Call Processing

Revised: March 1, 2018

The handling and processing of voice and video calls is a critical function provided by IP telephony systems. This functionality is handled by some type of call processing entity or agent. Given the critical nature of call processing operations, it is important to design unified communications deployments to ensure that call processing systems are scalable enough to handle the required number of users and devices and are resilient enough to handle various network and application outages or failures.

This chapter provides guidance for designing scalable and resilient call processing systems with Cisco call processing products. These products include Cisco Unified Communications Manager (Unified CM) and Cisco Unified Communications Manager Express (Unified CME). The discussions focus predominately on the following factors:

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

Specifically, this chapter focuses on the following topics:

- Call Processing Architecture, page 9-2

  This section discusses general call processing architecture and the various call processing hardware options. This section also provides information on Unified CM clustering.

- High Availability for Call Processing, page 9-13

  This section examines high availability considerations for call processing, including network redundancy, server redundancy, and load-balancing.

- Capacity Planning for Call Processing, page 9-23

  This section provides an overview of sizing for call processing deployments.

- Design Considerations for Call Processing, page 9-26

  This section provides a summarized list of high-level design guidelines and best practices for deploying call processing.

- Computer Telephony Integration (CTI), page 9-28

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

- Integration of Multiple Call Processing Agents, page 9-36

  This section discusses the integration of multiple call processing agents, which is typically done with Cisco Unified CM Session Management Edition (SME). It also covers direct integration of Cisco Unified CM with Cisco Unified Communications Manager Express (Unified CME).

What's New in This Chapter

Table 9-1 lists the topics that are new in this chapter or that have changed significantly from previous releases of this document.

Table 9-1 New or Changed Information Since the Previous Release of This Document

New or Revised Topic	Described in	Revision Date
Minor updates and corrections	Various sections of this chapter	March 1, 2018

Call Processing Architecture

In order to design and deploy a successful Unified Communications system, it is critical to understand the underlying call processing architecture that provides call routing functionality. This functionality is provided by the following Cisco call processing agents:

- Cisco Unified Communications Manager (Unified CM)

  Cisco Unified CM provides call processing services for small to very large single-site deployments, multi-site centralized call processing deployments, and/or multi-site distributed call processing deployments. Unified CM is at the core of a Cisco Collaboration solution, and it serves as a foundation to deliver voice, video, TelePresence, IM and presence, messaging, mobility, web conferencing, and security.

  Access to the enterprise collaboration network and to Unified CM from the internet to enable remote access and business-to-business secure telepresence and video communications, is also available through different collaboration edge solutions such as VPN and Cisco Expressway.

- Cisco TelePresence Video Communication Server (VCS)

  Cisco TelePresence VCS is a video application that provides video endpoint registration, call processing, and bandwidth management for SIP and H.323 endpoints. VCS acts as a SIP registrar, a SIP proxy server, an H.323 gatekeeper, and a SIP-to-H.323 gateway server to provide interworking between SIP and H.323 devices. Cisco TelePresence VCS also provides external communications using NAT/firewall traversal when combined with the VCS Expressway.

  Cisco recommends deploying Unified CM as the main call processing agent for all endpoints, including TelePresence endpoints and room-based TelePresence conferencing systems that support SIP, and use VCS only for full-featured interoperability with H.323 telepresence endpoints or integration with third-party video endpoints. This is to avoid the dial plan and call admission control complexities that dual call control introduces. Therefore, this chapter does not provide many details on VCS. For more information on VCS, refer to the Cisco Collaboration System 10.*x* SRND or the Cisco VCS product documentation.

- Cisco Business Edition 4000 and Cisco Unified Communications Manager Express (Unified CME)

  Cisco Business Edition 4000 (BE4K) is a new on-premises, completely cloud-managed audio telephony platform optimized for small to medium businesses. Powered by Cisco Unified Communications Manager Express (Unified CME) and Cisco Unity Express Virtual (vCUE), Business Edition 4000 provides affordable integrated IP telephony and voicemail solutions for up to 200 devices. As with Cisco Business Edition 6000 and 7000, Business Edition 4000 simplifies the quoting and ordering process and allows for rapid deployments by providing pre-configured hardware, pre-installed licenses, and pre-loaded Cisco Collaboration applications.

  Cisco Business Edition 4000 and Cisco Unified CME provide call processing services for small single-site deployments and larger distributed multi-site deployments. Cisco Unified CME also provides call processing services for deployments in which a local call processing entity at a remote site is needed to provide backup capabilities for a centralized call processing deployment of Cisco Unified CM.

Cisco Unified Communications Manager (Unified CM) and Cisco TelePresence Video Communication Server (VCS) are available as standard Cisco Collaboration products or through Cisco Business Edition 6000 and Cisco Business Edition 7000, which are packaged collaboration solutions that include call processing services and other services such as messaging, conferencing, and contact center.

The Cisco Business Edition 6000 and 7000 solutions simplify the quoting/ordering process and accelerate deployments by providing pre-configured hardware, pre-installed licensed hypervisor, and pre-loaded and/or pre-installed Cisco Collaboration applications. Cisco Business Edition 6000M and Cisco Business Edition 6000H are targeted for deployments with up to 1,000 users. Cisco Business Edition 7000 is targeted for deployments with more than 1,000 users. The design and sizing of the Cisco Collaboration applications have been simplified with Cisco Business Edition 6000. With Cisco Business Edition 7000, however, normal Unified CM design and sizing guidelines apply.

Call Processing Virtualization

Virtualization enables multiple Cisco Collaboration "servers" or "virtual machines" to run on one physical server. The Cisco Collaboration servers or virtual machines are also referred as VMs, nodes, or instances in this document.

This architecture has obvious benefits over traditional deployments where the applications are running directly on the hardware platform. For example, costs (such as server, electricity, cooling, and rack space costs) can be reduced significantly, and the operation and maintenance of the hardware platforms can be simplified. Virtualization is enabled by a hypervisor that is installed directly on the physical server and that manages the virtual machines. The hypervisor that is required with Cisco Collaboration is the VMware ESXi Hypervisor.

Each virtual machine has associated virtual hardware resources such as virtual CPU, virtual memory, and virtual disk. Those resources are defined for each Collaboration application in predefined templates that are distributed through an Open Virtualization Archive (OVA), an open standards-based method for packaging and distributing virtual machine templates. For many of the Cisco Collaboration applications, in order to provide different capacity options, several VM configuration options are available when deploying an OVA. OVAs must be used when installing a Cisco Collaboration application, not only to define the correct virtual hardware resources but also to ensure that the virtual disks are not misaligned with the host physical disks, which would impact the storage performance.

The virtualization support for the Cisco Collaboration call processing agents is as follows:

- Cisco Unified CM runs only as a virtual application; it cannot be deployed directly on a Cisco UCS server, for example.
- Cisco Unified CME runs within the Cisco IOS or IOS-XE software on Cisco Integrated Services Routers and does not support virtualization.
- Cisco Business Edition 4000 runs within the Cisco IOS-XE software on a Cisco 4321 Integrated Services Router (ISR) and does not support virtualization.

For more information on the considerations for designing and deploying virtualization of Cisco Unified Communications applications, refer to the information available at

https://www.cisco.com/go/virtualized-collaboration

Call Processing Hardware

There are three types of hardware options for Cisco Unified CM: Tested Reference Configurations, Cisco Business Edition 6000 and 7000, and Specifications-based hardware.

- Tested Reference Configuration (TRC)

  TRCs are selected hardware configurations based on the Cisco Unified Computing System (UCS) servers. They have a fixed hardware configuration, and they are tested and validated with Cisco Collaboration applications for specific advertised performance, capacity, and application co-residency scenarios. They are intended for customers who require explicitly validated infrastructure and/or customers who are not necessarily experienced with virtualization.

  The hardware configuration for each TRC is well defined, and allowed deviation from this hardware configuration is very limited. For example, changing the CPU model or number of cores, or changing the RAID configuration of a TRC, would change the server qualification, and the server would not be considered as a TRC anymore but rather as specifications-based hardware.

- Cisco Business Edition 6000 and 7000

  Cisco Business Edition 6000 and 7000 are packaged collaboration solutions that include the hardware platform, virtualization software, and Cisco applications. The hardware platform is pre-configured (for instance, firmware, drivers and the RAID controller are pre-configured at the factory). Just like the TRC, the hardware platform is tested and validated with Cisco Collaboration applications for specific capacity and performance.

  Cisco Business Edition 6000 is available with two hardware platform options: BE6000M and BE6000H. Cisco Business Edition 7000 also is available with two hardware platform options: BE7000M and BE7000H.

  For more details on the TRC and Cisco Business Edition 6000 and 7000 hardware platforms, refer to the documentation at https://www.cisco.com/go/virtualized-collaboration.

- Specifications-based hardware

  Specifications-based hardware (sometimes simply referred as "specs-based") provides more flexible hardware configurations. For example, it allows you to select a platform based on a Cisco UCS TRC and to change the CPU model, number of cores, and RAID configuration, and/or to use an iSCSI or NAS storage. If desired, it also allows you to use a server vendor other than Cisco. Any specifications-based hardware server, whether it is Cisco or not, must be listed in the following *VMware Compatibility Guide*:

  https://www.vmware.com/resources/compatibility/search.php

While specification-based hardware provides more flexible hardware configurations, some requirements must still be met. For example, there are requirements around the CPU model and minimum CPU speed, and vCenter is required in order to collect logs and statistics. With specifications-based hardware, it is important to understand that the hardware configuration has not been explicitly validated by Cisco with Cisco Collaboration applications. Therefore Collaboration applications cannot provide prescriptive guidance on hardware compatibility and cannot be guaranteed, and performance of the Cisco Collaboration applications is for guidance only (refer to the Troubleshooting TechNote at https://www.cisco.com/c/en/us/support/docs/voice-unified-communications/unified-communications-system/115955-uc-specs-tshoot-00.html).

To obtain guidance on the performance of Cisco Collaboration applications with specifications-based hardware, use the TRCs or Cisco Business Edition 6000 and 7000 hardware platforms as references. For more information, refer to the documentation at https://www.cisco.com/go/virtualized-collaboration.

Cisco Unified CME runs on Cisco Integrated Services Routers (ISR) such as the Cisco 2900, 3900, or 4000 Series ISRs. Cisco Unified CME does not run as a virtual application or as part of Cloud Services Router 1000V. Cisco Business Edition 4000 runs on Cisco 4321 Integrated Services Router (ISR) only. The voicemail powered by Cisco Unity Express Virtual (vCUE) can be run only within the Service Container of the Cisco ISR 4321 and not on a UCS-E module.

Determining the appropriate call processing type and platform for a particular deployment will depend on the scale, performance, and redundancy required. In general, Unified CM provides a very wide range of capacity options and higher availability, while Cisco Unified CME provides lower levels of capacity and redundancy. For specifics regarding redundancy and scalability, see the sections on High Availability for Call Processing, page 9-13, and Capacity Planning for Call Processing, page 9-23.

Unified CM Cluster Services

While Cisco Unified CME is a standalone call processing application, Unified CM supports the concept of clustering. The Unified CM architecture enables a group of server nodes to work together as a single call processing entity or IP PBX system. This grouping of server nodes is known as a *cluster*. A cluster of Unified CM server nodes may be distributed across an IP network, within design limitations, allowing for spatial redundancy and, hence, resilience to be designed into the Unified Communications System.

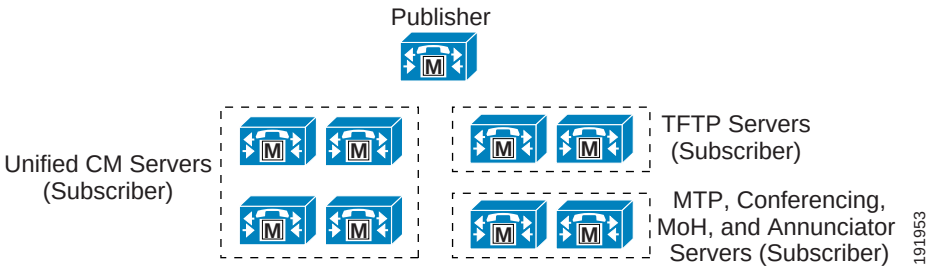
Within a Unified CM cluster, there are server nodes that provide unique services. Each of these services can coexist with others on the same server node. For example, in a small system it is possible to have a single server node providing database services, call processing services, and media resource services. As the scale and performance requirements of the cluster increase, many of these services should be moved to dedicated server nodes.

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

Cluster Server Nodes

Figure 9-1 illustrates a typical Unified CM cluster consisting of multiple server nodes. There are two types of Unified CM server nodes, publisher and subscriber. These terms are used to define the database relationship during installation.

Figure 9-1 Typical Unified CM Cluster



Publisher

The publisher is a required server node in all clusters, and as shown in Figure 9-1, there can be only one publisher per cluster. This server node is the first to be installed and provides the database services to all other subscribers in the cluster. The publisher node is the only server node 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 provide call processing or TFTP services running on the node. Instead, other subscriber nodes within the cluster provide these services.

The choice of the VM configuration for the publisher should be based on the desired scale and performance of the cluster. Cisco recommends that the publisher have the same server node performance capability as the call processing subscribers.

Subscriber

When the software is installed initially, only the database and network services are enabled. All subscriber nodes subscribe to the publisher to obtain a copy of the database information. However, in order to reduce initialization time for the Unified CM cluster, all subscriber nodes in the cluster attempt to use their local copy of the database when initializing. This reduces the overall initialization time for a Unified CM cluster. All subscriber nodes rely on change notification from the publisher or other subscriber nodes in order to keep their local copy of the database updated.

As shown in Figure 9-1, multiple subscriber nodes can be members of the same cluster. Subscriber nodes include Unified CM call processing subscriber nodes, TFTP subscriber nodes, and media resource subscriber nodes that provide functions such as conferencing and music on hold (MoH).

Call Processing Subscriber

A call processing subscriber is a server node that has the Cisco CallManager Service enabled. Once this service is enabled, the node 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. As shown in Figure 9-1, multiple call processing subscribers can be members of the same cluster. In fact, Unified CM supports up to eight call processing subscriber nodes per cluster.

TFTP Subscriber

A TFTP subscriber or server node performs two main functions as part of the Unified CM cluster:

- The serving of files for services, including 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, which are usually signed and in some cases encrypted before being available for download

The Cisco TFTP service that provides this functionality can be enabled on any server node 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 subscriber node to the TFTP service, as shown in Figure 9-1, for a cluster with more than 1250 users or any features that cause frequent configuration changes.

Cisco recommends that you use the same VM configuration for the TFTP subscribers as used for the call processing subscribers.

Media Resource Subscriber

A media resource subscriber or server node provides media services such as conferencing and music on hold to endpoints and gateways. These types of media resource services are provided by the Cisco IP Voice Media Streaming Application service, which can be enabled on any server node in the cluster.

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 7-17.)
- Annunciator service — Provides announcements in place of tones to indicate incorrectly dialed numbers or call routing unavailability. (See Annunciator, page 7-15.)
- Conference bridges — Provide software-based conferencing for instant and permanent conferences. (See Transcoding, page 7-5.)
- Media termination point (MTP) services — Provide features for H.323 clients, H.323 trunks, and Session Initiation Protocol (SIP) endpoints and trunks. (See Media Termination Point (MTP), page 7-7.)

Because of the additional processing and network requirements for media resource services, it is essential to follow all guidelines for running media resources within a cluster. Generally, Cisco recommends non-dedicated media resource subscribers for multicast MoH and annunciator services, but dedicated media resource subscribers as shown in Figure 9-1 are recommended for unicast MoH as well as large-scale software-based conferencing and MTPs unless those services are within the design guidelines detailed in the chapter on Media Resources, page 7-1.

Additional Cluster Services

In addition to the specific types of subscriber nodes within a Unified CM cluster, there are also other services that can be run on the Unified CM call processing subscriber nodes to provide additional functionality and enable additional features.

Computer Telephony Integration (CTI) Manager

The CTI Manager service acts as a broker between the Cisco CallManager service and TAPI or JTAPI integrated applications. This service is required in a cluster for any applications that utilize CTI. The CTI Manager service provides authentication of the CTI application and enables the application to monitor and/or control endpoint lines. CTI Manager can be enabled only on call processing subscribers, thus allowing for a maximum of eight nodes running the CTI Manager service in a cluster.

For more details on CTI Manager, see Computer Telephony Integration (CTI), page 9-28.

Unified CM Applications

Various types of application services can be enabled on Unified CM, such as Cisco Unified CM Assistant, Extension Mobility, and Web Dialer. For detailed design guidance on these applications, see the chapter on Cisco Unified CM Applications, page 18-1. The Cisco IM and Presence service can also be added (see the chapter on Collaboration Instant Messaging and Presence, page 20-1).

Mixing Unified CM VM Configurations

Mixing VM configurations within a Unified CM cluster is allowed, but Cisco recommends using the same VM configuration for all Unified CM nodes in a cluster. Cisco also recommends that the VM configuration used for the Unified CM publisher should not be smaller than any other Unified CM VM configuration used in the same cluster and that the VM configuration used for the backup subscribers should not be smaller than the VM configuration used for the primary subscribers.

When mixing VM configurations within a cluster, differences in capacity between the various VM configurations must be considered because the supported overall cluster capacity is limited by the cluster capacity corresponding to the smallest VM configuration within the cluster.

For example, if you mix one Unified CM call processing pair using the 7.5k VM configuration and two Unified CM call processing pairs using the 10k VM configuration, the overall cluster capacity that is supported corresponds to the cluster capacity of all nodes using the 7.5k VM configuration. With 3 call processing pairs in this example, the cluster capacity is limited to 22.5k endpoints (3 ∗ 7,500). To overcome this cluster capacity limitation, one option is to deploy separate clusters and connect those clusters with SIP trunks.

Mixing Hardware Platforms and Business Edition Platforms

Mixing different types of hardware platforms within a Unified CM cluster is also allowed, but because all VM configurations are not supported on all server hardware, this might result in mixing VM configurations and therefore might impact the overall cluster capacity. (See Mixing Unified CM VM Configurations, page 9-8, for details.) In addition to that, if Business Edition 6000 is part of the platform mix, the rules specific to the Business Edition 6000 solution must be taken into consideration.

Example 9-1 Mixing BE6000M and BE7000

Unified CM deployed as part of BE6000M is limited to 1,000 users and 1,200 devices, regardless of the number of cluster nodes. Adding other nodes is possible, whether or not they are part of BE7000 and whether or not those additional nodes are using larger VM configurations. It can provide redundancy and/or geographic distribution, but it does not increase the cluster capacity because of the BE6000 sizing rules. The Unified CM cluster capacity with BE6000M is still limited to 1,000 users and 1,200 devices when nodes are added. Similar restrictions apply with BE6000H, which is limited to a maximum of 1,000 users and 2,500 devices, regardless of the node count.

Example 9-2 Mixing Small TRC and BE7000

With the Small Tested Reference Configuration (TRC) that is not part of the Cisco Business Edition 6000M or 6000H solution, while the capacity of a node is limited to 1,000 users or devices, the capacity of the Unified CM cluster is not limited to 1,000 users or devices. If you add multiple nodes in a cluster, more than 1,000 users and devices can be supported. But on the Small TRC hardware platform, only the 1k-user VM configuration is supported. So if some Unified CM nodes running on the Small TRC and some running on BE7000 are mixed in the same cluster (as mentioned in the section on Mixing Unified CM VM Configurations, page 9-8), the overall cluster capacity that is supported is limited by the cluster capacity corresponding to the node using the smallest VM configuration – the 1,000-user VM configuration in this case. For example, if one Unified CM call processing pair is running on the Small TRC (with the 1k-user VM configuration) and another Unified CM call processing pair is running on the BE7000, the supported Unified CM cluster capacity is limited to 2,000 users and/or devices (2∗1,000), even if Unified CM VM configurations larger than 1k are deployed on BE7000.

Mixing servers from different vendors is allowed, but this would be under the specifications-based hardware policy, and Unified CM performance is not guaranteed on this type of platform mix.

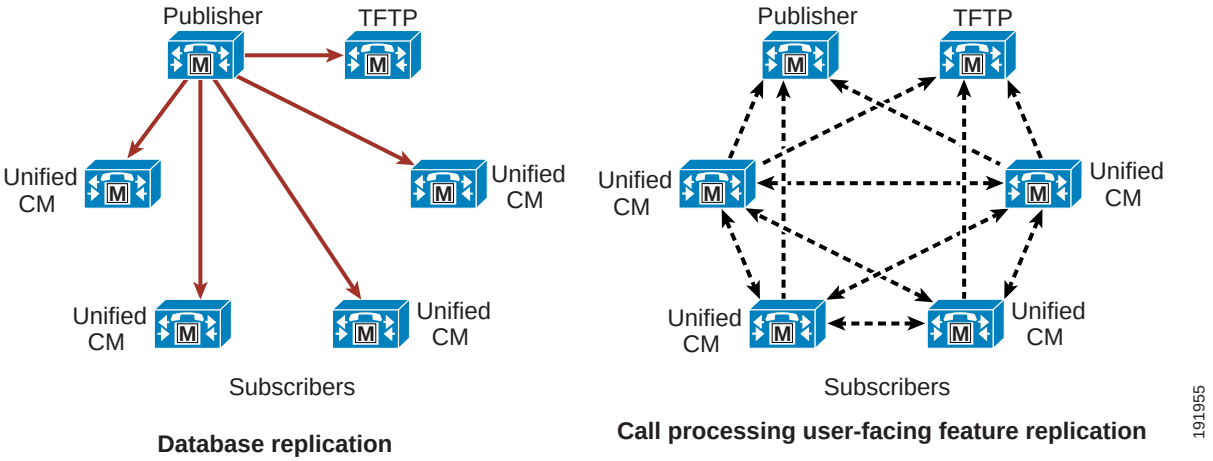
Intracluster Communications

There are two primary kinds of intracluster communications, or communications within a Unified CM cluster (see Figure 9-2 and Figure 9-3.) The first is a mechanism for distributing the database that contains all the device configuration information (see “Database replication” in Figure 9-2). The configuration database is stored on a publisher node, and a copy is replicated to the subscriber nodes 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 nodes to which an end-user device is registered. The subscriber nodes then replicate these database modifications to all the other nodes in the cluster, thus providing redundancy for the user-facing features. (See "Call processing user-facing feature replication" in Figure 9-2.) These features include:

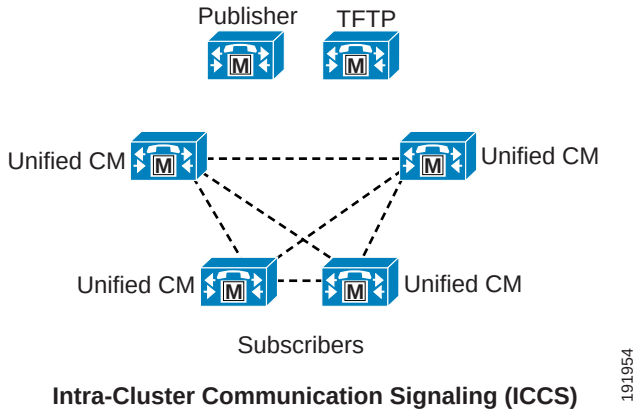
- Call Forward All (CFA)
- Message waiting indicator (MWI)
- Privacy Enable/Disable
- Extension Mobility login/logout
- Hunt Group login/logout
- Device Mobility
- Certificate Authority Proxy Function (CAPF) status for end users and applications users
- Credential hacking and authentication

Figure 9-2 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 9-3). This information is shared across all members of a cluster running the Cisco CallManager Service (call processing subscribers), and it ensures the optimum routing of calls between members of the cluster and associated gateways.

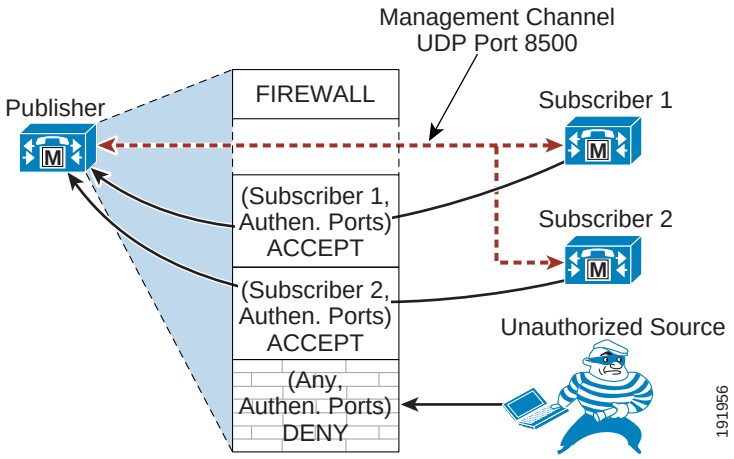
Figure 9-3 Intra-Cluster Communication Signaling (ICCS)



Intracluster Security

Each server node in a Unified CM cluster runs an internal dynamic firewall. The application ports on Unified CM are protected by source IP filtering. The dynamic firewall opens these application ports only to authenticated or trusted server nodes. (See Figure 9-4.)

Figure 9-4 Intracluster Security



This security mechanism is applicable only between server nodes 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 server nodes. During the installation process, a subscriber node is authenticated to the publisher using a pre-shared key authentication mechanism. The authentication process involves the following steps:

1. Install the publisher node using a security password.
2. Configure the subscriber node on the publisher by using Unified CM Administration.
3. Install the subscriber node using the same security password used during publisher server installation.
4. After the subscriber is installed, the server node 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 installation 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 nodes from the publisher. All the subscribers establish a management channel with each other, thus creating a mesh topology.

General Clustering Guidelines

The following guidelines apply to all Unified CM clusters:

- Cisco recommends using the same VM configuration for all nodes in a cluster. Mixing Unified CM VM configurations is allowed, but there are design implications and limits. For more details, refer to the section on Mixing Unified CM VM Configurations, page 9-8.
- Under normal circumstances, place all members of the cluster 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 10-43.
- A Unified CM cluster may contain as many as 20 server nodes, of which a maximum of eight call processing subscribers (nodes running the Cisco CallManager Service) are allowed. The other server nodes within the cluster may be configured as a dedicated database publisher, dedicated TFTP subscriber, or media resource subscriber.
- When deploying a two-node cluster, Cisco recommends that you do not exceed 1250 users in the cluster. Above 1250 users, a dedicated publisher and separate server nodes for primary and backup call processing subscribers is recommended.
- Business Edition 6000 provides a single instance of Unified CM (a Unified CM publisher that also handles call processing). Additional Business Edition 6000 server(s) may be deployed to provide subscriber redundancy either in an active/standby or load balancing fashion for Unified CM as well as some other co-resident applications. However, adding new nodes and new hardware platforms does not increase capacity. For example, the user and device capacities do not increase.
- Each Unified CM node instance can be a publisher node, call processing subscriber node, TFTP subscriber node, or media resource subscriber node. Only a single publisher node per cluster is supported.
- With virtualization, Unified CM no longer supports the Cisco Messaging Interface (CMI) service for Simplified Message Desk Interface (SMDI) integrations, fixed MoH audio source integration for live MoH audio feeds using the audio cards (MOH-USB-AUDIO=), or flash drives to these servers. The following alternate options are available:
  - For MoH live audio source feed, consider using Cisco IOS-based gateway multicast MoH for live audio source connectivity.
  - For saving system install logs, use virtual floppy softmedia.
  - There is no alternate option for the Cisco Messaging Interface (CMI) service for Simplified Message Desk Interface (SMDI) integrations.

High Availability for Call Processing

You should deploy the call processing services within a Unified Communications System in a highly available manner so that a failure of a single call processing component will not render all call processing services unavailable.

Hardware Platform High Availability

You should select the call processing platform based not only on the size and scalability of a particular deployment, but also on the redundant nature of the platform hardware.

When possible, choose platforms with dual power supplies to ensure that a single power supply failure will not result in the loss of a platform. Plug platforms with dual power supplies into two different power sources to avoid the failure of one power circuit causing the entire platform to fail. The use of dual power supplies combined with the use of uninterruptible power supply (UPS) sources will ensure maximum power availability. In deployments where dual power supply platforms are not feasible, Cisco still recommends the use of a UPS in situations where building power does not have the required level of power availability.

Providing hardware platform high availability is even more critical when deploying virtualization because a platform failure could result in the failure of all the virtual machines running on that hardware platform. When possible, avoid running multiple instances of the same application that have similar functions on the same physical server; instead, distribute those virtual machines across multiple servers and even across multiple chassis if possible when using Cisco UCS B-Series Blade Servers.

Network Connectivity High Availability

Connectivity to the IP network is also a critical consideration for maximum performance and high availability. With Cisco Unified CME, use a minimum of two ports to connect to the network. With Unified CM, high availability for the network connectivity is attained at the host level by configuring the hypervisor virtual switch with multiple uplinks and thus by using multiple physical ports on the hardware platform. Therefore, a single virtual NIC defined in the OVA setting is sufficient. If you are using the VMware vSphere virtual switch, for example, configure NIC teaming for the switch uplinks. Also connect those multiple ports to a minimum of two upstream switches to provide resiliency if an upstream switch fails.

Connect platforms to the network at the highest possible speed to ensure maximum throughput, typically 1 Gbps or even 10 Gbps when using the UCS B-Series platform. Ensure that platforms are connected to the network using full-duplex.

In addition to speed and duplex of IP network connectivity, equally important is the resilience of this network connectivity. Unified communications deployments are highly dependent on the underlying network connectivity for true redundancy. For this reason it is critical to deploy and configure the underlying network infrastructure in a highly resilient manner. For details on designing highly available network infrastructures, see the chapter on Network Infrastructure, page 3-1. In all cases, the network should be designed so that, given a switch or router failure within the infrastructure, a majority of users will have access to a majority of the services provided within the deployment.

To maximize call processing availability, locate and connect call processing platforms in separate buildings and/or separate network switches when possible to ensure that the impact to call processing will be minimized if there is a failure of the building or network infrastructure switch. With Unified CM

call processing, this means distributing cluster server nodes among multiple buildings or locations within the LAN or MAN deployment whenever possible. And at the very least, it means physically distributing network connections between different physical network switches in the same location.

Furthermore, even though Cisco Unified CME is a standalone call processing entity, providing physical distribution and therefore redundancy for this call processing entity still makes sense when deploying multiple call processing entities. Whenever possible in those scenarios, install each instance of Unified CME in a different physical location within the network, or at the very least physically attach them to different network switches.

Unified CM High Availability

Because of the underlying Unified CM clustering mechanism, a Unified Communications System has additional high availability considerations above and beyond hardware platform disk and power component redundancy, physical network location, and connectivity redundancy. This section examines call processing subscriber redundancy considerations, call processing load balancing, and redundancy of additional cluster services.

Call Processing Redundancy

Unified CM provides the following call processing redundancy configuration options or schemes:

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

These redundancy schemes are facilitated by the built-in registration failover mechanism within the Unified CM cluster architecture, which enables endpoints to re-register to a backup call processing subscriber node when the endpoint's primary call processing subscriber node fails. The registration failover mechanism can achieve failover rates for Skinny Client Control Protocol (SCCP) IP phones of approximately 125 registrations per second. The registration failover rate for Session Initiation Protocol (SIP) phones is approximately 40 registrations per second.

The call processing redundancy scheme you select determines not only the fault tolerance of the deployment, but also the fault tolerance of any upgrade.

With 1:1 redundancy, multiple primary call processing subscriber failures can occur without impacting call processing capabilities. With 2:1 redundancy, on the other hand, only one of the primary call processing subscribers out of the two primary call processing subscribers that share a backup call processing subscriber can fail without impacting call processing. However, if the total number of endpoints registered across both primary subscribers and the traffic to those two primary subscribers are within the capacity limits of the backup subscriber, then the backup subscriber is able to handle the failure of both primary subscribers.

Note Do not deploy 2:1 redundancy if the total capacity utilization across the two primary subscribers would exceed the capacity of the backup subscriber. For example, if the call processing capacity or endpoints capacity utilization exceeds 50% on both primary subscribers, the backup subscriber would not be able to handle call processing services properly if both primary subscribers fail. In these scenarios, for example, some endpoints might not be able to register, some new calls might not be established, and some services and features might not operate properly because the backup subscriber system capacity has been exceeded.

Likewise, with the 1:1 redundancy scheme, upgrades to the cluster can be performed with only a single set of endpoint registration failover periods impacting the call processing services. Whereas with the 2:1 redundancy scheme, upgrades to the cluster can require multiple registration failover periods.

A Unified CM cluster can be upgraded with minimal impact to the services. Two different versions (releases) of Unified CM may be on the same server node, 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 complete, the server nodes can be rebooted to switch the inactive partition to the active partition, thus running the new version of Unified CM.

With the 1:1 redundancy scheme, the following steps enable you to upgrade the cluster while minimizing downtime:

Step 1 Install the new version of Unified CM in the inactive partition, first on the publisher and then on all subscribers (call processing, TFTP, and media resource subscribers). Do not reboot.

Step 2 Reboot the publisher and switch to the new version.

Step 3 Reboot the TFTP subscriber node(s) one at a time and switch to the new version.

Step 4 Reboot any dedicated media resource subscriber nodes one at a time and switch to the new version.

Step 5 Reboot the backup call processing subscribers one at a time and switch to the new version.

Step 6 Reboot the primary call processing subscribers one at a time and switch to the new version. Device registrations will fail-over to the previously upgraded and rebooted backup call processing subscribers. After each primary call processing subscriber is rebooted, devices will begin to re-register to the primary call processing subscriber.

With this upgrade method, there is no period (except for the registration failover period) when devices are registered to subscriber nodes that are running different versions of the Unified CM software. All these steps can be automated using Cisco Prime Collaboration.

While the 2:1 redundancy scheme allows for fewer server nodes in a cluster, registration failover occurs more frequently during upgrades, increasing the overall duration of the upgrade as well as the amount of time call processing services for a particular endpoint will be unavailable. Because there is only a single backup call processing subscriber per pair of primary call processing subscribers, it might be possible to reboot to the new version on only one of the primary call processing subscribers in a pair at a time in order to prevent oversubscribing the single backup call processing subscriber. As a result, there may be a period of time after the first primary call processing subscriber in each pair is switched to the new version, in which endpoint registrations will have to be moved from the backup subscriber to the newly upgraded primary subscriber before the endpoint registrations on the second primary subscriber can be moved to the backup subscriber to allow a reboot to the new version. During this time, not only will endpoints on the second primary call processing subscriber be unavailable while they re-register to the backup subscriber, but until they re-register to a node running the new version, they will also be unable to reach endpoints on other subscriber nodes that have already been upgraded.

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.

Note Because an upgrade of a Unified CM cluster results in a period of time in which some or most devices lose registration and call processing services temporarily, you should plan upgrades in advance and implement them during a scheduled maintenance window. While downtime and loss of services to devices can be minimized by selecting the 1:1 redundancy scheme, there will still be some period of time in which call processing services are not available to some or all users.

For more information on upgrading Unified CM, refer to the installation and upgrade guides available at

https://www.cisco.com/c/en/us/support/unified-communications/unified-communications-manager-callmanager/products-installation-guides-list.html

Unified CM Redundancy with Survivable Remote Site Telephony (SRST)

Cisco IOS SRST provides highly available call processing services for endpoints in locations remote from the Unified CM cluster. Unified CM clustering redundancy schemes certainly provide a high level of redundancy for call processing and other application services within a LAN or MAN environment. However, for remote locations separated from the central Unified CM cluster by a WAN or other low-speed links, SRST can be used as a redundancy method to provide basic call processing services to these remote locations in the event of loss of network connectivity between the remote and central sites. Cisco recommends deploying SRST-capable Cisco IOS routers at each remote site where call processing services are considered critical and need to be maintained in the event that connectivity to the Unified CM cluster is lost. Endpoints at these remote locations must be configured with an appropriate SRST reference within Unified CM so that the endpoint knows what address to use to connect to the SRST router for call processing services when connectivity to Unified CM subscribers is unavailable.

Cisco Unified Enhanced SRST (E-SRST) on a Cisco IOS router can also be used at a remote site to provide backup call processing functionality in the event that connectivity to the central Unified CM cluster is lost. E-SRST provides more telephony features for the IP phones than are available with the regular SRST feature on a router. However, the endpoint capacities for Unified E-SRST are typically less than for basic SRST. Both SRST and E-SRST are supported with Cisco Unified SRST Manager, which synchronizes configurations from Unified CM with SRST and E-SRST, thus reducing manual configuration required in the branch SRST or E-SRST router and enabling users to have a similar calling experience in both SRST and normal modes.

Call Processing Subscriber Redundancy

Depending on the redundancy scheme chosen (see Call Processing Redundancy, page 9-14), 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 subscriber nodes. 1:1 redundancy uses dedicated pairs of primary and backup subscribers, while 2:1 redundancy uses a pair of primary subscribers that share one backup subscriber.

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

Figure 9-5 Basic Redundancy Schemes

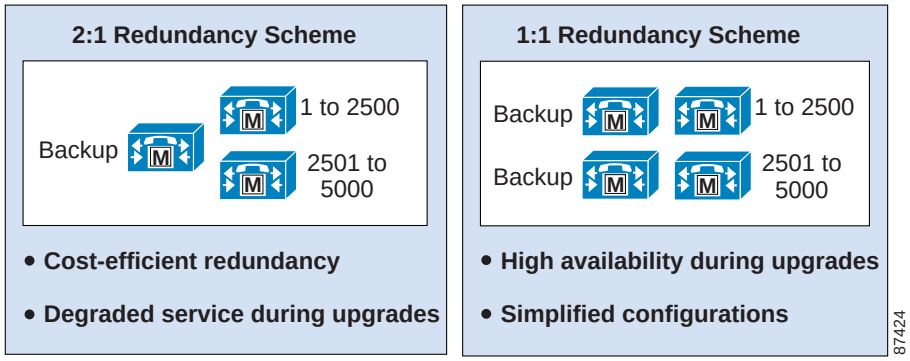


Figure 9-5 illustrates the two basic redundancy schemes available. In each case the backup server node must be capable of handling the capacity of at least a single primary call processing server node failure. In the 2:1 redundancy scheme, the backup might have to be capable of handling the failure of a single call processing server node or potentially both primary call processing server nodes, depending on the requirements of a particular deployment. For information on capacity sizing and choosing the VM configurations, see the section on Capacity Planning for Call Processing, page 9-23.

Note 2:1 redundancy is not supported with the 10,000-User VM configuration due to potential overload on the backup subscriber.

Figure 9-6 1:1 Redundancy Configuration Options

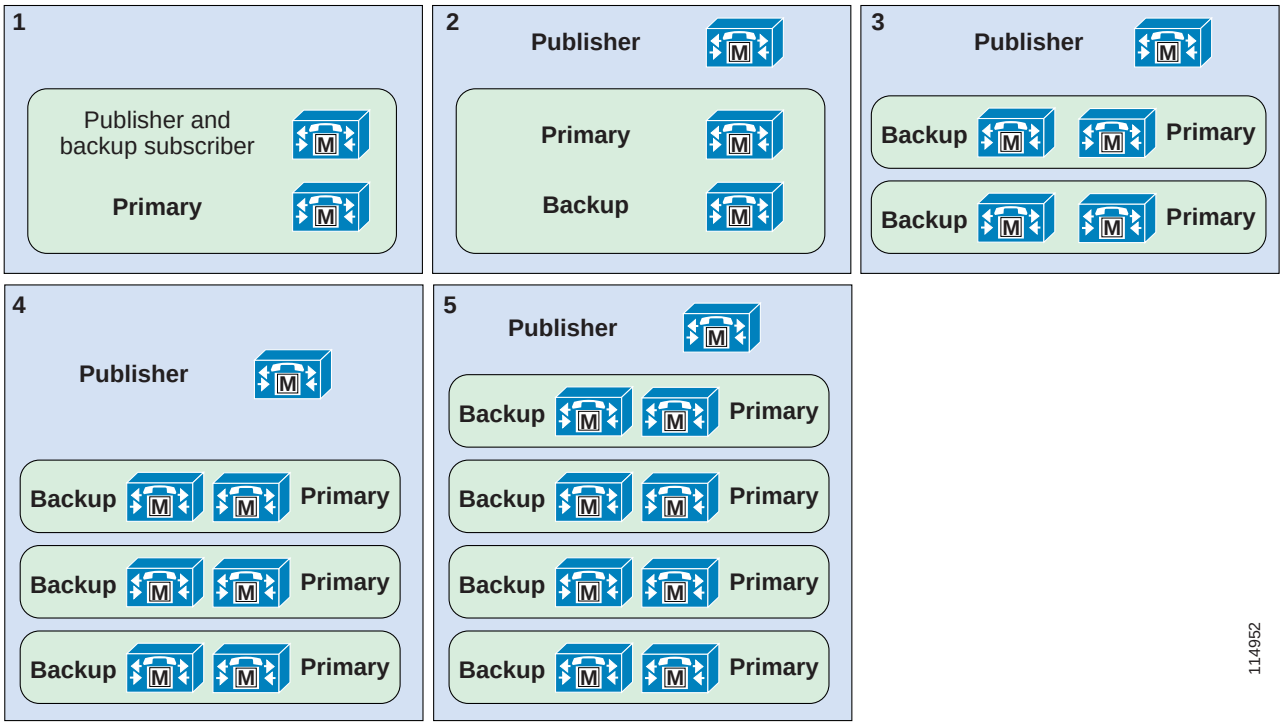
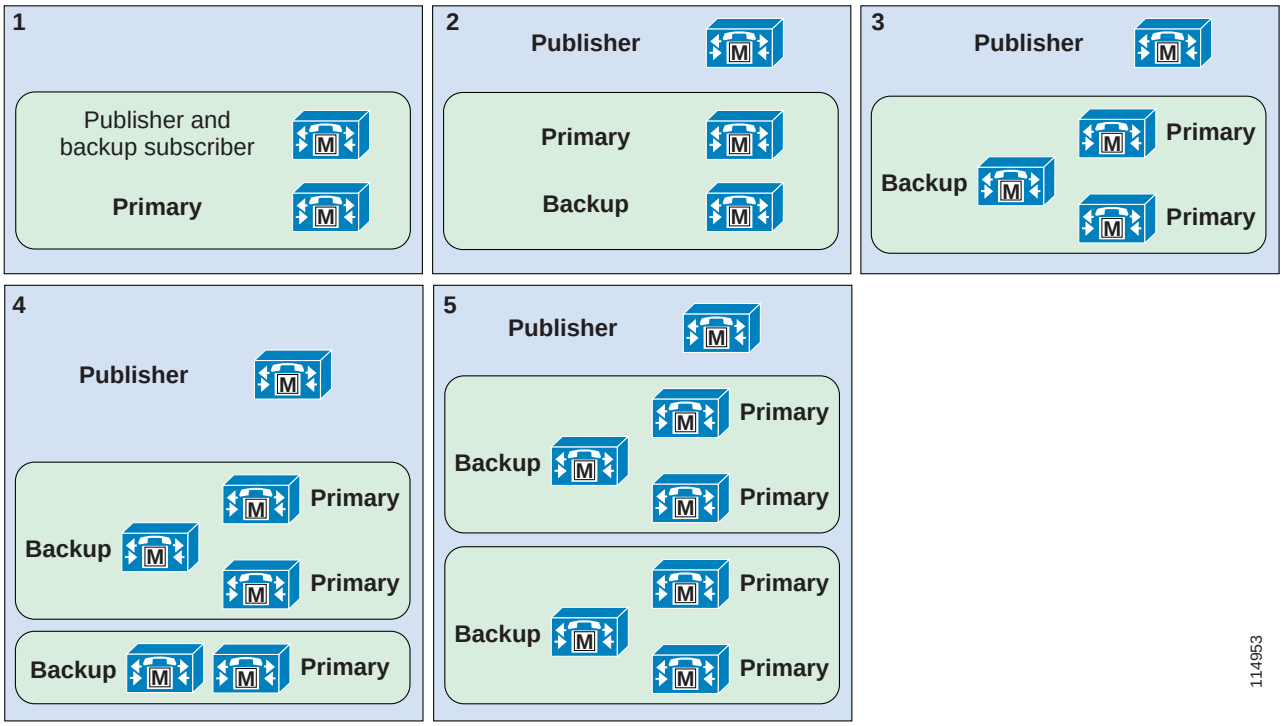


Figure 9-7 2:1 Redundancy Configuration Options



In Figure 9-6, the five options shown all indicate 1:1 redundancy. In Figure 9-7, the five options shown all indicate 2:1 redundancy. In both cases, Option 1 is used for clusters supporting less than 1,250 users and includes Unified CM deployments with Cisco Business Edition 6000. Options 2 through 5 illustrate increasingly scalable clusters for each redundancy scheme. The exact scale depends on the hardware platforms chosen or required.

These illustrations show only publisher and call processing subscribers. They do not account for other subscriber nodes such as TFTP and media resources.

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 nodes in deployments with clustering over the WAN (see Remote Failover Deployment Model, page 10-54), 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. The tertiary subscribers also count against the maximum number of call processing subscribers in a cluster (8 call processing subscriber nodes).

Although not shown in the Figure 9-6 or Figure 9-7, it is also possible to deploy a single-node cluster. The single-node cluster should not exceed 1000 endpoint configuration and registrations. Note that in a single-node configuration, there is no backup call processing subscriber and therefore no cluster redundancy mechanism. Survivable Remote Site Telephony (SRST) can be used as a redundancy mechanism in these types of deployments to provide minimal call processing services during periods when Unified CM is not available.

Load Balancing

In Unified CM clusters with the 1:1 redundancy scheme, device registration and call processing services can be load-balanced across the primary and backup call processing subscriber.

Normally a backup server node has no devices registered to it unless its primary is unavailable. This makes it easier to troubleshoot a deployment because there is a maximum of four primary call processing subscriber nodes that will be handling the call processing load at a given time. Further, this potentially simplifies configuration by reducing the number of Unified CM redundancy groups and device pools.

In a load-balanced deployment, up to half of the device registration and call processing load can be moved from the primary to the secondary subscriber by using the Unified CM redundancy groups and device pool settings. In this way each primary and backup call processing subscriber pair provides device registration and call processing services to as many as half of the total devices serviced by this pair of call processing subscribers. This is referred to as 50/50 load balancing. The 50/50 load balancing model provides the following benefits:

- Load sharing — The registration and call processing load is distributed on multiple server nodes, 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 node 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 node, do not let the total load on the primary and secondary subscribers exceed that of a single subscriber node.

Note During upgrades of a Unified CM cluster with 50/50 load balancing, upgrades to the backup call processing subscriber will result in devices registered to that subscriber (half of the total devices serviced by the primary and backup subscriber pair) failing over to the primary call processing subscriber.

TFTP Redundancy

Cisco recommends deploying more than one dedicated TFTP subscriber node for a large Unified CM cluster, thus providing redundancy for TFTP services. While two TFTP subscribers are typically sufficient, more than two TFTP server nodes can be deployed in a cluster.

In addition to providing one or more redundant TFTP subscribers, you must configure endpoints to take advantage of these redundant TFTP nodes. When configuring the TFTP options using DHCP or statically, define a TFTP subscriber node IP address array containing the IP addresses of both TFTP subscriber nodes within the cluster. In this way, by creating two DHCP scopes with two different IP address arrays (or by manually configuring endpoints with two different TFTP subscriber node IP addresses), you can assign half of the endpoint devices to use TFTP subscriber A as the primary and TFTP subscriber B as the backup, and the other half to use TFTP subscriber B as the primary and TFTP subscriber A as the backup. In addition to providing redundancy during a failure of one TFTP subscriber, this method of distributing endpoints across multiple TFTP subscribers provides load balancing so that one TFTP subscriber is not handling all the TFTP service load.

Note When adding a specific binary or firmware load for a phone or gateway, you must add the file(s) to each TFTP subscriber node in the cluster.

CTI Manager Redundancy

All CTI integrated applications communicate with a call processing subscriber node running the CTI Manager service. Further, most CTI applications have the ability to specify redundant CTI Manager service nodes. For this reason, Cisco recommends activating the CTI Manager service on at least two call processing subscribers within the cluster. With both a primary and backup CTI Manager configured, in the event of a failure the application will switch to a backup CTI Manager to receive CTI services.

As stated previously, the CTI Manager service can be enabled only on call processing subscribers, therefore there is a maximum of eight CTI Managers per cluster. Cisco recommends that you load-balance CTI applications across the enabled 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 a CTI application with the same server node pair used for the CTI Manager service. 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 node A as the primary call processing subscriber and server node B as the backup subscriber. The other two ports would have a Unified CM redundancy group of server node B as the primary subscriber and server node A as the backup subscriber.
- The IVR application would be configured to use the CTI Manager on subscriber A as the primary and subscriber B as the backup.

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

For more details on CTI and CTI Manager, see Computer Telephony Integration (CTI), page 9-28.

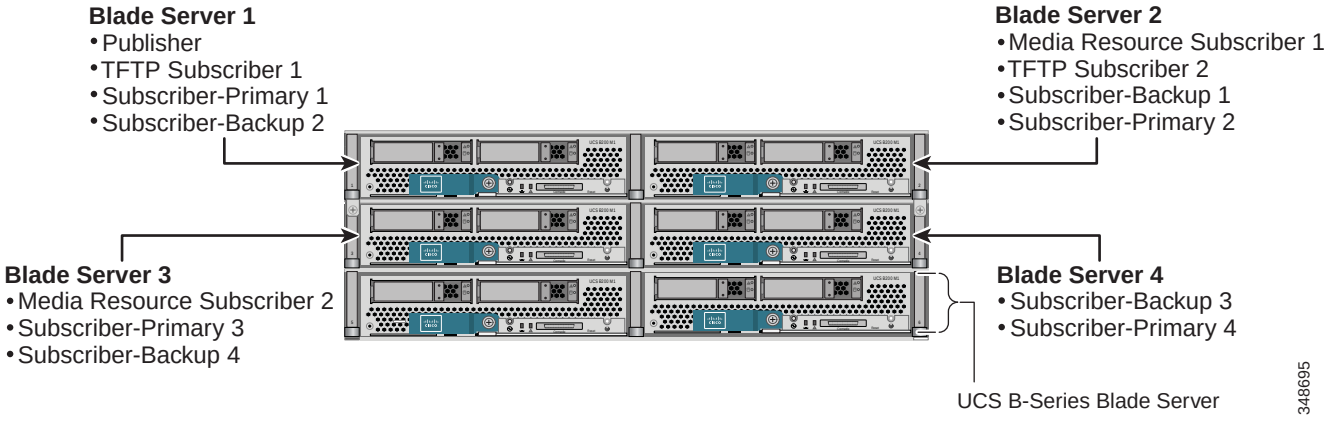
Virtual Machine Placement and Hardware Platform Redundancy

With virtualization there are redundancy considerations because of the virtual nature of server nodes: namely, the installation and residency of Unified CM server node instances across physical servers.

As illustrated by the example in Figure 9-8, observe the following guidelines when deploying Unified CM to ensure the highest level of call processing redundancy:

- Each primary call processing subscriber node instance should reside on a different physical server than its backup call processing subscriber node instance. This ensures that the failure of a server containing the primary call processing node instance does not impact the system's ability to provide endpoints with access to their backup call processing subscriber node.
- When deploying multiple TFTP or media resource subscriber nodes instances for redundancy of those services, always distribute redundant subscriber nodes across more than one server to ensure that a failure of a single server does not eliminate those services. This ensures that, given the failure of a blade containing a TFTP or media resource subscriber, endpoints will still be able to access TFTP and media resource services on a subscriber node residing on another server. Endpoints can also be distributed among redundant TFTP and media resource subscriber node instances to balance system load in non-failure scenarios.
- When deploying CTI applications, always make sure that call processing subscriber node instances running the CTI Manager service are distributed across more than one server to ensure that a failure of a single server does not eliminate CTI services. Further, CTI applications should be configured to use the CTI Manager service running on the subscriber node instance on one server as the primary CTI Manager and the CTI Manager service running on the subscriber node on another server as the backup CTI Manager.

Figure 9-8 Unified CM Server Node Distribution on UCS



When using blade servers with a chassis (for example, B-Series blade servers with a Cisco UCS 5100 Blade chassis), in addition to distributing subscriber node instances across multiple blades, you may distribute subscriber node instances across multiple blade chassis for additional redundancy and scalability.

For more information about redundancy and provisioning of host resources for virtual machines, refer to the documentation at https://www.cisco.com/go/virtualized-collaboration.

Cisco Business Edition High Availability

With Cisco Business Edition 6000M, Cisco Business Edition 6000H, and Cisco Business Edition 7000, high availability is provided by clustering additional Cisco Unified CM node(s). Additional Business Edition server(s) can be deployed to provide high availability for call processing as well as other applications and services.

Note More than two physical servers may be clustered to provide additional redundancy and/or geographic distribution as with a clustering over the WAN deployment. However, with Cisco Business Edition 6000, the additional server(s) only provides redundancy and not a capacity increase. For example, with BE6000M and BE6000H, the total number of users across the cluster may not exceed 1,000. A deployment exceeding this limit is considered to be a standard Unified CM cluster, and as such the deployment must follow high availability design guidance for standard Unified CM. (See Unified CM High Availability, page 9-14.) With Cisco Business Edition 7000, the capacity is not limited to 1,000 users; rather, the standard application capacity planning and design rules apply.

Capacity Planning for Call Processing

Call processing capacity planning is critical for successful unified communications deployments. This section discusses capacity planning for Cisco Unified CM, whether or not it is part of the Cisco Business Edition 6000 or 7000 solution. It also covers Cisco Business Edition 4000 and Unified CME.

Unified CM Capacity Planning

Unified CM capacity depends on the hardware platform, the VM configuration, and the deployment requirements. It also depends on whether or not Unified CM is deployed as part of Cisco Business Edition 6000. Table 9-2 lists some general Unified CM capacity limits.

Table 9-2 Cisco Unified CM Capacity Limits

Capacity Information	Cisco Business Edition 6000M	Cisco Business Edition 6000H	Cisco Business Edition 7000 and Enterprise Cisco Unified CM
Maximum number of users	1,000 per cluster	1,000 per cluster	Up to 10,000 per node; up to 40,000 per cluster[1]
Maximum number of endpoints	1,200 per cluster	2,500 per cluster	Up to 10,000 per node; up to 40,000 per cluster[1]
How to perform capacity planning	Capacity information in product documentation[2]	Capacity information in product documentation[2]	Product documentation, SRND guidelines, and Cisco Collaboration Sizing Tool[3]

1. Could be higher with a Megacluster deployment.
2. When necessary, capacity planning can be based on the Cisco Collaboration Sizing Tool instead of the fixed capacity numbers in the Business Edition product documentation. However, the Unified CM cluster is still limited to 1,000 users.
3. For deployments with up to 10,000 users or endpoints (whichever limit is reached first), capacity planning can be done with the simplified sizing available in the Cisco Preferred Architecture for Enterprise Collaboration.

Cisco Business Edition 6000M/H Capacity Planning

With the Cisco Business Edition 6000, Unified CM is deployed with a specific VM configuration and Unified CM capacity is fixed. Cisco Unified CM capacity planning is simple and does not rely on the Cisco Collaboration Sizing Tool.

Unified CM deployed with Cisco Business Edition 6000M supports up to 1,000 users, 1,200 devices, and 5,000 BHCA. With Business Edition 6000H, up to 1,000 users, 2,500 devices and 5,000 BHCA are supported.

With BE6000, adding nodes or hardware platforms is supported to provide high availability, but that does not increase capacity. The VM configurations specific to the Business Edition 6000 must be used. The larger VM configurations for 2,500, 7,500, or 10,000 users cannot be used with Business Edition 6000.

Unified CM deployed as part of Business Edition 6000 has some additional restrictions. For example, up to 50 sites and up to 100 contact center agents are supported with Cisco Business Edition 6000M/H. (For more details, refer to the Cisco Business Edition 6000 product documentation available at https://www.cisco.com/go/be6000.) If those requirements cannot be met but the number of users is still under 1,000, it is possible to take an alternate approach to the capacity planning with Cisco Business Edition 6000M/H. Instead of relying on fixed capacity specific to Business Edition 6000, the sizing could be done the same way as with Business Edition 7000 and enterprise Unified CM, based on the sizing guidelines in the product documentation, this SRND, and the Cisco Collaboration Sizing Tool. The 1,000-user VM configuration would have to be selected when using the Sizing Tool.

Cisco Business Edition 7000M/H and Cisco Unified CM Capacity Planning

With Cisco Business Edition 7000 or the enterprise (non-Business-Edition) version of Cisco Unified CM, various VM configurations with different corresponding capacities are available, and the capacity increases when nodes are added. Capacity planning is done based on the guidelines in this document and the Cisco Collaboration Sizing Tool. However, there is a simplified capacity planning method described in the *Cisco Preferred Architecture for Enterprise Collaboration CVD* (available at https://www.cisco.com/go/pa). This simplified sizing method can be used if the Unified CM deployment has less than 10,000 users or devices (whichever limit is reached first) and if specific sizing assumptions are met. If those sizing assumptions cannot be met, then simplified capacity planning cannot be used, and normal capacity planning based on the guidelines in the product documentation, this SRND, and the Cisco Collaboration Sizing Tool must be done instead. The sizing tool takes into account many more parameters; for example, the type of phone (SCCP, SIP, or mobile) and the phone security mode are taken into consideration. It is a more complex sizing process, but it can be customized to your specific deployment.

Enabling and increasing utilization of some Unified CM functions can have an impact on the call processing capabilities of the system and in some cases can reduce the overall capacity. These functions include tracing, call detail recording, highly complex dial plans, and other services that are co-resident on the Unified CM platform. 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, call forwarding, co-resident services, and other co-resident applications. All of these functions can consume additional resources within the Unified CM system.

You can use the following technique to improve system performance:

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.

Unified CM Capacity Planning Guidelines and Endpoint Limits

The following capacity guidelines apply to Cisco Unified CM as part of Cisco Business Edition 7000 or outside of Business Edition:

- Within a cluster, a maximum of 8 call processing subscriber nodes can be enabled with the Cisco CallManager Service. Other server nodes may be used for more dedicated functions such as publisher, TFTP subscribers, and media resources subscribers.
- Each Unified CM node can support registration for a maximum of 10,000 secured or unsecured SCCP or SIP endpoints. Each cluster can support configuration and registration for a maximum of 40,000 secured or unsecured SCCP or SIP endpoints.
- There are several VM configuration options for Cisco Unified CM available in the OVA, depending on the required capacity. The names of the VM configurations correspond to the maximum number of users per node, assuming that each user has one phone. When the ratio of number phones per user is different than one, the VM configuration names actually correspond to the maximum number of endpoints per node. Depending on different variables such as BHCA and feature set used, the actual number of users or endpoints could be lower. To validate the sizing of a deployment, use the Cisco Collaboration Sizing Tool, available at https://www.cisco.com/go/cst.

- The default trace setting for the CallManager service is 1,500 files of 10 MB for Signaling Distribution Layer (SDL) traces. 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.

For more information about Unified CM capacity planning considerations, including sizing limits as well as a complete discussion of system sizing, capacity planning, and deployment considerations, see the chapter on Collaboration Solution Sizing Guidance, page 25-1.

Megacluster

The term *megacluster* defines and identifies certain Unified CM deployments that allow for further increases in scalability. A megacluster provides more device capacity through the support of additional Unified CM subscriber nodes, with a maximum of eight Unified CM subscriber pairs (1:1 redundancy) per megacluster, thus allowing for a maximum of 80,000 devices.

A megacluster can also be deployed where customers simply require non-locally redundant call processing functionality, rather than using Survivable Remote Site Telephony (SRST), to scale beyond the maximum eight sites allowed in a standard cluster deployment and up to 16 Unified CM subscriber nodes per megacluster. For example, consider a large hospital that has twelve locations and each location has only 1,000 devices. This total of 12,000 devices could be accommodated within a standard cluster, which has a maximum device capacity of 40,000 devices. However, in this case it is the need for additional Unified CM subscribers, rather than additional device capacity, that requires a megacluster deployment. In this example, a Unified CM subscriber node could be deployed in each location, and each Unified CM subscriber could serve as the primary subscriber for the local endpoints and as a backup subscriber for endpoints from another location.

When considering a megacluster deployment, the primary areas impacting capacity are as follows:

- The megacluster may contain a total of 21 server nodes consisting of 16 subscriber nodes, 2 TFTP server nodes, 2 music on hold (MoH) server nodes, and 1 publisher node.
- Unified CM must be deployed with the 7,500-user or 10,000-user VM configuration options.
- Redundancy model must be 1:1.

All other capacities relating to a standard cluster also apply to a megacluster. Note that support for a megacluster deployment is granted only following the successful review of a detailed design, including the submission of the results from the Cisco Collaboration Sizing Tool. For more information about the Cisco Collaboration Sizing Tool and the sizing of Unified CM standard clusters and megaclusters, see the chapter on Collaboration Solution Sizing Guidance, page 25-1.

Due to the many potential complexities surrounding megacluster deployments, customers who wish to pursue such a deployment must engage either their Cisco Account Team, Cisco Advanced Services, or their certified Cisco Unified Communications Partner.

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

For more information about call processing sizing and for a complete discussion of system sizing, capacity planning, and deployment considerations, see the chapter on Collaboration Solution Sizing Guidance, page 25-1.

Cisco Business Edition 4000 Capacity Planning

Cisco Business Edition 4000 supports a maximum of 200 endpoints. For more information, refer to the chapter on Collaboration Solution Sizing Guidance, page 25-1, and to the Business Edition 4000 documentation available at:

https://www.cisco.com/c/en/us/products/unified-communications/business-edition-4000/index.html

Unified CME Capacity Planning

When deploying Unified CME, it is critical to select a Cisco IOS router platform that provides the desired capacity in terms of number of supported endpoints required. In addition, platform memory capacity should also be considered if the Unified CME router is providing additional services above and beyond call processing, such as IP routing, DNS lookup, dynamic host configuration protocol (DHCP) address services, or VXML scripting.

Unified CME can support a maximum of 450 endpoints on a single Cisco IOS platform; however, each router platform has a different endpoint capacity based on the size of the system. Because Unified CME is not supported within the Cisco Collaboration Sizing Tool, it is imperative to follow capacity information provided in the Unified CME product data sheets available at

https://www.cisco.com/c/en/us/products/unified-communications/unified-communications-manager-express/datasheet-listing.html

Design Considerations for Call Processing

Observe the following design recommendations and guidelines when deploying Cisco call processing:

Cisco Unified CM

- Cisco Unified CM runs only as a virtualized application on the VMware Hypervisor. It does not run directly on a hardware platform without the VMware Hypervisor.
- You can enable a maximum of 8 call processing subscriber nodes (nodes running the Cisco CallManager Service) within a Cisco Unified CM cluster. Additional server nodes may be dedicated and used for publisher, TFTP, and media resources services. An approved megacluster deployment supports a maximum of 16 call processing subscriber nodes.
- Each Unified CM cluster can support configuration and registration for a maximum of 40,000 secured or unsecured endpoints. For additional information about Unified CM capacity planning, including per-platform sizing limits, see the chapter on Collaboration Solution Sizing Guidance, page 25-1.
- When deploying a two-node cluster, Cisco recommends that you do not exceed 1,250 users in the cluster. Above 1,250 users, Cisco recommends a dedicated publisher and separate nodes for primary and backup call processing subscribers.
- Cisco recommends using the same VM configuration for all nodes in a cluster. Mixing Unified CM VM configurations is allowed, but there are design implications and limits. For more details, refer to the section on Mixing Unified CM VM Configurations, page 9-8.
- 2:1 redundancy is not supported when using the 10,000-user VM configuration option due to potential overload on the backup subscriber

- Use multiple physical ports in the hardware platform for the virtual machine network traffic, and use a minimum of two upstream switches to provide network connectivity redundancy. If using the VMware vSphere virtual switch, use VMware NIC teaming.
- Whenever possible, distribute the hardware platforms across multiple physical switches within the network and across multiple physical locations within the same network to minimize the impact of a switch failure or the loss of a particular network location.
- Deploy SRST or E-SRST on Cisco IOS routers at remote locations to provide fallback call processing services in the event that these locations lose connectivity to the Unified CM cluster.
- Cisco recommends that you leave voice activity detection (VAD) disabled within the Unified CM cluster. VAD is disabled by default in the Unified CM service parameters, and you should disable it on H.323 and SIP dial peers configured on Cisco IOS gateways by using the **no vad** command.
- Ensure that the Unified CM nodes are distributed across different physical servers so that backup or redundant subscriber nodes are on different physical servers than the primary subscriber nodes.
- Servers with slower CPUs can be restricted in terms of the Open Virtualization Archive (OVA) VM configuration that they support. We refer to those servers as servers with *restricted UC performance* CPUs. For example, the Cisco Unified CM 1000-user OVA VM configuration is the only Unified CM OVA VM configuration that can be installed on those servers. However, some smaller servers support only smaller VM configurations. For information on proper VM configuration selection as well as the use of the Cisco Collaboration Sizing Tool, see the chapter on Collaboration Solution Sizing Guidance, page 25-1.
- Access to the USB and serial ports on the hardware platform is not supported with Unified CM virtual machines. Therefore, attaching fixed live audio sources for MoH, making a serial SMDI connection to a legacy voicemail system, or attaching a USB flash drive for writing log files are also not supported. The following alternative options are available:
  - For MoH live audio source feed, consider using Cisco IOS-based gateway multicast MoH for live audio source connectivity.
  - For saving system installation logs, use virtual floppy softmedia.
  - There is no support for SMDI serial connection.
- With Cisco Business Edition 6000, Unified CM is deployed as a single Unified CM publisher node that also handles call processing. To provide Unified CM redundancy, SRST can be deployed or additional hardware server(s) hosting Unified CM subscriber node(s) can be deployed.

Note More than two servers may be clustered for a BE6000 deployment to provide additional redundancy and/or geographic distribution; however, the capacity limits are not increased. For example, the total number of users across the cluster may not exceed 1,000 with BE6000M or BE6000H.

- If multiple Business Edition 6000 servers are required in the same deployment, distribute them across multiple physical switches.
- Use an uninterruptible power supply (UPS) to provide maximum availability, especially if the server has only one power supply.
- When deploying Business Edition 6000 with two servers for high availability, a Unified CM node should run on each server to provide high availability in case one of the servers fails. Furthermore, Cisco recommends configuring the Unified CM cluster with the subscriber node as the primary call processing server and the publisher node as the backup call processing server.

- With Cisco Business Edition 7000, Unified CM has the same rules, capacities, and design considerations as an enterprise (not part of Cisco Business Edition) Unified CM deployment.
- Applications that are not part of the Business Edition 6000 solution and that are running on separate hardware can be integrated to a Business Edition 6000 deployment, but you must ensure that those applications do not exceed the Business Edition 6000 capacity limits. For example, the overall BHCA and the number of contact center agents should not exceed the Business Edition 6000 capacity limit of Unified CM. For more information on the Business Edition 6000 capacity limits, refer to the chapter on Collaboration Solution Sizing Guidance, page 25-1. Also ensure that those applications support the Cisco Collaboration VM configuration provided by Business Edition 6000. For example, Cisco Unified Contact Center Enterprise requires the Unified CM 7.5k-user or larger VM configuration, so it cannot be integrated with a Unified CM deployment that is running on Business Edition 6000.

Cisco Unified CME

- Unified CME supports a maximum of 450 endpoints. However, depending on the Cisco IOS router model, endpoint capacity could be significantly lower. For additional information about Unified CME platforms and capacities, refer to the Cisco Unified Communications Manager Express compatibility information available at https://www.cisco.com/c/en/us/support/unified-communications/unified-communications-manager-express/products-device-support-tables-list.html.
- When possible, dual-attach the Unified CME router to the network using multiple IP interfaces to provide maximum network availability. Likewise, if multiple instances of Unified CME are required in the same deployment, distribute them across multiple physical switches or locations.
- When possible, deploy the Unified CME router with dual power supplies and/or an uninterruptible power supply (UPS) in order to provide maximum availability of the platform.

Computer Telephony Integration (CTI)

Cisco Computer Telephony Integration (CTI) extends the rich feature set available on Cisco Unified CM to third-party applications. The 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 Jabber, 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 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.

The following devices can be monitored or controlled through CTI:

- CTI Route Point
- CTI Port
- Cisco Unified IP Phones supporting CTI
- CTI Remote Device

CTI Remote Device provides the ability for a CTI application to have monitoring and limited call control capabilities over phones that do not support CTI, such as traditional PSTN phones, mobile phones, third-party phones, or phones attached to a third-party PBX.

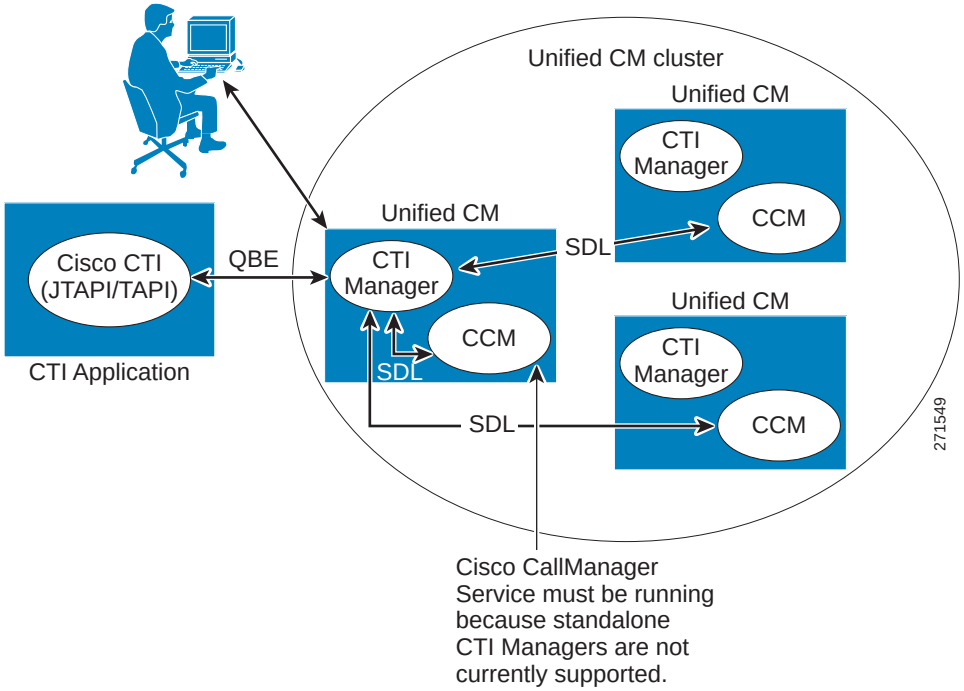
CTI Architecture

Cisco CTI consists of the following components (see Figure 9-9), 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.

- Signaling Distribution Layer (SDL) — Unified CM internal communication messages.
- Publisher and subscriber — Cisco Unified Communications Manager (Unified CM) server nodes.
- 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 9-9 Cisco CTI Architecture



Once an application is authenticated and authorized, the CTIM acts as the broker between the telephony application and the Cisco CallManager Service. (This service is the call control agent and should not be confused with the overall product name Cisco Unified Communications Manager.) The CTIM responds to requests from telephony applications and converts them to Signaling Distribution Layer (SDL) messages used internally in the Unified CM system. Messages from the Cisco CallManager Service are also received by the CTIM and directed to the appropriate telephony application for processing.

The CTIM may be activated on any of the Unified CM subscriber nodes in a cluster that have the Cisco CallManager Service active. This allows up to eight CTIMs to be active within a Unified CM cluster. Standalone CTIMs are currently not supported.

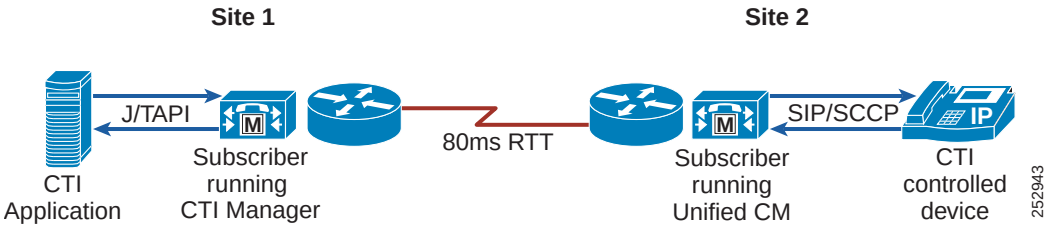
CTI Applications and Clustering Over the WAN

Deployments that employ clustering over the WAN are supported in the following two scenarios:

- CTI Manager over the WAN (see Figure 9-10)

  In this scenario, the CTI application and its associated CTI Manager are on one side of the WAN (Site 1), and the monitored or controlled devices are on the other side, registered to a Unified CM subscriber (Site 2). The round-trip time (RTT) must not exceed the currently supported limit of 80 ms for clustering over the WAN. To calculate the necessary bandwidth for CTI traffic, use the formula in the section on Local Failover Deployment Model, page 10-47. Note that this bandwidth is in addition to the Intra-Cluster Communication Signaling (ICCS) bandwidth calculated as described in the section on Local Failover Deployment Model, page 10-47, as well as any bandwidth required for audio (RTP traffic).

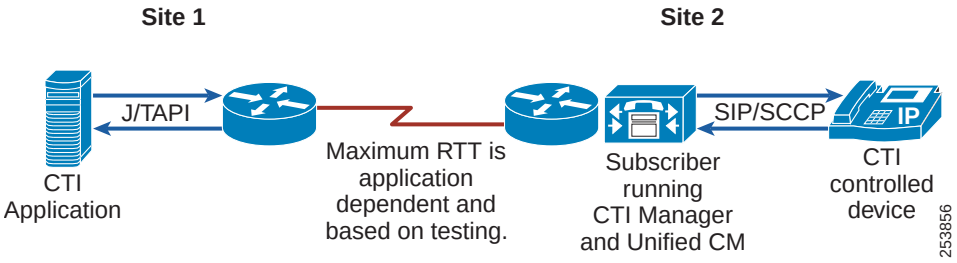
Figure 9-10 CTI Over the WAN



- TAPI and JTAPI applications over the WAN (CTI application over the WAN; see Figure 9-11)

  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 9-11 JTAPI Over the WAN



Note Support for TAPI and JTAPI over the WAN is application dependent. Both customers and application developers or providers should ensure that their applications are compatible with any such deployment involving clustering over the WAN.

Capacity Planning for CTI

The maximum number of supported CTI-controlled devices is 40,000 per cluster. For more information on CTI capacity planning, including per-platform node and cluster CTI capacities as well as CTI resource calculation formulas and examples, see the chapter on Collaboration Solution Sizing Guidance, page 25-1.

High Availability for CTI

This section provides some guidelines for provisioning CTI for high availability.

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 server nodes running the CTIM service. For CTI application redundancy, Cisco recommends having the CTIM service activated on at least two Unified CM server nodes in a cluster, as shown in Figure 9-12.

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 node 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 9-12.

Figure 9-12 Redundancy and Load Balancing

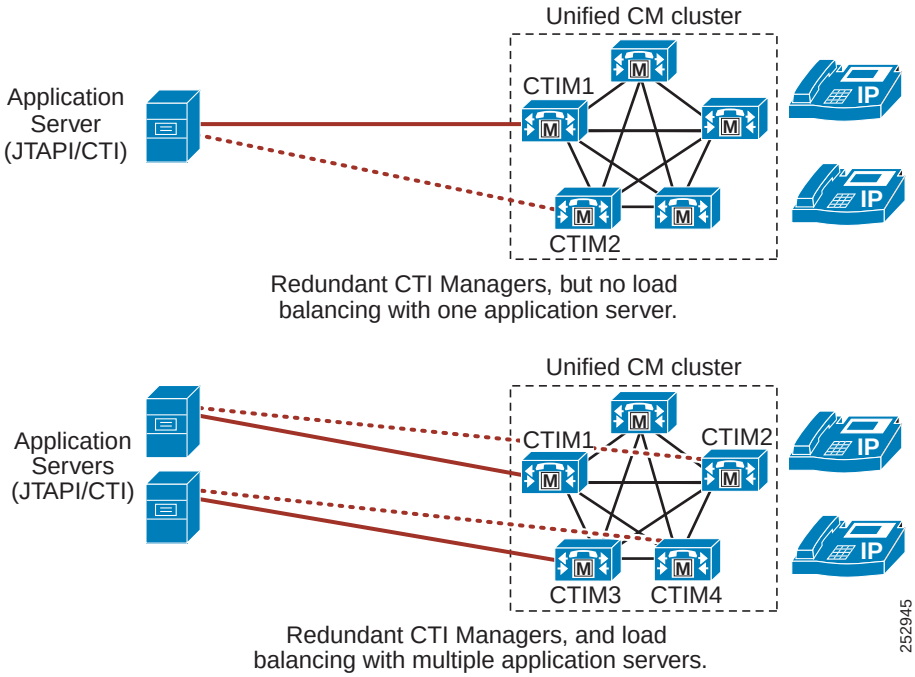


Figure 9-13 shows an example of this type of configuration for Cisco Unified Contact Center Enterprise (Unified CCE). This type of configuration has the following characteristics:

- 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 9-13 CTI Redundancy with Cisco Unified Contact Center Enterprise

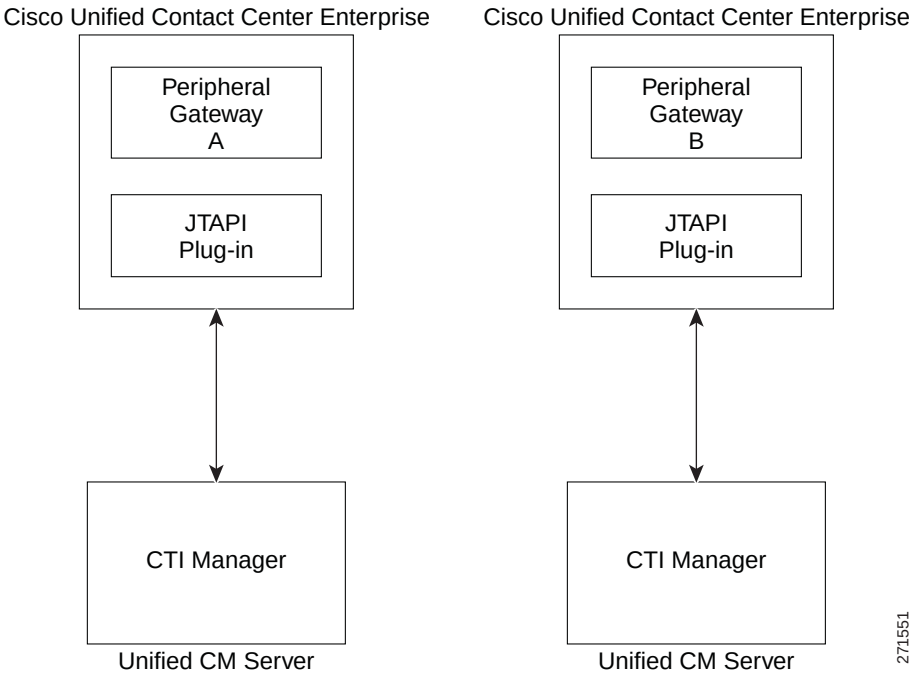
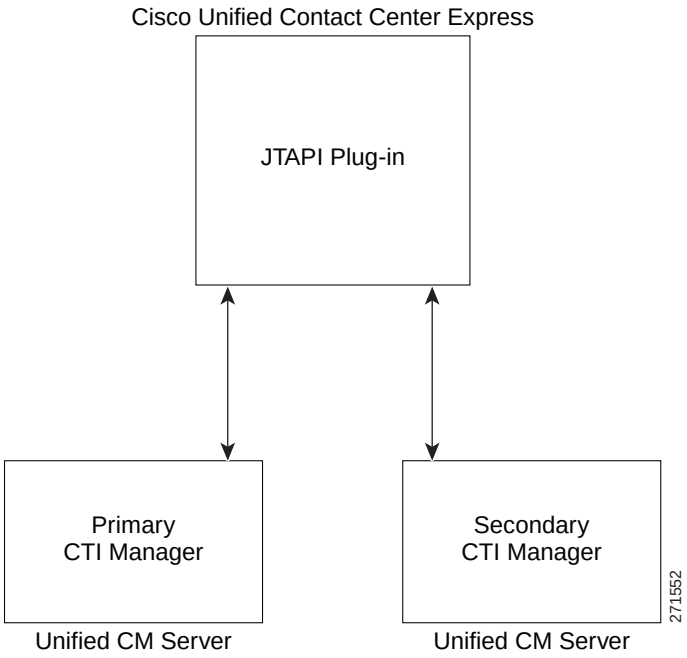


Figure 9-14 shows an example of this type of configuration for Cisco Unified Contact Center Express (Unified CCX). This type of configuration has the following characteristics:

- 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 9-14 CTI Redundancy with Cisco Unified Contact Center Express



Implementation

For guidance and support on writing applications, application developers should consult the Cisco Developer Network (DevNet), located at

https://developer.cisco.com/site/devnet/home

Integration of Multiple Call Processing Agents

To integrate multiple Unified CM clusters together or to integrate Unified CM clusters with the Cisco TelePresence Video Communication Server (VCS), use Cisco Unified CM Session Management Edition (SME). SME is the recommended trunk and dial plan aggregation platform in multi-site distributed call processing deployments. SME is essentially a Unified CM cluster with trunk interfaces only and no IP endpoints. It enables aggregation of multiple Unified Communications systems, referred to as *leaf* systems.

Unified CM Session Management Edition may also be used to connect to third-party unified communications systems such as PSTN connections, PBXs, and centralized unified communications applications.

For more information on SME, see the section on Unified CM Session Management Edition, page 10-26.

Direct integration of multiple call processing agents is also possible. This section explains the requirements for interoperability and internetworking of Cisco Unified CM with Cisco Unified Communications Manager Express (Unified CME) using 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 9-36
- Unified CM and Unified CME Interoperability via SIP in a Multisite Deployment with Distributed Call Processing, page 9-38

Cisco Unified CM and Cisco Unified Communications Manager Express (Unified CME) could also be integrated using H.323, but this section does not cover this integration in detail. For more information on the H.323 integration, refer to the *Cisco Collaboration 9.x SRND*, available at

https://www.cisco.com/go/srnd

Overview of Interoperability Between Unified CM and 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 after 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.

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

https://www.cisco.com/c/en/us/support/unified-communications/unified-communications-manager-express/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.

Instant and Permanent Hardware Conferencing

Hardware DSP resources are required for both instant and permanent conferences. Whether connected via SIP, H.323, or PSTN, both Unified CM and Unified CME phones can be invited or added to an instant conference to become conference participants as along 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.

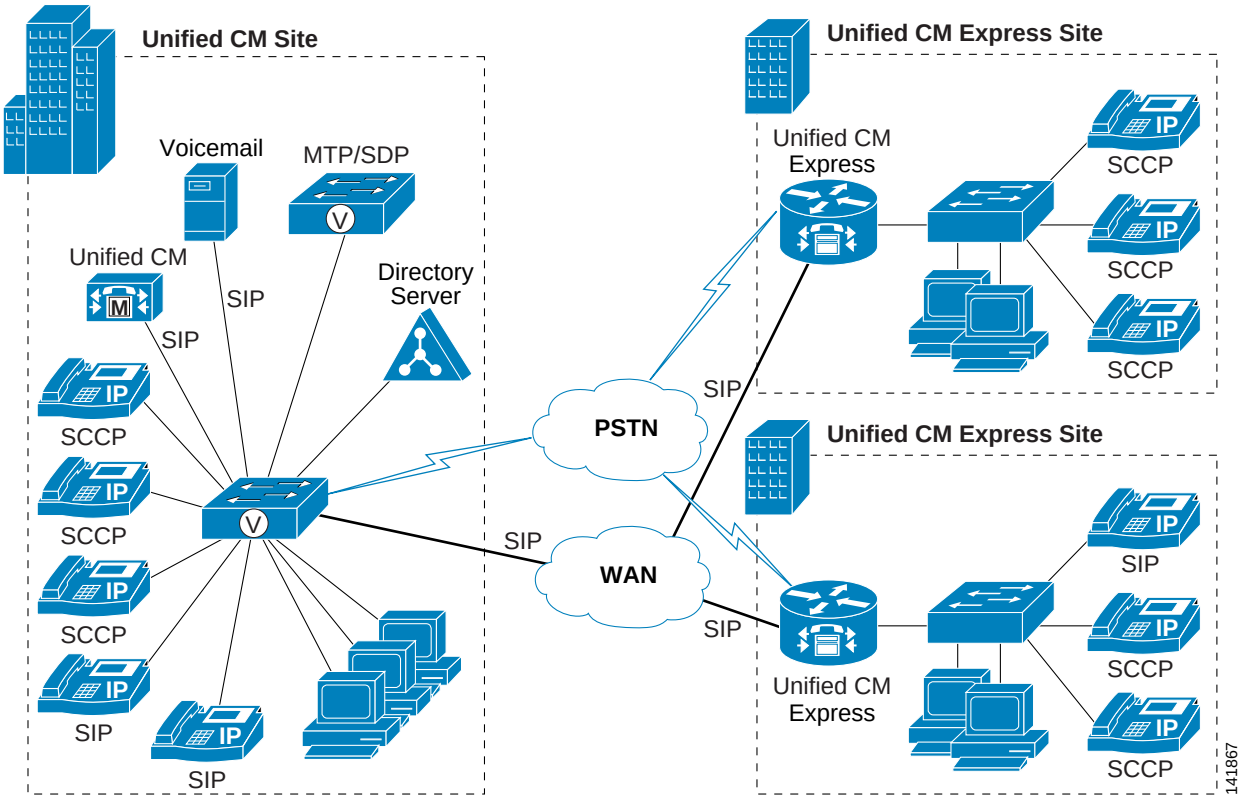
For information on required and supported DSP resources and the maximum number of conference participants allowed for instant or permanent conferences, refer to the Unified CME product documentation available at

https://www.cisco.com/c/en/us/support/unified-communications/unified-communications-manager-express/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 9-15 shows a Cisco Unified Communications multisite deployment with Unified CM networked directly with Cisco Unified CME using a SIP trunk.

Figure 9-15 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 9-15:

- 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.

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

- 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.
- SRTP over a SIP trunk is a gateway feature in Cisco IOS 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**.