)
- Gatekeepers in a cluster pass state information and provide load balancing

Figure 8-15 shows three sites with Unified CM distributed call processing and three distributed gatekeepers configured in a local cluster.

Figure 8-15 Gatekeeper Clustering



In Figure 8-15, Gatekeeper 2 is the backup for Gatekeeper 1, Gatekeeper 3 is the backup for Gatekeeper 2, and Gatekeeper 1 is the backup for Gatekeeper 3.

Example 8-3 shows the configuration for Gatekeeper 1 (SJC), and Example 8-4 shows the configuration for Gatekeeper 2 (CHC). The configuration for Gatekeeper 3 (NYC) is not shown because it is very similar to the other two.

Example 8-3 Gatekeeper Clustering Configuration for Gatekeeper 1

```
gatekeeper
zone local SJC cisco.com 10.1.1.1
zone local CHC_GK1 cisco.com
zone local NYC_GK1 cisco.com
!
```

90622

```

zone cluster local SJC_Cluster SJC
  element SJC_GK2 10.1.2.1 1719
  element SJC_GK3 10.1.3.1 1719
!
zone cluster local CHC_Cluster CHC_GK1
  element CHC 10.1.2.1 1719
  element CHC_GK3 10.1.3.1 1719
!
zone cluster local NYC_Cluster NYC_GK1
  element NYC 10.1.3.1 1719
  element NYC_GK2 10.1.2.1 1719
!
zone prefix SJC 40852.....
zone prefix NYC_GK1 21251.....
zone prefix CHC_GK1 72067.....
gw-type-prefix 1#* default-technology
load-balance cpu 80 memory 80
bandwidth interzone SJC 192
bandwidth interzone NYC_GK1 160
bandwidth interzone CHC_GK1 160
arq reject-unknown-prefix
no shutdown

```

Example 8-4 Gatekeeper Clustering Configuration for Gatekeeper 2

```

gatekeeper
zone local CHC cisco.com 10.1.2.1
zone local SJC_GK2 cisco.com
zone local NYC_GK2 cisco.com
!
zone cluster local CHC_Cluster CHC
  element CHC_GK3 10.1.3.1 1719
  element CHC_GK1 10.1.1.1 1719
!
zone cluster local SJC_Cluster SJC_GK2
  element SJC 10.1.1.1 1719
  element SJC_GK3 10.1.3.1 1719
!
zone cluster local NYC_Cluster NYC_GK2
  element NYC_GK1 10.1.1.1 1719
  element NYC 10.1.3.1 1719
!
zone prefix SJC_GK2 40852.....
zone prefix NYC_GK2 21251.....
zone prefix CHC 72067.....
gw-type-prefix 1#* default-technology
load-balance cpu 80 memory 80
bandwidth interzone CHC_Voice 160
bandwidth interzone SJC_Voice2 192
bandwidth interzone NYC_Voice3 160
arq reject-unknown-prefix
no shutdown

```

The following notes also apply to [Example 8-3](#) and [Example 8-4](#):

- Each Unified CM cluster has a local zone configured to support Unified CM trunk registrations.
- A cluster is defined for each local zone, with backup zones on the other gatekeepers listed as elements. Elements are listed in the order in which the backups are used.
- A zone prefix is configured for each zone to allow inter-zone and inter-cluster call routing.

- The **gw-type-prefix 1# default-technology** command allows all locally unresolved calls to be forwarded to a device registered with a technology prefix of 1# in the local zone. In this example, all Unified CM trunks have been configured to register with a 1# prefix.
- The **load-balance cpu 80 memory 80** command limits CPU and memory usage. If the router hits either limit, all new requests are denied and the first backup in the list is used until utilization drops below the threshold.
- Bandwidth statements are configured for each site. Cisco recommends that you use the **bandwidth interzone** command because the **bandwidth total** command does not work in some configurations.
- The **arq reject-unknown-prefix** command guards against potential call routing loops across redundant Unified CM trunks.

All gatekeepers in the cluster display all Unified CM trunk registrations. For trunks that use the gatekeeper as a primary resource, the flag field is blank. For trunks that use another gatekeeper in the cluster as their primary gatekeeper, the flag field is set to A (alternate). Having all endpoints registered as primary or alternate allows all calls to be resolved locally without having to send a location request (LRQ) to another gatekeeper.

[Example 8-5](#) shows the output from the **show gatekeeper endpoints** command on Gatekeeper 1 (SJC).

Example 8-5 Output for Gatekeeper Endpoints

```

                                GATEKEEPER ENDPOINT REGISTRATION
                                =====
CallSignalAddr  Port  RASSignalAddr  Port  Zone Name          Type      Flags
-----
10.1.1.12       1307  10.1.1.12      1254  SJC                 VOIP-GW
H323-ID: SJC-to-GK-trunk_1
10.1.1.12       4422  10.1.1.12      4330  SJC                 VOIP-GW
H323-ID: SJC-to-GK-trunk_2
10.1.2.12       4587  10.1.2.12      4330  CHC_GK1             VOIP-GW  A
H323-ID: CHC-to-GK-trunk_1
10.1.3.21       2249  10.1.3.21      1245  NYC_GK1             VOIP-GW  A
H323-ID: NYC-to-GK-trunk_1
Total number of active registrations = 4

```

Directory Gatekeeper Redundancy

You can implement directory gatekeeper redundancy by using HSRP or by configuring multiple identical directory gatekeepers. When a gatekeeper is configured with multiple remote zones using the same zone prefix, the gatekeeper can use either of the following methods:

- Sequential LRQs (default)

Redundant remote zones (matching zone prefixes) are assigned a cost, and LRQs are sent to the matching zones in order based on the cost values. Using sequential LRQs saves WAN bandwidth by not blasting LRQs to all matching gatekeepers.

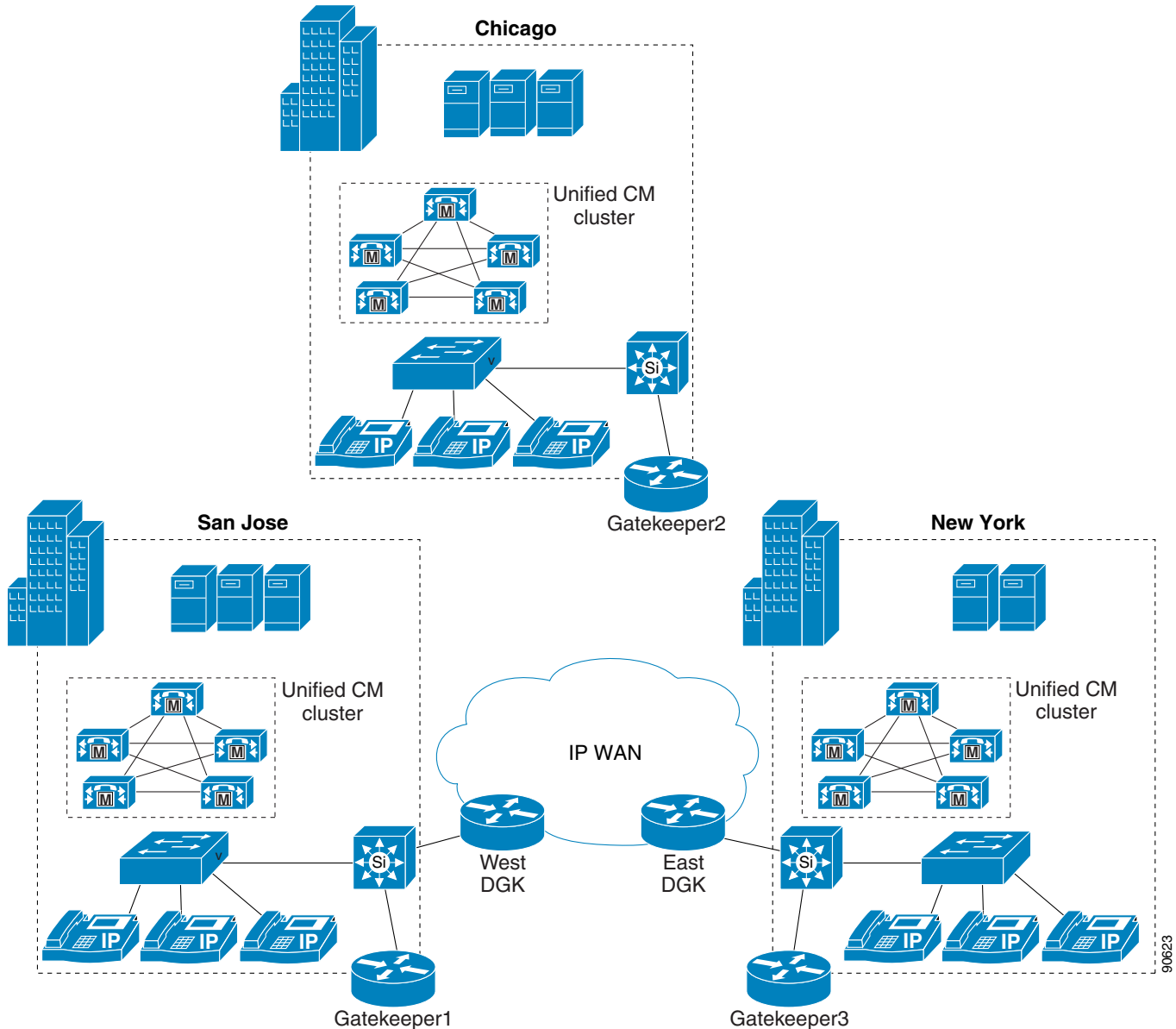
- LRQ Blast

LRQs are sent to redundant zones (matching zone prefixes) simultaneously. The first gatekeeper to respond with an Location Confirm (LCF) is the one that is used.

Cisco recommends that you use multiple active directory gatekeepers with sequential LRQs, thus allowing directory gatekeepers to be placed in different locations. Using HSRP requires both directory gatekeepers to be located in the same subnet, and only one gatekeeper can be active at any time.

Figure 8-16 illustrates a Unified CM distributed call processing environment with two active directory gatekeepers.

Figure 8-16 Redundant Directory Gatekeepers



Example 8-6 and Example 8-7 show the configurations for the two directory gatekeepers in Figure 8-16.

Example 8-6 Configuration for West Directory Gatekeeper

```
gatekeeper
zone local DGKW customer.com 10.1.10.1
zone remote SJC customer.com 10.1.1.1
zone remote CHC customer.com 10.1.2.1
zone remote NYC customer.com 10.1.3.1
zone prefix SJC 408.....
```

```

zone prefix CHC 720.....
zone prefix NYC 212.....
lrq forward-queries
no shutdown

```

Example 8-7 Configuration for East Directory Gatekeeper

```

gatekeeper
zone local DGKE customer.com 10.1.12.1
zone remote SJC customer.com 10.1.1.1
zone remote CHC customer.com 10.1.2.1
zone remote NYC customer.com 10.1.3.1
zone prefix SJC 408.....
zone prefix CHC 720.....
zone prefix NYC 212.....
lrq forward-queries
no shutdown

```

The following notes also apply to [Example 8-6](#) and [Example 8-7](#):

- Both directory gatekeepers are configured exactly the same.
- A local zone is configured for the directory gatekeeper.
- Remote zones are configured for each remote gatekeeper.
- Zone prefixes are configured for both remote zones for inter-zone call routing. The wildcard (*) could be used in the zone prefix to simplify configuration, but the use of dots (.) is more specific. Calls are not routed to the DGK zone, so a prefix is not required for it.
- The **lrq forward-queries** command allows the directory gatekeeper to forward an LRQ received from another gatekeeper.



Note

Directory gatekeepers do not contain any active endpoint registrations and do not supply any bandwidth management.

[Example 8-8](#), [Example 8-9](#), and [Example 8-10](#) show the configurations for Gatekeepers 1 to 3 in [Figure 8-16](#).

Example 8-8 Configuration for Gatekeeper 1 (SJC)

```

zone local SJC customer.com 10.1.1.1
zone remote DGKW customer.com 10.1.10.1
zone remote DGKE customer.com 10.1.12.1
zone prefix SJC 408.....
zone prefix DGKW .....
zone prefix DGKE .....
bandwidth remote 192
gw-type-prefix 1# default-technology
arq reject-unknown-prefix
no shutdown

```

Example 8-9 Configuration for Gatekeeper 2 (CHC)

```

gatekeeper
zone local GK-CHC customer.com 10.1.2.1
zone remote DGKE customer.com 10.1.12.1
zone remote DGKW customer.com 10.1.10.1

```

```

zone prefix CHC 720.....
zone prefix DGKE .....
zone prefix DGKW .....
bandwidth remote 160
gw-type-prefix 1# default-technology
arq reject-unknown-prefix
no shutdown

```

Example 8-10 Configuration for Gatekeeper 3 (NYC)

```

gatekeeper
zone local NYC customer.com 10.1.3.1
zone remote DGKE customer.com 10.1.12.1
zone remote DGKW customer.com 10.1.10.1
zone prefix NYC 212.....
zone prefix DGKE .....
zone prefix DGKW .....
bandwidth remote 160
gw-type-prefix 1# default-technology
arq reject-unknown-prefix
no shutdown

```

The following notes also apply to [Example 8-8](#), [Example 8-9](#), and [Example 8-10](#):

- Each Unified CM cluster has a local zone configured to support Unified CM trunk registrations.
- Remote zones are configured for each directory gatekeeper.
- Zone prefixes are configured for the local zone and both remote zones for inter-zone call routing. Both directory gatekeeper prefixes are 10 dots. Sequential LRQs are used by default when matching zone prefixes are configured. The gatekeeper sends an LRQ to the directory gatekeeper with the lowest cost; if there is no response, the gatekeeper tries the second directory gatekeeper.
- The **bandwidth remote** command is used to limit bandwidth between the local zone and any other remote zone.
- The **gw-type-prefix 1# default-technology** command allows all locally unresolved calls to be forwarded to a device registered with a technology prefix of 1# in the local zone. In this example, all Unified CM trunks have been configured to register with a 1# prefix.
- The **arq reject-unknown-prefix** command guards against potential call routing loops across redundant Unified CM trunks.

Interoperability of Unified CM and Unified CM Express

This section explains the requirements for interoperability and internetworking of Cisco Unified CM with Cisco Unified Communications Manager Express (Unified CME) using H.323 or SIP trunking protocol in a multisite IP telephony deployment. This section highlights the recommended deployments between phones controlled by Unified CM and phones controlled by Unified CME.

This section covers the following topics:

- [Overview of Interoperability Between Unified CM and Unified CME, page 8-37](#)
- [Unified CM and Unified CME Interoperability via SIP in a Multisite Deployment with Distributed Call Processing, page 8-38](#)
- [Unified CM and Unified CME Interoperability via H.323 in a Multisite Deployment with Distributed Call Processing, page 8-42](#)

Overview of Interoperability Between Unified CM and Unified CME

Cisco Unified CME supports Cisco IP SIP Phones 7905G, 7906G, 7911G, 7912G, 7940G, 7960G, 7941G, 7942G, 7945G, 7961G, 7962G, 7965G, 7970G, 7971G, 7975G, 8961, 9951, and 9971 in addition to Cisco IP SCCP Phones 6921, 6941, 6961, 7905G, 7906G, 7911G, 7912G, 7914, 7920, 7921G, 7931G, 7935, 7936, 7940G, 7960G, 7941G, 7942G, 7945G, 7961G, 7962G, 7965G, 7970G, 7971G, 7975G, 7985G, and Cisco IP Communicator.

All call signaling is sent through Unified CME, regardless of the endpoint being used. However, with SCCP endpoints on the same Unified CME, the media can flow around Unified CME. With SIP endpoints on the same Unified CME, the media can flow around Unified CME with Unified CME release 4.1 and later, but the media flows through Unified CME with releases prior to 4.1.

**Note**

Flow around means the media is sent directly between the phones. *Flow through* means the media is not sent directly to the phones, but is sent through Unified CME.

Either H.323 or SIP can be used as a trunking protocol to interconnect Unified CM and Unified CME. When deploying Unified CM at the headquarters or central site in conjunction with one or more Unified CME systems for branch offices, network administrators must choose either the SIP or H.323 protocol by taking careful consideration of protocol specifics and supported features across the WAN trunk. Using H.323 trunks to connect Unified CM and Unified CME has been the predominant method in past years, until more enhanced capabilities for SIP phones and SIP trunks were added in Unified CM and Unified CME. This section first describes some of the features and capabilities that are independent of the trunking protocol for Unified CM and Unified CME interoperability, then it explains some of the most common design scenarios and best practices for using SIP trunks and H.323 trunks.

Call Types and Call Flows

In general, Unified CM and Unified CME interworking allows all combination of calls from SCCP IP phones to SIP IP phones, or vice versa, across a SIP trunk or H.323 trunk. Calls can be transferred (blind or consultative) or forwarded back and forth between the Unified CM and Unified CME SIP and/or SCCP IP phones.

When connected to Unified CM via H.323 trunks, Unified CME can auto-detect Unified CM calls. When a call terminating on Unified CME is transferred or forwarded, Unified CME regenerates the call and routes the call appropriately to another Unified CME or Unified CM by hairpinning the call. Unified CME hairpins the call legs from Unified CM for the VoIP calls across SIP or H.323 trunks when needed. For more information on allowing auto-detection on a non-H.450 supported Unified CM network and for enabling or disabling supplementary services for H450.2, H450.3, or SIP, refer to the Unified CME product documentation available at http://www.cisco.com/en/US/products/sw/voicesw/ps4625/tsd_products_support_series_home.html.

When connected to Unified CM via SIP trunks, Unified CME does not auto-detect Unified CM calls. By default, Unified CME always tries to redirect calls using either a SIP Refer message for call transfer or a SIP 302 Moved Temporarily message for call forward; if that fails, Unified CME will then try to hairpin the call.

Music on Hold

While Unified CM can be enabled to stream MoH in both G.711 and G.729 formats, Unified CME streams MoH only in G.711 format. Therefore, when Unified CME controls the MoH audio on a call placed on hold, it requires a transcoder to transcode between a G.711 MoH stream and a G.729 call leg.

Ad Hoc and Meet Me Hardware Conferencing

Hardware DSP resources are required for both Ad Hoc and Meet Me conferences. Whether connected via SIP, H.323, or PSTN, both Unified CM and Unified CME phones can be invited or added to an Ad Hoc conference to become conference participants as long as the phones are reachable from the network. When calls are put on hold during an active conference session, music will not be heard by the conference participants in the conference session.

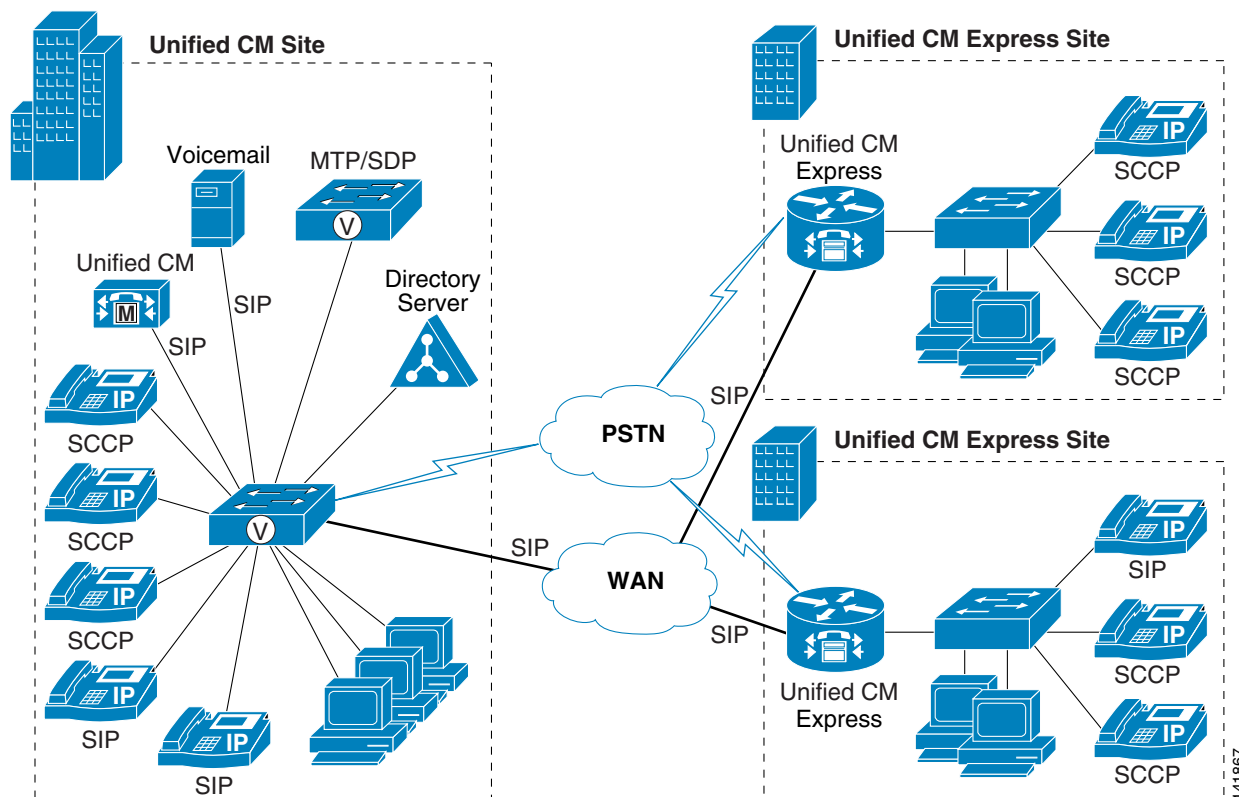
For information on required and supported DSP resources and the maximum number of conference participants allowed for Ad Hoc or Meet Me conferences, refer to the Unified CME product documentation available at

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

Unified CM and Unified CME Interoperability via SIP in a Multisite Deployment with Distributed Call Processing

Unified CM can communicate directly with Unified CME using a SIP interface. [Figure 8-17](#) shows a Cisco Unified Communications multisite deployment with Unified CM networked directly with Cisco Unified CME using a SIP trunk interface.

Figure 8-17 Multisite Deployment with Unified CM and Unified CME Using SIP Trunks



Best Practices

Follow these guidelines and best practices when using the deployment model illustrated in [Figure 8-17](#):

- Configure a SIP Trunk Security Profile with **Accept Replaces Header** selected.
- Configure a SIP trunk on Unified CM using the SIP Trunk Security Profile created, and also specify a ReRouting CSS. The ReRouting CSS is used to determine where a SIP user (transferor) can refer another user (transferee) to a third user (transfer target) and which features a SIP user can invoke using the SIP 302 Redirection Response and INVITE with Replaces.
- For SIP trunks there is no need to enable the use of media termination points (MTPs) when using SCCP endpoints on Unified CME. However, SIP endpoints on Unified CME require the use of media termination points on Unified CM to be able to handle delayed offer/answer exchanges with the SIP protocol (that is, the reception of INVITEs with no Session Description Protocol).
- Route calls to Unified CME via a SIP trunk using the Unified CM dial plan configuration (route patterns, route lists, and route groups).
- Use Unified CM device pools and regions to configure a G.711 codec within the site and the G.729 codec for remote Unified CME sites.
- Configure the **allow-connections sip to sip** command under **voice services voip** on Unified CME to allow SIP-to-SIP call connections.
- For SIP endpoints, configure the **mode cme** command under **voice register global**, and configure **dtmf-relay rtp-nte** under the **voice register pool** commands for each SIP phone on Unified CME.
- For SCCP endpoints, configure the **transfer-system full-consult** command and the **transfer-pattern .T** command under **telephony-service** on Unified CME.
- Configure the SIP WAN interface voip dial-peers to forward or redirect calls, destined for Unified CM, with **session protocol sipv2** and **dtmf-relay [sip-notify | rtp-nte]** on Unified CME.

See [Example 8-11](#) for a sample Unified CME configuration using this SIP deployment model.

Design Considerations

This section first covers some characteristics and design considerations for Unified CM and Unified CME interoperability via SIP in some main areas such as supplementary services for call transfer and forward, presence service for busy lamp field (BLF) notification for speed-dial buttons and directory call lists, and out-of-dialog (OOD-Refer) for integration with partner applications and third-party phone control for click-to-dial between the Unified CM phones and Unified CME phones. The section also covers some general design considerations for Unified CM and Unified CME interoperability via SIP.

Supplementary Services

SIP Refer or SIP 302 Moved Temporarily messages can be used for supplementary services such as call transfer or call forward on Unified CME or Unified CM to instruct the transferee (referee) or phone being forwarded (forwardee) to initiate a new call to the transfer-to (refer-to) target or forward-to target. No hairpinning is needed for call transfer or call forward scenarios when the SIP Refer or SIP 302 Moved Temporarily message is supported.

However, **supplementary-service** must be disabled if there are certain extensions that have no DID mapping or if Unified CM or Unified CME does not have a dial plan to route the call to the DID in the SIP 302 Moved Temporarily message. When **supplementary-service** is disabled, Unified CME hairpins the calls or sends a re-invite SIP message to Unified CM to replace the media path to the new called party ID. Both signaling and media are hairpinned, even when multiple Unified CMEs are involved for further

call forwards. The **supplementary-service** can also be disabled for transferred calls. In this case, the SIP Refer message will not be sent to Unified CM, but the transferee (referee) party and transfer-to party (refer-to target) are hairpinned.

**Note**

Supplementary services can be disabled with the command **no supplementary-service sip moved-temporarily** or **no supplementary-service sip refer** under **voice service voip** or **dial-peer voice xxxx voip**.

The following examples illustrate the call flows when supplementary services are disabled:

- Unified CM phone B calls Unified CME phone A, which is set to call-forward (all, busy, or no answer) to phone C (either a Unified CM phone, a Unified CME phone on the same or different Unified CME, or a PSTN phone).

Unified CME does not send the SIP 302 Moved Temporarily message to Unified CM, but hairpins the call between Unified CM phone B and phone C.

- Unified CM phone B calls Unified CME phone A, which transfer the call to phone C (either a Unified CM phone, a Unified CME phone, or a PSTN phone).

Unified CME does not send the SIP Refer message to Unified CM, but hairpins the call between Unified CM phone B and phone C.

General Design Considerations for Unified CM and Unified CME Interoperability via SIP

- Use Unified CME 4.1 and later releases for basic calls and supplementary features through SIP trunks to Unified CM.
- Disable **supplementary-service** if SIP 302 Moved Temporarily or SIP Refer messages are not supported by Unified CM, otherwise Unified CM cannot route the call to the transfer-to or forward-to target.
- In a SIP-to-SIP call scenario, a Refer message is sent by default from the transferor to the transferee, the transferee sets up a new call to the transfer-to target, and the transferor hears ringback tone by default while waiting for the transfer at connect. If **supplementary-service** is disabled on Unified CME, Unified CME will provide in-band ringback tone right after the call between the transferee and transfer-to target is connected.
- Presence service is supported on Unified CM and Unified CME via SIP trunk only.
- The OOD-Refer feature allows third-party applications to connect two endpoints on Unified CM or Unified CME through the use of the SIP REFER method. Consider the following factors when using OOD-Refer:
 - Both Unified CM and Unified CME must be configured to enable the OOD-Refer feature.
 - Call Hold, Transfer, and Conference are not supported during an OOD-Refer transaction, but they are not blocked by Unified CME.
 - Call transfer is supported only after the OOD-Refer call is in the connected state and not before the call is connected; therefore, call transfer-at-alert is not supported.
- Control signaling in TLS is supported, but SRTP is not supported over the SIP trunk.
- Video is not supported on SIP phones, nor is it supported over SIP trunks.
- SRTP over a SIP trunk is a gateway feature in Cisco IOS Release 12.4(15)T for Unified CM. SRTP support is not available with Unified CM and Unified CME interworking via SIP trunks.

**Note**

When multiple PSTN connections exist (one for Unified CM and one for Unified CME), fully attended transfer between a Unified CM endpoint and a Unified CME endpoint to a PSTN endpoint will fail. The recommendation is to use blind transfer when using multiple PSTN connections, and it is configured under **telephony-service** as **transfer-system full-blind**.

**Note**

Cisco Unified CME supports video calls across SIP trunks between multiple Unified CMEs. This capability applies in distributed call processing deployments that use Unified CMEs only. Video calls between Unified CM and Unified CME over SIP trunks are not supported currently. For configuration details, 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.

Configuration Examples

The following examples illustrate some of the design considerations and best practices discussed in this section.

Example 8-11 Configuration for Cisco Unified CME with SIP

```
voice service voip
  allow-connections sip to sip
  sip
  registrar server
dial-peer voice 1 voip      /* To Unified CM endpoints */
  destination-pattern xxxx
  session protocol sipv2
  session target ipv4:10.10.10.20
  session transport udp    /* tcp can be used here also */
  dtmf-relay rtp-nte
  codec g729r8             /* Voice class can also be used */
  no vad
voice register global
  mode cme
  source-address 10.10.10.21 port 5060
  create profile
voice register pool 1
  id mac 0007.0E8B.5777
  type 7940
  number 1 dn 1
  codec g729r8             /* Voice class can also be used */
  dtmf-relay rtp-nte
telephony-service
  load 7960-7940 POS3-07-04-11
  ip source-address 10.10.10.22 port 2000
  create cnf-files
  keepalive 45
  max-conferences 8 gain -6
  moh music-on-hold.au
  transfer-system full-consult /* full-blind can also be used */
  transfer-pattern .T
```

Unified CM and Unified CME Interoperability via H.323 in a Multisite Deployment with Distributed Call Processing

There are two deployment options to achieve interoperability between Unified CM and Unified CME via H.323 connections in a multisite WAN deployment with distributed call processing. The first option is to deploy a Cisco Unified Border Element as a front-end device of Unified CM, which has a peer-to-peer H.323 connection with a remote Unified CME system. The Cisco Unified Border Element performs dial plan resolution between Unified CM and Unified CME, and it also terminates and re-originates call signaling messages between the two. The Cisco Unified Border Element acts as a proxy device for a system that does not support H.450 for its supplementary services, such as Unified CM, which uses Empty Capability Sets (ECS) to invoke supplementary services. The Cisco Unified Border Element can also act as the PSTN gateway for the Unified CM cluster so that a separate PSTN gateway is not needed.

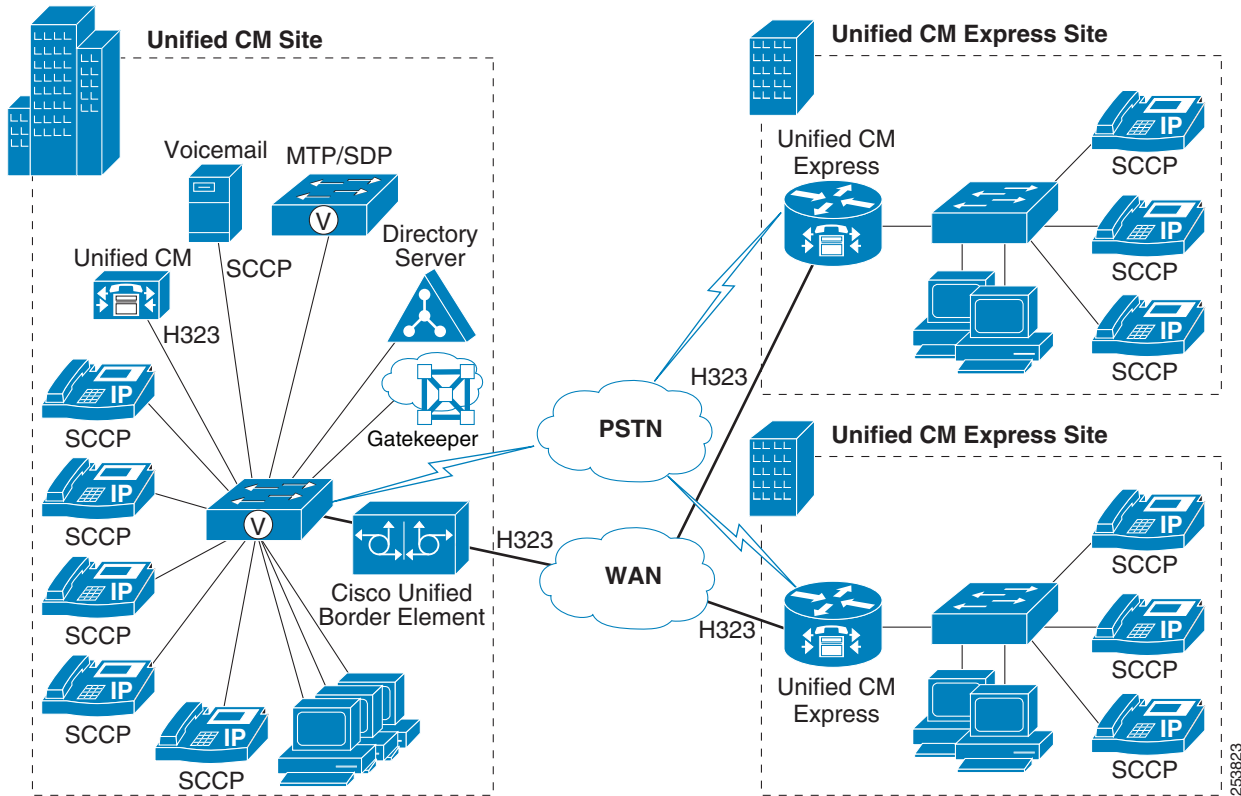
The second option is to deploy a via-zone gatekeeper. Unified CM, Unified CME, and the Cisco Unified Border Element all register with the via-zone gatekeeper as VoIP gateway devices. The via-zone gatekeeper performs dial plan resolution and bandwidth restrictions between Unified CM and Unified CME. The via-zone gatekeeper also inserts a Cisco Unified Border Element in the call path to interwork between ECS and H.450 to invoke the supplementary services. For detailed information on via-zone gateway and Cisco Unified Border Element, see the chapter on [Call Admission Control](#), page 9-1.

These two deployment options have the following differences:

- With the first option, the Cisco Unified Border Element registers with Unified CM as an H.323 gateway device; with the second option, it registers with via-zone gatekeeper as a VoIP gateway device.
- With the first option, the Cisco Unified Border Element performs dial plan resolution based on the VoIP dial-peer configurations on the Cisco Unified Border Element; with the second option, the via-zone gatekeeper performs dial plan resolution based on the gatekeeper dial plan configuration.
- With the first option, there is no call admission control mechanism that oversees both call legs; with the second option, the via-zone gatekeeper performs gatekeeper zone-based call admission control.
- With the second option, the via-zone gatekeeper can also act as an infrastructure gatekeeper for Unified CM, to manage all dial plan resolution and bandwidth restrictions between Unified CM clusters, between a Unified CM cluster and a network of H.323 VoIP gateways, or between a Unified CM cluster and a service provider's H.323 VoIP transport network.

[Figure 8-18](#) shows H.323 integration between Unified CM and Unified CME using a via-zone gatekeeper and Cisco Unified Border Element.

Figure 8-18 Multisite Deployment with Unified CM and Unified CME Using a Cisco Unified Border Element or Via-Zone Gatekeeper



Best Practices

This section discusses configuration guidelines and best practices when using the deployment model illustrated in [Figure 8-18](#) with the second deployment option (via-zone gatekeeper):

- Configure a gatekeeper-controlled H.225 trunk between Unified CM and the via-zone gatekeeper. Media termination point (MTP) resources are required over the trunk only when Unified CME tries to initiate an outbound H.323 fast-start call.
- The **Wait For Far End H.245 Terminal Capability Set (TCS)** option must be unchecked to prevent stalemate situations from occurring when the H.323 devices at both sides of the trunk are waiting the far end to send TCS first and the H.245 connection times out after a few seconds.
- Configure the Unified CM service parameter **Send H225 user info message to H225 info for Call Progress Tone**, which will make Unified CM send the H.225 Info message to Unified CME to play ringback tone or tone-on-hold.
- Use the Unified CM dial plan configuration (route patterns, route lists, and route groups) to send calls destined for Unified CME to the gatekeeper-controlled H.225 trunk.
- Register Unified CME and the Cisco Unified Border Element as H.323 gateways with the via-zone gatekeeper.

- Configure the **allow-connection h323 to h323** command on the Cisco Unified Border Element to allow H.323-to-H.323 call connections. This command is optional to configure on Unified CME. Configure **allow-connection h323 to sip** if Cisco Unity Connection is used on Unified CME.
- Supplementary services such as transfer and call forward will result in calls being media hairpinned when the two endpoints reside in the same Unified CME branch location.

**Note**

The only configuration difference between the two deployment options is that the first option requires configuring the Cisco Unified Border Element as an H.323 gateway device in Unified CM. The rest of the configuration guidelines listed above are the same for both options.

**Note**

When multiple PSTN connections exist (one for Unified CM and one for Unified CME), fully attended transfer between a Unified CM endpoint and a Unified CME endpoint to a PSTN endpoint will fail. The recommendation is to use blind transfer when using multiple PSTN connections, and it is configured under **telephony-service** as **transfer-system full-blind**.

See [Example 8-12](#) and [Example 8-13](#) for sample configurations of the Cisco Unified Border Element and Unified CME.

Design Considerations

In an H.323 deployment, Unified CME supports call transfer, call forward with H.450.2, and H.450.3 as part of the H.450 standards. However, Unified CM does not support H.450, and supplementary services such as call transfer, call forward, call hold or resume are done using the Empty Capabilities Set (ECS). Therefore, when calls are transferred or forwarded between Unified CM and Unified CME, they are hairpinned and routed with a Cisco Unified Border Element and with or without a gatekeeper, as described as the two deployment models in the previous section. This section lists some of the design considerations and best practices for Unified CM and Unified CME interoperability via H.323.

Supplementary Services Such as Call Transfer and Call Forward

Unified CME can auto-detect Unified CM, which does not support H.450, by using H.450.12 protocol to automatically discover the H.450.x capabilities. Unified CME uses VoIP hairpin routing for calls between Unified CM and Unified CME. When the call is terminated, Unified CME hairpins the call from the Unified CM phone by re-originating and routing the call as appropriate.

**Note**

When Unified CME detects that Unified CM does not support H.450, Unified CME hairpins the calls by hairpinning both signaling and media at Unified CME. This causes double the amount of bandwidth to be consumed when calls are transferred or forwarded across the WAN. (For example, if a Unified CM phone calls a Unified CME phone and the Unified CME phone transfers the call to a second Unified CM phone, Unified CME hairpins both the signaling and media even though the call is between two Unified CM phones.) To avoid this double bandwidth consumption on the WAN, Cisco recommends using the Cisco Unified Border Element to act as an H.450 tandem gateway and to allow for H.450-to-ECS mapping for supplementary services such as call transfer or call forward.

Supported Call Flows

Unified CME is a back-to-back user agent (B2BUA), thus call flows work from SCCP phone to SCCP phone and from SCCP phone to SIP phone. SIP phone calls work over H.323 trunks, but supplementary features are not supported.

Security

Unified CME provides secure signaling via TLS, and Unified CME 4.2 adds the support of media encryption via SRTP. Unified CM also supports secure signaling via TLS and secure media via SRTP. However, interworking between secure Unified CM and secure Unified CME is not supported.

Video

Observe the following design considerations when implementing video functionality with Unified CME:

- All endpoints on Unified CM and Unified CME must be configured as video-capable endpoints. The video codec and formats for all the video-capable endpoints must match.
- Unified CM and Unified CME support basic video calls; however, supplementary services such as call transfer and call forward are not supported for video calls between Unified CM and Unified CME. To support supplementary services with Unified CME, H.450 must be enabled on all Unified CMEs and voice gateways. Because Unified CM does not support H.450, video calls will revert to audio-only calls when supplementary services are needed between Unified CM phones and Unified CME phones.
- Conference calls revert to audio only.
- WAN bandwidth must meet the minimum video bit rate of 384 kbps for video traffic to traverse the WAN.
- Video basic calls are supported on SCCP phones only and are not supported on SIP phones.

H.320 Video via ISDN

Observe the following design considerations when implementing H.320 video functionality via ISDN:

- When directly connected to an H.320 endpoint via a PRI or BRI interface, Unified CME and Cisco IOS routers currently support only 128 kbps video calls.
- When H.320 is enabled on Unified CME and PSTN gateways to interwork with Unified CM, use a separate dial-peer for video calls to differentiate them from voice-only calls. Configure **bear-cap speech** under the **voice-port** configuration on Unified CME.
- H.320 does not support supplementary services.

General Design Considerations for Unified CM and Unified CME Interoperability via H.323

- Configure Unified CME to auto-detect Unified CM by using H.450.12 to hairpin the calls between Unified CM and Unified CME phones.
- For SCCP-to-SCCP calls or SCCP-to-SIP calls, an H.323 trunk can be deployed between Unified CM and Unified CME.
- Deploy Unified CME 4.0 and later releases for secure signaling with TLS, but deploy Unified CME 4.2 and later releases for secure media via SRTP. However, conferencing is not secured, nor is security interoperability supported between Unified CM and Unified CME phones.
- Deploy Unified CME 4.1 and later releases for integrated voice, video, and data transport via ISDN with support of H.320.
- Deploy video only for SCCP phones (with support of basic calls), and not for SIP phones.
- MTP functionality is not compatible with video; for video calls to work, the MTP feature must be disabled (unchecked).
- Make sure that IP connectivity between Unified CM and Unified CME works properly.
- Make sure the local video setup works correctly for each Unified CME local zone and Unified CM location (local SCCP).

- Use the existing voice dial-plan infrastructure.
- Observe the following guidelines for video traffic shaping:
 - Mark the video and audio channels of a video call with CoS 4 to preserve lip-sync and to separate video from audio-only calls.
 - Place voice and video traffic in different queues.
 - Use Priority Queuing (PQ) for voice and video traffic. Two different policies are required for voice-only calls and video (voice stream + video stream) calls based on Classifications. Voice calls are protected from video calls because the voice stream in a video call is marked the same as the video stream in the video call.
- Video should not be deployed in links with less than 768 kbps of bandwidth.
- With link speeds greater than 768 kbps and with proper call admission control to avoid oversubscription, placing video traffic in a PQ does not introduce a noticeable increase in delay to the voice packets.
- There is no need to configure fragmentation for speeds greater than 768 kbps.
- cRTP is not recommended for video packets. (Because video packets are large, cRTP is of no help with video.)
- Voice and video traffic should occupy no more than 33% of the link capacity.
- When calculating video bandwidth, add 20% to the total video data rate of the call to account for overhead.

For more details on integrating Unified CME with Unified CM via H.323, refer to the *Cisco Unified CME Solution Reference Network Design Guide*, available at

http://www.cisco.com/en/US/products/sw/voicew/ps4625/products_implementation_design_guides_list.html

Configuration Examples

The following examples illustrate some of the design considerations and best practices discussed in this section.

Example 8-12 Configuration for Cisco Unified Border Element

```
voice service voip
  allow-connections h323 to h323
  supplementary-service h450.2
  supplementary-service h450.3
  supplementary-service h450.12
  h323
    emptycapability
    h225 id-passthru
    h225 connect-passthru
h245 passthru tcsonstd-passthru
interface Loopback0
  ip address 2.1.1.1 255.255.255.0
  h323-gateway voip interface
h323-gateway voip id BORDERGW-zone ipaddr 1.1.1.1 1719
  h323-gateway voip h323-id BORDERGW
  h323-gateway voip bind srcaddr 2.1.1.1
dial-peer voice 1 voip          /* To Unified CM endpoints */
  destination-pattern 4...
  session target ras
  dtmf-relay h245-alphanumeric
  codec g729r8
```

```

no vad
dial-peer voice 1 voip          /* To Unified CME endpoints */
  destination-pattern 3...
  session target ras
  dtmf-relay h245-alphanumeric
  codec g729r8
  no vad

```

Example 8-13 Configuration for Cisco Unified CME with H.323

```

voice service voip
  h323
interface Loopback0
  ip address 3.1.1.1 255.255.255.0
  h323-gateway voip interface
  h323-gateway voip id CME-zone ipaddr 1.1.1.1 1719
  h323-gateway voip h323-id CME
  h323-gateway voip bind srcaddr 3.1.1.1
dial-peer voice 1 voip          /* To Unified CM endpoints */
  destination-pattern 4...
  session target ras
  session transport tcp
  codec g729r8                  /* Voice class can also be used */
  no vad
telephony-service
  ip source-address 3.1.1.1 port 2000
  create cnf-files
  keepalive 45
  max-conferences 8 gain -6
  moh music-on-hold.au
  transfer-system full-blind /* Used with multiple PSTN connections */
  transfer-pattern .T

```

Example 8-14 Configuration for Via-Zone Gatekeeper

```

gatekeeper
  zone local CCM-zone customer.com 1.1.1.1 outvia BORDERGW-zone
  zone local CME-zone customer.com outvia BORDERGW-zone
  zone local BORDERGW-zone customer.com
  no zone subnet CCM-zone default enable
  no zone subnet CME-zone default enable
  no zone subnet BORDERGW-zone default enable
  zone subnet CCM-zone 4.1.1.1/32 enable
  zone subnet CME-zone 3.1.1.1/32 enable
  zone subnet BORDERGW-zone 2.1.1.1/32 enable
  zone prefix CCM-zone 4...
  zone prefix CME-zone 3...
  bandwidth interzone zone CCM-zone <bandwidth value>
  bandwidth interzone zone CME-zone <bandwidth value>
  bandwidth interzone zone BORDERGW-zone <bandwidth value>
  no shutdown

```

Example 8-15 Configuration for Unified CME Video

```
voice service voip
  supplementary-service h450.12 advertise-only

  h323
    call start slow
  ...

telephony-service
  video
    maximum bit-rate <0-10000000>
    load 7970 TERM70.6-0-2-0s
    ip source-address 10.10.10.1 port 2000
    service phone videoCapability 1 !!! Enable Video Capability Service, case sensitive
  create cnf-files
  ...

ephone 1
```