



CHAPTER 8

Call Processing

Last revised on: February 13, 2008

This chapter provides design guidance for scalable and resilient call processing systems with Cisco Unified CallManager 4.2. This chapter also discusses how to choose the appropriate hardware and deployment scenario for Cisco Unified CallManager based on specific requirements that include the following:

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

This chapter focuses on the following topics:

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

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

- [Cisco Unified CallManager Platform Capacity Planning, page 8-12](#)

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

- [Gatekeeper Design Considerations, page 8-18](#)

This section explains how gatekeepers can be used in an IP Telephony 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 Cisco Unified CallManager and CallManager Express, page 8-27](#)

This section explains the H.323 integration between Cisco Unified CallManager and Cisco Unified CallManager Express in a distributed call processing deployment.

Cisco Unified CallManager Cluster Guidelines

The Cisco Unified CallManager 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 Cisco Unified CallManager 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 Cisco Unified CallManager cluster, and it provides guidelines for deploying the servers in ways that achieve the desired scale, performance, and resilience.

Hardware Platforms

Cisco Unified CallManager 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 8-1 lists the general types of servers you can use in a cluster, along with their main characteristics.

Table 8-1 *Types of Cisco Unified CallManager Servers*

Server Type	Cisco Server Model	Characteristics
Standard server (not high availability)	MCS 7815 or equivalent	<ul style="list-style-type: none"> • Single processor • Single power supply • Non-RAID hard disk
High-availability standard server	MCS 7825 or equivalent	<ul style="list-style-type: none"> • Single processor • Multiple power supplies • Single SCSI RAID hard disk array
High-performance server	MCS 7835 and MCS 7845 or equivalent	<ul style="list-style-type: none"> • Multiple processors • Multiple power supplies • Multiple SCSI RAID hard disk arrays

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

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

The server platforms have particular memory requirements, which are documented in *Product Bulletin 2864, Physical Memory Recommendations For Cisco CallManager Version 4.0 and Later*, available at

http://www.cisco.com/en/US/products/sw/voicesw/ps556/prod_bulletin0900aecd80284099.html

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 Cisco Unified CallManager 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.

**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.

Server platforms with dual Ethernet NICs can support NIC Teaming. This feature will depend on the manufacturer, and it allows a server to be connected to the Ethernet via two NICs and, hence, two cables. In some circumstances, this feature can be useful to provide additional levels of redundancy; however, careful design of the IP infrastructure is required prior to successful deployment. For additional information regarding the installation of teaming drivers on Cisco and Hewlett-Packard (HP) server platforms, refer to the documentation at

http://www.cisco.com/en/US/products/hw/voiceapp/ps378/prod_installation_guide09186a00803a0032.html

General Clustering Guidelines

The following guidelines apply to all Cisco Unified CallManager clusters:

**Note**

A cluster may contain a mix of server platforms, but all servers in the cluster must run the same Cisco Unified CallManager 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 2-17](#).

Cisco Unified CallManager Cluster Services

Within a Cisco Unified CallManager 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.

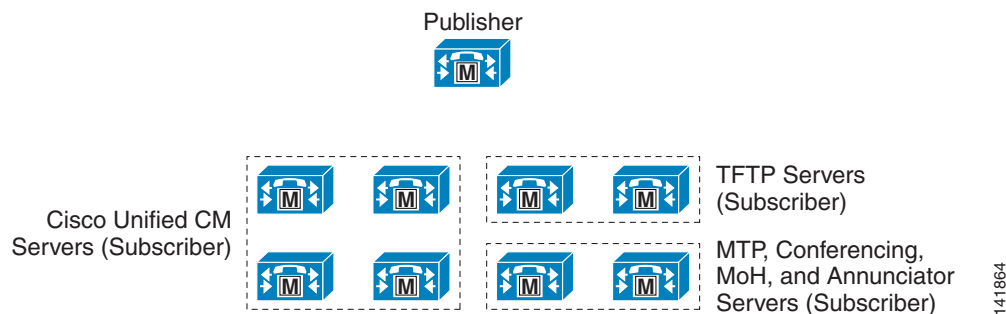
A Cisco Unified CallManager 4.2 cluster may contain as many as 20 servers, of which a maximum of eight may run the Cisco CallManager service that provides call processing. The other servers may be configured as a dedicated database publisher, dedicated Trivial File Transfer Protocol (TFTP) server, or music on hold (MoH) servers. Media streaming applications (conference bridge or media termination point) may also be installed on a separate server that registers with the cluster.

When deploying a cluster with Cisco MCS 7815 or equivalent servers, there is a minimum 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 300 phones is supported in this configuration on a Cisco MCS 7815. 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 Cisco Unified CallManager 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 Cisco Unified CallManager cluster.

Figure 8-1 Typical Cisco Unified CallManager Cluster



Intracuster Communications

There are two primary kinds of communication within a Cisco Unified CallManager cluster (intracuster communications). (See Figure 8-2.) 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. Changes made on the publisher are communicated to the subscriber databases, ensuring that the configuration is consistent across the members of the cluster, as well as facilitating spatial redundancy of the database.

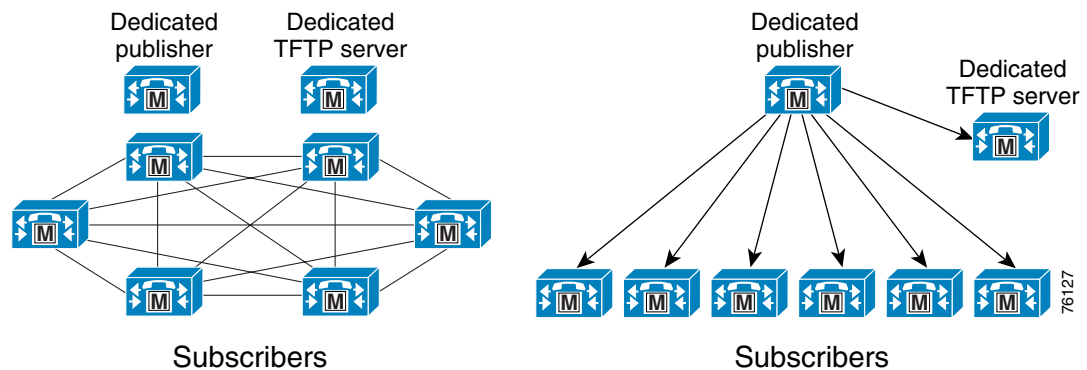
The second type of Intracuster communication is the propagation and replication of run-time data such as registration of devices, locations bandwidth, and shared media resources (see ICCS in Figure 8-2). This information is shared across all members of a cluster running the Cisco Unified CallManager Service, and it assures the optimum routing of calls between members of the cluster and associated gateways.

Lightweight Directory Access Protocol (LDAP) directory information is also replicated between all servers in a cluster. The LDAP directory is replicated by the publisher to all other servers. Cisco Unified CallManager may be integrated into a corporate LDAP directory, such as Microsoft Active Directory or Netscape Directory. The replication is then dependant on the integration method deployed and is outside the scope of this document. For more information on directory integration, refer to the chapter on [LDAP Directory Integration, page 18-1](#).

Figure 8-2 Intracluster Communications

Intra-Cluster Communication Signaling (ICCS)

SQL and directory replication



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 read and write access to the configuration database. When configuration changes are made, other members of the cluster have a read-only copy of the database. On larger systems with more than 1250 users, Cisco recommends a dedicated publisher to prevent administrative operations from affecting the telephony users. A dedicated publisher does not have any call processing services or TFTP services running on the server. Other servers run the TFTP and Cisco Unified CallManager services.

Servers in the cluster will attempt to use the publisher's database when initializing. If the publisher is not available, they will use the local read-only copy from their hard drives.

If the system is running but the publisher is not available, the following operations will not be available:

- Call forwarding changes
- Configuration changes
- Extension Mobility login or logout operations, and Device Mobility

Extension Mobility will not function without the publisher because it requires access to the read/write database. Therefore, Cisco recommends running this service on the publisher only. Device Mobility will not update the device pool settings for the phone when roaming to another site. The phone uses its home device pool settings if the publisher is not available.

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 Cisco Unified CallManager 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 read-only 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.

Cisco Unified CallManager 4.2 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 8-6](#)), 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 [Cisco Unified CallManager Platform Capacity Planning, page 8-12](#), and redundancy is covered in the section on [Call Processing Redundancy, page 8-6](#).

Call Processing Redundancy

With Cisco Unified CallManager 4.2, you can choose from the following redundancy configurations:

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

The 1:1 redundancy scheme allows upgrades with only the failover periods impacting the cluster. The 1:1 redundancy scheme enables you to upgrade the cluster using the following method:

-
- | | |
|---------------|---|
| Step 1 | Upgrade the publisher server. |
| Step 2 | Upgrade dedicated TFTP servers one at a time; reboot the server and wait for configuration files to be rebuilt before upgrading any further servers in the cluster. |
| Step 3 | Upgrade dedicated music on hold (MoH) and other media resource servers one at a time. |
| Step 4 | Upgrade the backup subscribers one at a time. This step will impact some users if 50/50 load balancing is implemented. |
| Step 5 | Fail-over the primary subscribers to their backups, and stop the Cisco CallManager Service on the primaries. |
| Step 6 | Upgrade the primaries one at a time, then re-enable the Cisco CallManager Service. |
-

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 Cisco Unified CallManager software. This factor can be important because the Intra-Cluster Communication Signaling (ICCS) protocol that communicates between subscribers can detect a different software version and shut down communications to that subscriber.

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 10,000 or more IP phones are registered on the two primary subscribers because there cannot be more than 10,000 backup registrations on a single backup subscriber.

**Note**

Before you do an upgrade, Cisco recommends that you back up the Cisco Unified CallManager 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 Cisco Unified CallManager.

Figure 8-3 Basic Redundancy Schemes

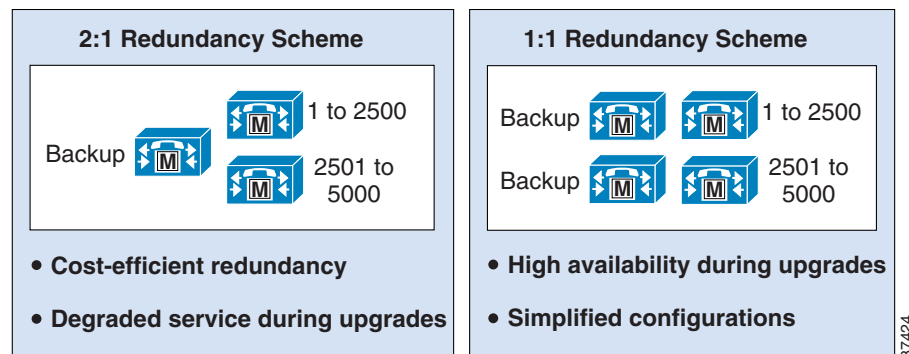
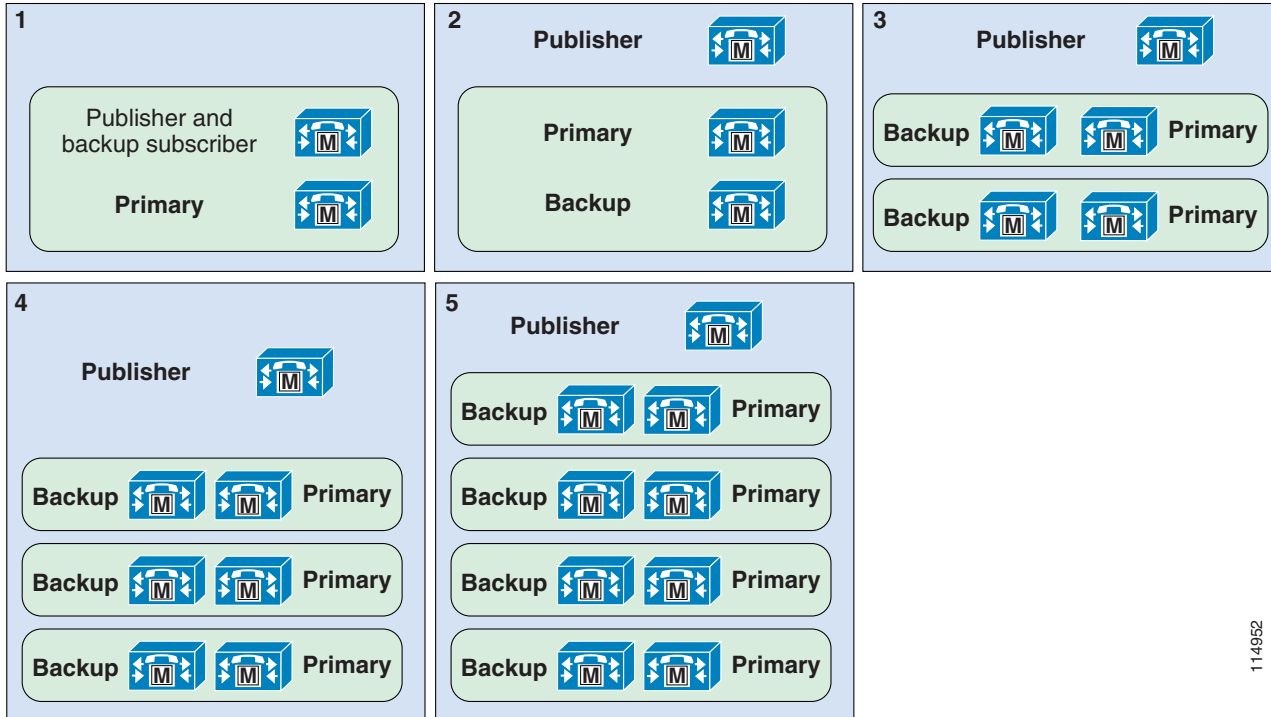


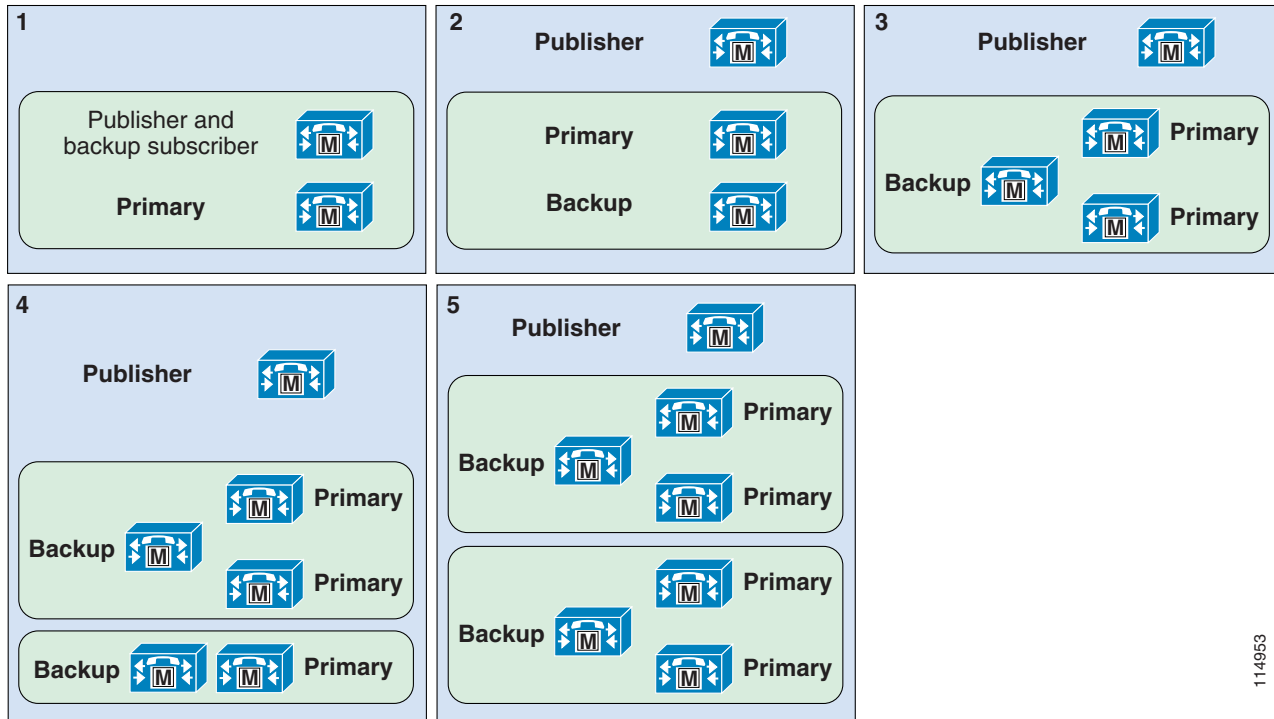
Figure 8-3 illustrates the two basic redundancy schemes available. In each case the backup server must be capable of handling the capacity of at least a single primary call processing server failure. In the 2:1 redundancy scheme, the backup might have to be capable of handling the failure of a single 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 [Cisco Unified CallManager Platform Capacity Planning, page 8-12](#).

Figure 8-4 1:1 Redundancy Configuration Options

114952

In [Figure 8-4](#) 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 8-5 2:1 Redundancy Configuration Options

114953

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 Cisco Unified CallManager 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 Cisco Unified CallManager 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 Cisco Unified CallManager 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.

By default, Cisco Unified CallManager Release 3.3 and later versions cache the configuration files in memory and do not store them on the hard drive of the TFTP server. This default setting can be changed to place the configuration files on the hard drive of the TFTP server, but doing so will impact TFTP performance. Therefore, Cisco recommends that you do not change this default setting.

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 Unified 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 Cisco Unified CallManager redundancy group of server A as the primary and server B as the backup (or secondary). The other two ports would have a Cisco Unified CallManager 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 Cisco Unified CallManager server failure.

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 Unified 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 6-17](#).)
- Conference bridges — Provide software-based conferencing for ad-hoc and meet-me conferences. (See [Audio Conferencing, page 6-7](#).)
- 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 6-12](#).)

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 a dedicated media resource server is recommended for 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 Cisco CallManager service parameters, and you should disable it on H.323 dial peers by using the **no vad** command.

Extension Mobility

When Extension Mobility is used in a cluster, the expected capacity and performance can impact the choice of the publisher hardware platform. As a user logs in or logs out from a phone, the configuration must be updated in the configuration database, the TFTP service has to regenerate the configuration file, and then the device is reset to make the change. Most of this takes place on the publisher.

Currently supported limits for Extension Mobility are as follows:

- Cisco Unified CallManager releases 3.3 and 4.0
 - 33 sequential logins/logouts per minute
- Cisco Unified CallManager release 4.1 and later 4.x releases
 - 33 sequential logins/logouts per minute
 - A Cisco MCS 7845 publisher can support 75 sequential logins/logouts per minute.

Cisco Unified CallManager Platform Capacity Planning

Many types of devices can register with Cisco Unified CallManager; 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 Cisco Unified CallManager Capacity Calculator 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 Cisco Unified CallManager 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 a performance monitor 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. For specific server memory recommendations, refer to *Product Bulletin 2864, Physical Memory Recommendations For Cisco CallManager Version 4.0 and Later*, available at http://www.cisco.com/en/US/products/sw/voicesw/ps556/prod_bulletin0900aecd80284099.html
- On supported hardware, make sure the battery-backed write cache (BBWC) is enabled with 50% write capability. This feature might require the addition of a battery to the Small Computer System Interface (SCSI) controller card.
- On MCS-7845 or equivalent servers, set the trace files location to the F: drive. This setting is a Cisco Unified CallManager service parameter.
- Reformat the trace file partition to use smaller block sizes. (Refer to the release notes for the version of the operating system you have installed.) A utility is provided for this purpose.
- A Cisco Unified CallManager 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, there are service parameters that can be modified to allow additional time for the configuration to initialize. For details on the service parameters, refer to the online help for Service Parameters in Cisco Unified CallManager Administration.

The following guidelines apply to Cisco Unified CallManager Release 4.2:

- Within a cluster, you may enable a maximum of 8 servers with the Cisco Unified CallManager Service. Other servers may be used for more dedicated functions such as TFTP, publisher, music on hold, and so forth.
- You can configure a maximum of 800 CTI connections or associations per standard server, or a maximum of 3200 per cluster if they are equally balanced among the servers.
- You can configure a maximum of 2500 CTI connections or associations per high-performance server, or a maximum 10,000 per cluster if they are equally balanced among the servers.
- Each cluster can support a maximum of 30,000 SCCP phones.
- Each cluster can support a maximum of 600 H.323 devices (gateways, trunks, and clients), digital MGCP devices, and SIP trunks

Capacity Calculations

The Cisco CallManager Capacity Tool for Cisco Unified CallManager Release 4.x enables you to calculate the capacity of the system for various configurations. The capacity planning tool is currently available to all www.cisco.com users with a login account. If your system does not meet the following guidelines, or if you consider the system to be more complex and would like to verify the capacity but you do not have access to the Cisco CallManager Capacity Tool, please contact your Cisco Systems Engineer (SE) or the Cisco Technical Assistance Center (TAC).

The Cisco CallManager Capacity Tool is available (with proper login authentication) at

<http://www.cisco.com/partner/WWChannels/technologies/resources/CallManager/>

If your system meets the following requirements, you should not need to verify your configuration with the Cisco CallManager Capacity Tool:

- System contains less than 25% of the maximum number of users for the server platforms
- Average number of lines per phone does not exceed 1.1
- Average busy hour call attempts (BHCA) per user is below 4 calls per hour
- Cluster security is not enabled
- No more than 20% trunking (5 users per trunk) either via gateways or trunks
- No more than 5% voicemail ports (20 users per voicemail port)
- No more than 20 MoH streams per server
- No CTI, JTAPI, or TAPI devices
- No more than 5% conference bridges (20 users per port on a bridge)
- No transcoders
- Only MTPs are used to support the maximum calls required for trunking
- Less than 20 locations
- System contains nothing else that is not defined above and is not an IP Phone, IP Communicator, gateway, media resource, voicemail port, or trunk

The maximum number of users Cisco Unified CallManager can support depends on the server platform, as indicated in Table 8-2.

Table 8-2 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 (All supported models) ³	300	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 servers support only N+1 redundancy and may not be a member of a cluster.

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>



Note

The maximum supported non-redundant installation on platforms that are not highly available is 500 IP phones.

Cisco CallManager Capacity Tool

The Cisco CallManager Capacity Tool requires various pieces of information to provide a calculation of the minimum size and type of servers required for a system. The information includes the type and quantity of devices, such as IP phones, gateways, and media resources. For each device type, the capacity tool also requires the average BHCA and the average utilization time. For example, if all IP phones make an average of 5 calls per hour and the average call lasts 3 minutes, then the BHCA is 5 and the utilization is 0.25. (Five calls of 3 minutes each equals 15 minutes per hour on the phone, which is 0.25 hour.)

In addition to the device information, the capacity tool also requires information regarding the dial plan, such as route patterns and translation patterns. When all the details have been entered, the capacity tool will calculate how many primary servers of the desired server type are required, as well as the number of clusters if the required capacity exceeds a single cluster.

The capacity tool replaces the previous mechanisms known as device weights, BHCA multipliers, call type multipliers, and dial plan weights.

When you use the Cisco CallManager Capacity Tool, the results will indicate the highest capacity of the various resources calculated for the configuration. The resources include hard limits, memory, processor utilization, and disk I/O activity. If you add more devices, it might look as if no more capacity is being used because the additional configuration is using resources that are below the current highest capacity indicated.

For example, if 2500 third-party controlled SCCP IP phones were added to the Cisco CallManager Capacity Tool, it would show that a single MCS 7845 subscriber is required and is at 100% capacity. If an additional 1000 phones are added, the results still indicate a single MCS 7845 subscriber is required and is at 100% capacity. This result is due to the fact that CTI capacity for this server is at 100%, and

additional IP phones do not add CTI requirements. Until the number of other devices exceeds one of the other limits, it will appear as if no additional capacity is being used. This is normal and expected behavior for the Cisco CallManager Capacity Tool.

Use the guidelines in the following sections to provide information for the capacity calculator.

Phone Calculations

The main Telephony section (not the Contact Center section) of the Capacity Tool lists the following phone types:

- SCCP Phones (Non-secure)
- Secure SCCP Phones

For each phone type, you can enter quantity, BHCA, utilization, and line appearance values, which are further explained in the following sections.

Quantity

This value is the total number of IP phones of each type that will be configured in the cluster. The quantity includes all Cisco 7900 Series IP Phones, VG248 ports, VG224, IP Communicator, and other third-party SCCP endpoint devices. The quantity must include all configured phones even if they are not active.

BHCA

This value is the average BHCA of all phones of each type. If you have shared lines across multiple phones, the BHCA should include one call for each and every phone that shares that line. Shared lines across multiple phone types will affect the BHCA for each type. (That is, one call to a shared line will be calculated as multiple calls, one to each phone that rings.) If you have different groups of phones that generate different BHCAs, use the following method to provide a BHCA value for use in the Cisco CallManager Capacity Tool.

For example, assume there are two classes of user with the following characteristics:

- 100 phones at 20 BHCA = 2,000 BHCA total
- 5000 phones at 4 BHCA = 20,000 BHCA total

The total BHCA for all phone devices in this case is 22,000.

Then divide the total BHCA by the total number of phone devices, to yield:

$$\text{Average BHCA per phone device} = 22,000 / 5,100 = 4.31 \text{ BHCA}$$

Utilization

This value is the average call utilization per phone, and it includes all the time that a call is on the phone. The actual utilization of a phone can exceed 100% if that phone allows multiple calls per line or has multiple line appearances that also are frequently utilized. Utilization is measured as a percentage of an hour. For example, a phone that is on a call for 3 minutes in the busiest hour is considered to be 5% utilized. The same method for calculating BHCA can also be used to calculate the average utilization for various groups of phone. If the phone has shared lines, then only the expected utilization for that phone should be calculated and not the actual utilization for the line that is shared. The Cisco CallManager Capacity Tool currently accepts an average utilization for all phones.

Lines Appearances

This value is the average number of lines for all the phones. If a phone has multiple appearances of the same DN in different partitions, it will be counted as multiple line appearances. Shared lines should be counted as one line, but ensure that the correct BHCA and utilization are calculated. Multiple calls per

line do not increase the line appearance count, but they do affect the BHCA and utilization calculations. For example, if it is normal for a phone to have two calls, one active and the other on hold for various amounts of time, then the utilization will be higher for that device because two calls are actually connected.

Gateways

Quantity of Gateways

The quantity of gateways is split into several entries because it depends on gateway type, as follows:

- MGCP T1/E1 gateways

This value is the total number of gateways that need to be configured in the Cisco Unified CallManager database. For example, a Cisco IOS MGCP gateway can have multiple T1s or E1s, but they would be added as a single gateway. A Cisco WS-6608 module would be added as a gateway per port that is configured as a T1 or E1 gateway, up to the maximum of eight per module.

- MGCP analogue gateways

This value is the total number of analogue gateways to be added in the Cisco Unified CallManager database. Generally, an entire module (WS-6624 or Cisco Unified CallManager) or hardware platform (Cisco IOS Router platform) would be added as a single gateway.

- H.323 gateways

An entire module or hardware platform would be added as a single gateway. This quantity does not include H.323 gateways that are not defined in Cisco Unified CallManager but are used via H.323 trunks.

Number of DS0s

This value is the total number of DS0s or analogue ports supported by each type of gateway. The number of DS0s is split between the different types of gateways as follows:

- T1 CAS:
 $24 * (\text{Number of T1 CAS spans})$
- T1 (E1) PRI:
 $(23 \text{ or } 30) * (\text{Total number of PRIs})$
- H.323 gateways
 $(\text{Total number of DS0s}) / (\text{Number of calls supported on all digital, analogue, or IP interfaces})$

BHCA

The BHCA is the average for all DS0s or analogue ports on a gateway during the busiest hour. The method for calculating this average is the same as the method used for the phone BHCA.

Utilization

This value is the average utilization of all DS0s or analogue ports during the busiest hour.

EM Profiles

Extension Mobility (EM) profiles contain line appearances that should also be included in the phone calculations. They will not affect the quantity of devices, but they will increase the average number of line appearances per phone. The BHCA and utilization of an EM user should already have been calculated for the phone they will log into.

H.323 and SIP Trunks

Quantity of Trunks

This value is the total number of trunks that will be configured in the Cisco Unified CallManager database. A gatekeeper controlled trunk is not affected by the possible number of destinations and, therefore, will be counted once for each configured gatekeeper controlled trunk. This situation is also the case for Session Initiation Protocol (SIP) trunks using a SIP proxy.

Quantity of Calls

This value is the total number of concurrent calls expected to be allowed on all trunks. The number of allowed calls will generally be controlled by locations or gatekeeper call admission control. Keep in mind that regions and codecs can also affect the total number of calls allowed.

Utilization

This value is the average utilization of all calls across all trunks. It is a percentage of the busiest hour, where 75% utilization means that calls will be active for 45 minutes per hour.

MTP Requirements

MTP requirements must also be considered separately in the calculator. If you have MTP required on an H.323 trunk, then an MTP resource is required for every concurrent call on that trunk. MTP resources are always required for SIP trunks.

CTI Route Points, Ports, and Third-Party Controlled Lines

CTI route point calculations include the total quantity, average BHCA, and utilization values. Any assigned directory numbers should be added to the Dial Plan section of the tool.

CTI ports can be used in a variety of ways by the applications controlling them. For simplicity they have been divided into two groups:

- Simple call or redirect call

A simple call is a call either to or from the port to another device, with no supplementary service being invoked. It is the typical call scenario where one party calls another, the called party answers, the call proceeds, and the parties hang up.

A redirect is a call to a CTI port that is not actually answered. Instead, it is forwarded to another destination or redirected. This is even simpler than a simple call because no call is connected, so no media is connected to the CTI port.

- Transfer or conference

These call types are typical of an IVR or auto-attendant application.

A transfer type call is characterized as a call that is connected to a CTI port, then at some point the CTI port transfers the caller without consultation (or blind transfer) to another destination.

A conference is a variation of the transfer, whereby a call connected to the CTI port is either transferred with consultation (three-way call) or is conferenced with another party for the duration of the call.

Third-Party Controlled Lines

Each application that requires third-party control or monitoring of a device or line must be counted and entered in this field of the calculator. If more than one application is monitoring or controlling the same device, it will need to be counted more than once.

Gatekeeper Design Considerations

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

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, 3600, 3700, 3800, 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/sw/iosswrel/ps5207/products_configuration_guide_book09186a00801fcee1.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 3700, 3800, or 7200 Series Router. A higher number of concurrent calls requires more memory. For the latest information on 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. Prior to Cisco Unified CallManager Release 3.3, the only method for providing gatekeeper redundancy was Hot Standby Router Protocol (HSRP); however, beginning with Cisco Unified CallManager Release 3.3, gatekeeper clustering and redundant gatekeeper trunks are also available as methods of providing gatekeeper redundancy. 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)

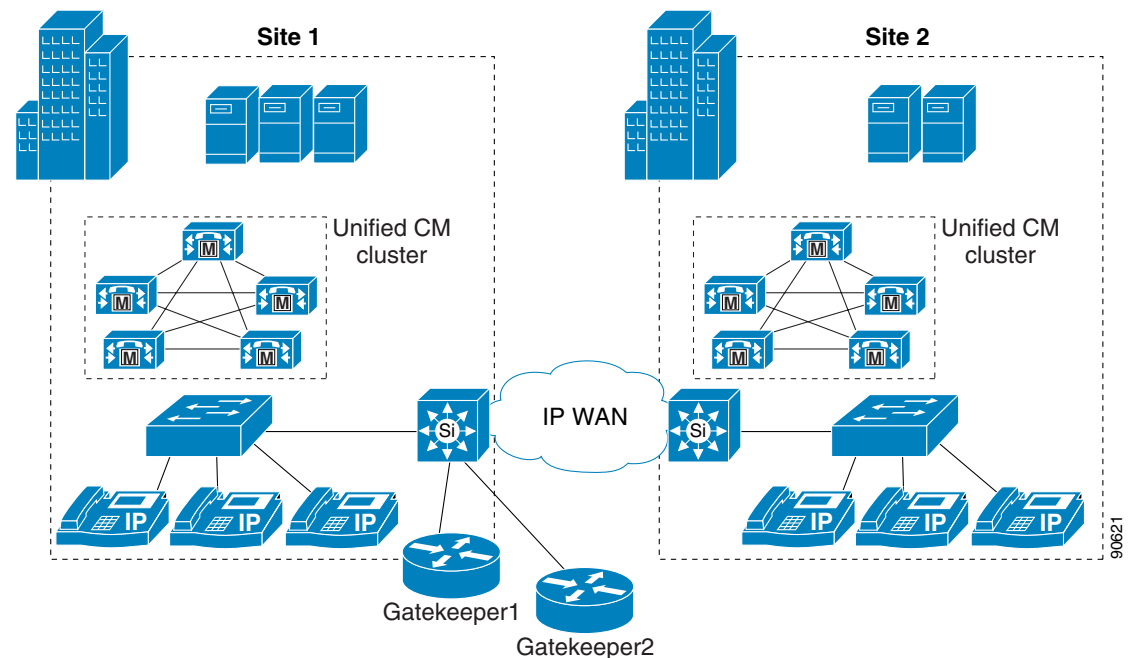
Hot Standby Router Protocol (HSRP) is the only option for gatekeeper redundancy with Cisco Unified CallManager prior to Release 3.3. HSRP does not provide the features required to build a redundant and scalable gatekeeper network, therefore it should be used only in Cisco Unified CallManager environments prior to Release 3.3.

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 8-6 show a network configuration using HSRP for gatekeeper redundancy.

Figure 8-6 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 8-6. 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 Cisco Unified CallManager cluster has a local zone configured on each router to support Cisco Unified CallManager 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** 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 Cisco Unified CallManager 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 Cisco Unified CallManager 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)

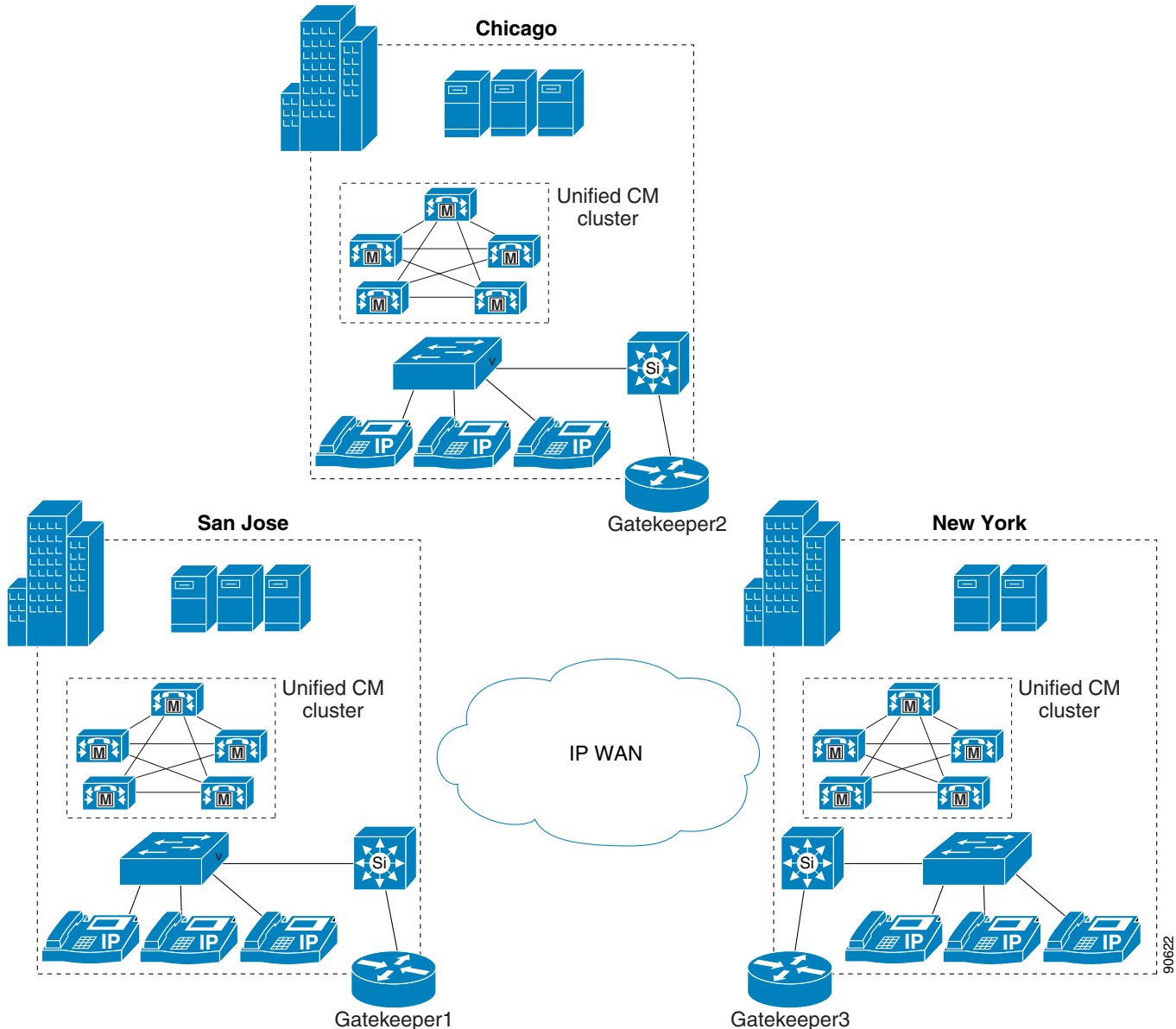
Gatekeeper clustering (alternate gatekeeper) enables the configuration of a "local" gatekeeper cluster, with each gatekeeper acting as primary for some Cisco Unified CallManager trunks and an alternate for others. Gatekeeper Update Protocol (GUP) is used to exchange state information between gatekeepers 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) and Cisco Unified CallManager Release 3.3 or later, stateful redundancy and load balancing is available. Gatekeeper clustering provides the following features:

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

Figure 8-7 shows three sites with Cisco Unified CallManager distributed call processing and three distributed gatekeepers configured in a local cluster.

Figure 8-7 Gatekeeper Clustering

In [Figure 8-7](#), 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
!
```

```

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 Cisco Unified CallManager cluster has a local zone configured to support Cisco Unified CallManager 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** 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 Cisco Unified CallManager 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 Cisco Unified CallManager trunks.

All gatekeepers in the cluster display all Cisco Unified CallManager 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  RASignalAddr  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

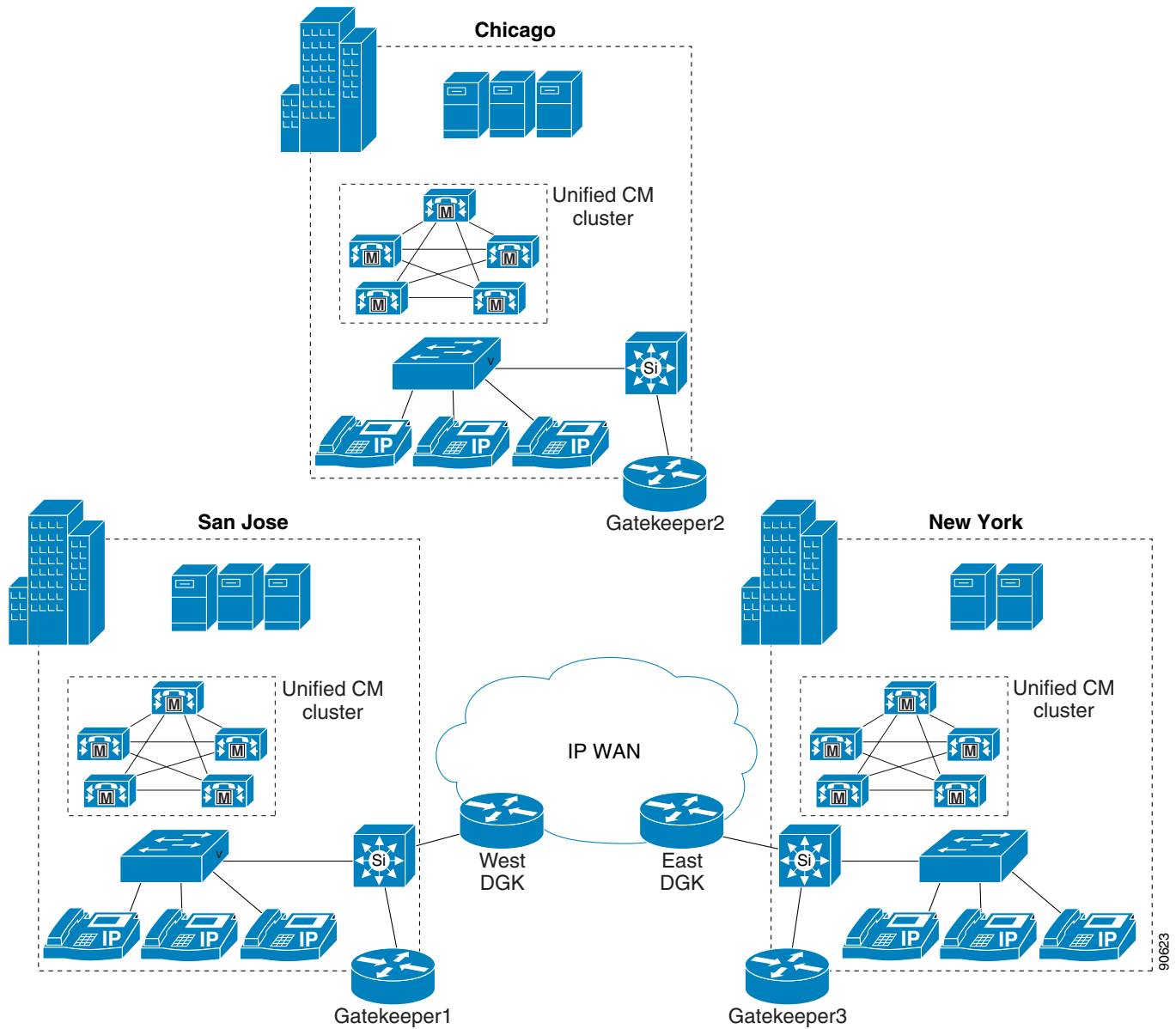
You can implement directory gatekeeper redundancy by using HSRP or by configuring multiple identical directory gatekeepers. When a gatekeeper is configured with multiple remote zones using the same zone prefix, the gatekeeper can use either of the following methods:

- 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 8-8 illustrates a Cisco Unified CallManager distributed call processing environment with two active directory gatekeepers.

Figure 8-8 Redundant Directory Gatekeepers



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

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 8-8](#).

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 Cisco Unified CallManager cluster has a local zone configured to support Cisco Unified CallManager 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** 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 Cisco Unified CallManager trunks have been configured to register with a 1# prefix.
- The **arq reject-unknown-prefix** command guards against potential call routing loops across redundant Cisco Unified CallManager trunks.

Interoperability of Cisco Unified CallManager and CallManager Express

This section explains the requirements for interoperability and internetworking of Cisco Unified CallManager with Cisco Unified CallManager Express (CME, formerly called Cisco IOS Telephony Services, or ITS) using H.323 protocol in a multisite IP telephony deployment. This section highlights the recommended deployments between phones controlled by Cisco Unified CallManager and phones controlled by CME.

Cisco CME 3.4 supports Cisco IP SCCP Phones 7902, 7905, 7910, 7912, 7920, 7935, 7936, 7940, 7960, 7970, 7971, and Cisco IP Communicator. Cisco Unified CallManager 4.2 supports CME with SIP phones.

The following section provides guidelines for achieving interoperability between Cisco Unified CallManager and Cisco Unified CallManager Express:

- [Multisite IP Telephony Deployments Using Cisco Unified CallManager and CME, with H.323 Trunks and an IP-to-IP Gateway, page 8-28](#)

For more information about CME, refer to the product documentation for Cisco Unified CallManager Express, available online at

<http://www.cisco.com>

Multisite IP Telephony Deployments Using Cisco Unified CallManager and CME, with H.323 Trunks and an IP-to-IP Gateway

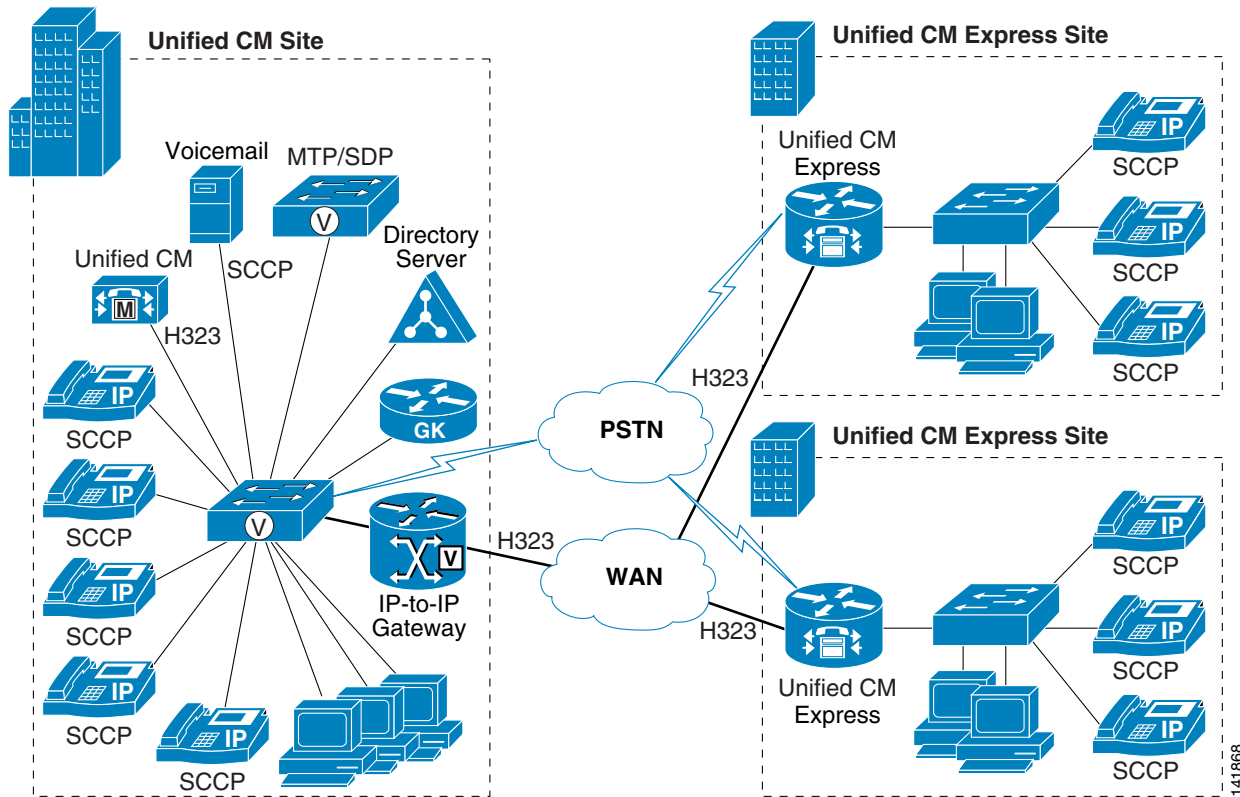
An IP-to-IP gateway is a separate router that provides a proxy (or front end) for a system that does not support H.450, such as Cisco Unified CallManager. The IP-to-IP gateway can be placed between Cisco Unified CallManager and a CME router, and it provides H.323-to-H.323 call connections to terminate and re-originate calls for transfers and forwards to endpoints that do not support H.450.

Cisco Unified CallManager Express can be directly integrated with Cisco Unified CallManager using an H.323 interface without the IP-to-IP gateway. However, because Cisco Unified CallManager does not support H.450 specifications, supplementary services such call transfer and call forward initiated from Cisco Unified CallManager phones will require the media to be hair-pinned (over the WAN in some cases) through Cisco Unified CallManager even though the Cisco Unified CallManager phones might not be involved in the call. The IP-to-IP gateway ensure that this media hair-pinning does not occur over the WAN.

The IP-to-IP gateway can also act as the PSTN gateway for the remote Cisco CME systems as well as for Cisco Unified CallManager. A separate PSTN gateway is not required in such cases.

The IP-to-IP gateway must run a Cisco IOS release that is compliant with Cisco CME 3.4 and that supports H.450, such as Cisco IOS Release 12.4(6)T or later with the IP VOICE feature set.

[Figure 8-9](#) shows an IP Telephony deployment with Cisco Unified CallManager connected to Cisco CME through an IP-to-IP gateway.

Figure 8-9 Multisite IP Telephony Deployment with Cisco Unified CallManager, CME, and an IP-to-IP Gateway

Best Practices

Follow these guidelines and best practices when using the deployment model illustrated in [Figure 8-9](#):

- Configure a gatekeeper controlled H.225 trunk between Cisco Unified CallManager and the IP-to-IP gateway, with **Media Termination Point Required** checked and **Wait For Far End H.245 Terminal Capability Set** unchecked.
- In Cisco Unified CallManager, set the service parameter **Send H225 user info message to H225 info for Call Progress Tone**.
- Use the Cisco Unified CallManager dial plan configuration (route patterns, route lists, and route groups) to send calls to the H.225 trunk that connects to the IP-to-IP gateway.
- Register Cisco CME and the IP-to-IP gateway as H.323 gateways on the gatekeeper.
- A media termination point (MTP) is required for H.450 to perform Empty Capability Set (ECS) signaling conversion. The MTP should be configured on the IP-to-IP gateway and registered to Cisco Unified CallManager. An MTP can cause media to be hair-pinned over the WAN in certain rare cases. For more details on media termination points, see the chapter on [Media Resources](#), [page 6-1](#).
- Configure the **allow-connection h323 to h323** command on the IP-to-IP gateway to allow H.323-to-H.323 call connections. Remote CME routers need not have this command enabled.

**Note**

When using H.323 trunks between CME and Cisco Unified CallManager, configure the **allow-connection h323 to h323** command on CME to allow H.323-to-H.323 hair-pinned call connections. The Cisco Unified CallManager Express auto-detection feature (starting with CME 3.1) disables all H.450-based communications on the interface connected to Cisco Unified CallManager. To complete the call transfer or forward between Cisco Unified CallManager and CME phones, an H.323-to-H.323 connection is required.

- Define VoIP dial peers on the IP-to-IP gateway to route calls to the Cisco Unified CallManager and CME endpoints.
- Define VoIP dial peers on CME to forward calls, destined for Cisco Unified CallManager endpoints, to the IP-to-IP gateway.
- Supplementary services such as transfer and call forward will result in calls being media hair-pinned when the two endpoints reside in the same CME branch location.

**Note**

When multiple PSTN connections exist (one for Cisco Unified CallManager and one for CME), fully attended transfer between a Cisco Unified CallManager endpoint and a 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**.

[Example 8-11](#) lists a sample configuration for the IP-to-IP gateway.

Example 8-11 Configuration for IP-to-IP Gateway

```
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 tcsnonstd-passthru
dial-peer voice 1 voip          /* To Cisco Unified CallManager endpoints */
  destination-pattern xxxx
  session target ipv4:y.y.y.y
  dtmf-relay h245-alphanumeric
  codec g729r8
  no vad
dial-peer voice 1 voip          /* To Cisco Unified CallManager endpoints */
  destination-pattern zzzz
  session target ras            /* "ras" if gatekeeper is used, otherwise "ipv4:a.b.c.d" */
  dtmf-relay h245-alphanumeric
  codec g729r8
  no vad
```

Example 8-12 lists a sample configuration for the in this H.323 deployment model.

Example 8-12 Configuration for Cisco CME 3.4 with H.323

```
voice service voip
  h323
interface FastEthernet0/1
  ip address 10.10.10.23 255.255.255.0
  h323-gateway voip interface
  h323-gateway voip id cme ipaddr 10.10.10.30 1719
dial-peer voice 1 voip          /* To Cisco Unified CallManager endpoints */
  destination-pattern xxxx
  session target ras
  session transport tcp
  codec g729r8                 /* Voice class can also be used */
  no vad
telephony-service
  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-blind /* Used with multiple PSTN connections */
  transfer-pattern .T
```

