CHAPTER 7

Music on Hold

Last revised on: February 13, 2008

Music on hold (MoH) is an integral feature of the Cisco Unified Communications system. This feature provides music to callers when their call is placed on hold, transferred, parked, or added to an ad-hoc conference. Implementing MoH is relatively simple but requires a basic understanding of unicast and multicast traffic, MoH call flows, configuration options, server behavior and requirements. This chapter describes how to design and provision MoH resources for a Cisco Enterprise IP Telephony deployment.

Cisco Unified CallManager provides access to a variety of media resources. A media resource is a software-based or hardware-based entity that performs some media processing function on the voice data streams that are connected to it. Media processing functions include mixing multiple streams to create one output stream, passing the stream from one connection to another, or transcoding the data stream from one compression type to another.

Cisco Unified CallManager allocates and uses the following types of media resources:

- Media termination point (MTP) resources
- Transcoding resources
- Unicast conferencing resources
- Annunciator resources
- Music on hold resources

For more information about media resources in general, see the chapter on Media Resources, page 6-1.

This chapter examines the following design aspects of the MoH feature:

Deployment Basics of MoH

For callers to hear music on hold, Cisco Unified CallManager must be configured to support the MoH feature. The MoH feature has two main requirements:

- An MoH server to provide the MoH audio stream sources
- Cisco Unified CallManager configured to use the MoH streams provided by the MoH server when a call is placed on hold

The integrated MoH feature allows users to place on-net and off-net users on hold with music streamed from a streaming source. This source makes music available to any on-net or off-net device placed on hold. On-net devices include station devices and applications placed on hold, consult hold, or park hold by an interactive voice response (IVR) or call distributor. Off-net users include those connected through Media Gateway Control Protocol (MGCP) and H.323 gateways. The MoH feature is also available for plain old telephone service (POTS) phones connected to the Cisco IP network through Foreign Exchange Station (FXS) ports. The integrated MoH feature includes media server, database administration, call control, media resource manager, and media control functional areas. The MoH server provides the music resources and streams.

You can configure the MoH feature via the Cisco Unified CallManager Administration interface. When an end device or feature places a call on hold, Cisco Unified CallManager connects the held device to an MoH media resource. Essentially, Cisco Unified CallManager instructs the end device to establish a connection to the MoH server. When the held device is retrieved, it disconnects from the MoH resource and resumes normal activity.

Unicast and Multicast MoH

Cisco Unified CallManager supports two types of MoH transport mechanisms:

- Unicast
- Multicast

Unicast MoH consists of streams sent directly from the MoH server to the endpoint requesting an MoH audio stream. A unicast MoH stream is a point-to-point, one-way audio Real-Time Transport Protocol (RTP) stream between the server and the endpoint device. Unicast MoH uses a separate source stream for each user or connection. As more endpoint devices go on hold via a user or network event, the number of MoH streams increases. Thus, if twenty devices are on hold, then twenty streams of RTP traffic are generated over the network between the server and these endpoint devices. These additional MoH streams can potentially have a negative effect on network throughput and bandwidth. However, unicast MoH can be extremely useful in those networks where multicast is not enabled or where devices are not capable of multicast, thereby still allowing an administrator to take advantage of the MoH feature.

Multicast MoH consists of streams sent from the MoH server to a multicast group IP address that endpoints requesting an MoH audio stream can join as needed. A multicast MoH stream is a point-to-multipoint, one-way audio RTP stream between the MoH server and the multicast group IP address. Multicast music on hold conserves system resources and bandwidth because it enables multiple users to use the same audio source stream to provide music on hold. Thus, if twenty devices are on hold, then potentially only a single stream of RTP traffic is generated over the network. For this reason, multicast is an extremely attractive technology for the deployment of a service such as MoH because it greatly reduces the CPU impact on the source device and also greatly reduces the bandwidth consumption for delivery over common paths. However, multicast MoH can be problematic in situations where a network is not enabled for multicast or where the endpoint devices are not capable of handling multicast.

For information about IP multicast network design, refer to the *IP Multicast SRND* document, available online at

http://www.cisco.com/go/designzone

Recommended Unicast/Multicast Gateways

The following recommended gateways support both unicast and multicast MoH:

- Cisco 6624 and 6608 gateway modules with MGCP and Cisco Unified CallManager Release 3.3(3) or later
- Cisco Communication Media Module (CMM) with MGCP or H.323 and Cisco Unified CallManager Release 4.0, Cisco IOS Release 12.2(13)ZP3 or later, and Catalyst OS Release 8.1(1) or later
- Cisco 2600, 2800, 3600, 3700, and 3800 Series Routers with MGCP or H.323 and Cisco IOS Release 12.2(8)T or later

Co-resident and Standalone MoH Servers

The MoH feature requires the use of a server that is part of a Cisco Unified CallManager cluster. You can configure the MoH server in either of the following ways:

- Co-resident deployment

  In a co-resident deployment, the MoH feature runs on any server (either publisher or subscriber) in the cluster that is also running the Cisco Unified CallManager software. Because MoH shares server resources with Cisco Unified CallManager in a co-resident configuration, this type of configuration drastically reduces the number of simultaneous streams that an MoH server can send.

- Standalone deployment

  A standalone deployment places the MoH feature on a dedicated server within the Cisco Unified CallManager cluster. The sole function of this dedicated server is to send MoH streams to devices within the network. A standalone deployment allows for the maximum number of streams from a single MoH server.

Fixed and Audio File MoH Sources

You can set up the source for MoH in any of the following ways:

- MoH from an audio file on the Cisco Unified CallManager or MoH server
  - Unicast MoH from an audio file
  - Multicast MoH from an audio file
- MoH from a fixed music source (via sound card)
  - Unicast MoH from a fixed source
  - Multicast MoH from a fixed source

MoH can be generated from an audio file stored on the MoH server. Audio files must be in one of the following formats:

- G.711 A-law or mu-law
- G.729 Annex A
- Wideband

You can create these files with the Cisco MoH Audio Translator service, which transcodes and formats audio source files (such as .wav or .mp3 files) into the appropriate MoH source file for the specified codec type(s). The MoH server requests these files based on the audio sources configured. When an MoH event occurs, the configured audio source file is streamed to the requesting device on hold.

If recorded or live audio is needed, MoH can be generated from a fixed source. For this type of MoH, a sound card is required. The fixed audio source is connected to the audio input of the local sound card.

This mechanism enables you to use radios, CD players, or any other compatible sound source. The stream from the fixed audio source is transcoded in real-time to support the codec that was configured through Cisco Unified CallManager Administration. The fixed audio source can be transcoded into G.711 (A-law or mu-law), G.729 Annex A, and Wideband, and it is the only audio source that is transcoded in real-time.

The following sound cards are supported for a fixed or live audio source:

- Griffin Technologies iMic USB

  A USB sound device supported in Cisco CallManager Release 3.3(5) and later, with Microsoft Windows 2000 (OS 2000 version 2.7 or later). This device is supported on all Cisco MCS-78*xx*H or MCS-78*xx*I servers with 3.0 GHz or greater processor.

- Telex P-800 USB

  A USB sound device supported in Cisco CallManager Release 3.3(3) with Microsoft Windows 2000 (OS 2000 version 2.5). This device is supported on all Cisco MCS-78*xx*H or MCS-78*xx*I servers with 2.2 GHz or greater processor.

Note The Telex P-800 USB card is no longer available. While existing P-800 cards are still supported as indicated above, the Griffin iMic USB card should be used in new deployments.

Note Prior to using a fixed audio source to transmit music on hold, you should consider the legalities and the ramifications of re-broadcasting copyrighted audio materials. Consult your legal department for potential issues.

MoH Server as Part of the Cisco Unified CallManager Cluster

The MoH feature requires that each MoH server must be part of a Cisco Unified CallManager cluster. All MoH servers must share their configurations with the publisher server and participate in the database replication schema. Specifically, the MoH server must share the following information (configured through Cisco Unified CallManager Administration) by means of the database:

- Audio sources — The number and identity of all configured MoH audio sources
- Multicast or unicast — The transport nature (multicast or unicast) configured for each of these sources
- Multicast address — The multicast base IP address of those sources configured to stream as multicast

The MoH server becomes part of the Cisco Unified CallManager cluster and participates in the database replication automatically. To configure a standalone MoH server, start with a normal Cisco Unified CallManager installation on that server, then disable the Cisco CallManager service (on the standalone MoH server only) and enable the Cisco IP Voice Media Streaming Application.

Basic MoH and MoH Call Flows

This section describes the basic operation of MoH as implemented in Cisco Unified CallManager, as well as typical call flow scenarios.

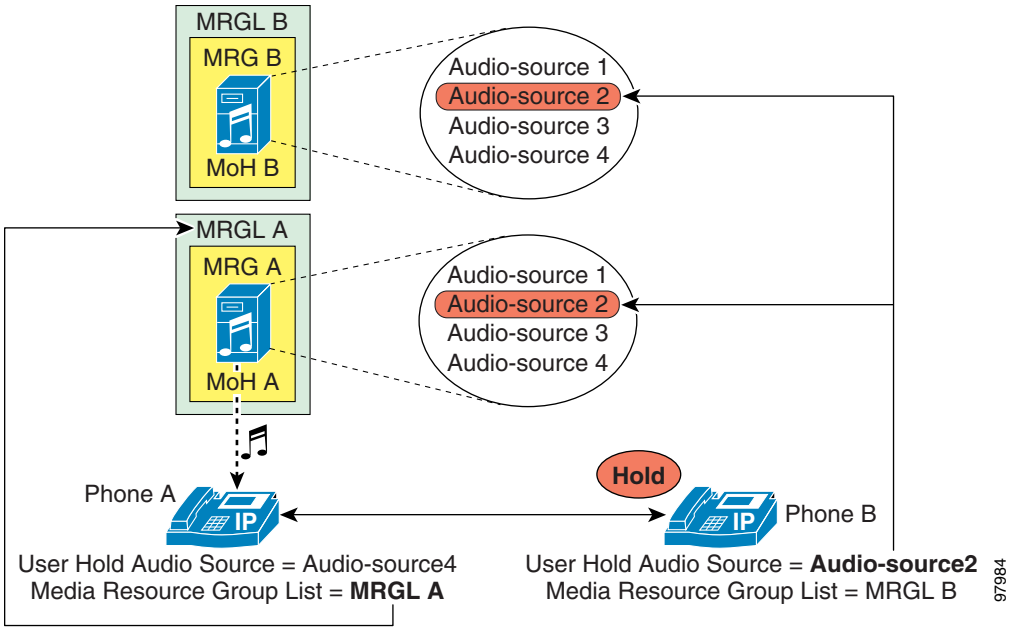
Basic MoH

The basic operation of MoH in a Cisco Unified Communications environment consists of a holder and a holdee. The *holder* is the endpoint user or network application placing a call on hold, and the *holdee* is the endpoint user or device placed on hold.

The MoH stream that an endpoint receives is determined by a combination of the User Hold MoH Audio Source of the device placing the endpoint on hold (holder) and the configured media resource group list (MRGL) of the endpoint placed on hold (holdee). The User Hold MoH Audio Source configured for the holder determines the audio file that will be streamed when the holder puts a call on hold, and the holdee's configured MRGL indicates the resource or server from which the holdee will receive the MoH stream.

In simplest terms, the holder's configuration determines which audio file to play, and the holdee's configuration determines which resource or server will play that file. As illustrated by the example in Figure 7-1, if phones A and B are on a call and phone B (holder) places phone A (holdee) on hold, phone A will hear the MoH audio source configured for phone B (Audio-source2). However, phone A will receive this MoH audio stream from the MRGL (resource or server) configured for phone A (MRGL A).

Figure 7-1 User Hold Audio Source and Media Resource Group List (MRGL)



Because the configured MRGL determines the server from which a unicast-only device will receive the MoH stream, you must configure unicast-only devices with an MRGL that points to a unicast MoH resource or media resource group (MRG). Likewise, a device capable of multicast should be configured with an MRGL that points to a multicast MRG.

MoH Configuration Settings

You can configure the settings for MRGLs and User and Network Hold Audio Sources in several places within Cisco Unified CallManager Administration, and you can configure different (and potentially conflicting) settings in each place.

To determine which User and Network Audio Source configuration setting to apply in a particular case, Cisco Unified CallManager interprets these settings for the *holder* device in the following priority order:

1. Directory or line setting (Devices with no line definition, such as gateways, do not have this level.)
2. Device setting
3. Common Profile setting
4. Cluster-wide default setting

When attempting to determine the audio source for a particular holder, Cisco Unified CallManager first looks at the User (or Network) Audio Source configured at the directory or line level. If this level is not defined, Cisco Unified CallManager looks at the User (or Network) Audio Source configured on the holder device. If this level is not defined, Cisco Unified CallManager looks at the User (or Network) Audio Source configured for the common profile of the holder device. If this level is not defined, then Cisco Unified CallManager looks at the cluster-wide default audio source ID configured under the Cisco Unified CallManager system parameters. (By default, this audio source ID is set to 1 for both User and Network Hold Audio Sources, which is the SampleAudioSource.)

Cisco Unified CallManager also interprets the MRGL configuration settings of the *holdee* device in the following priority order:

1. Device setting
2. Device pool setting
3. System default MoH resources

When attempting to determine the MRGL for a particular holdee, Cisco Unified CallManager looks at the MRGL configured at the device level. If this level is not defined, Cisco Unified CallManager looks at the MRGL configured for the device pool of the holdee device. If this level is not defined, then Cisco Unified CallManager uses the system default MoH resources. System default MoH resources are those resources not assigned to any MRG, and they are always unicast.

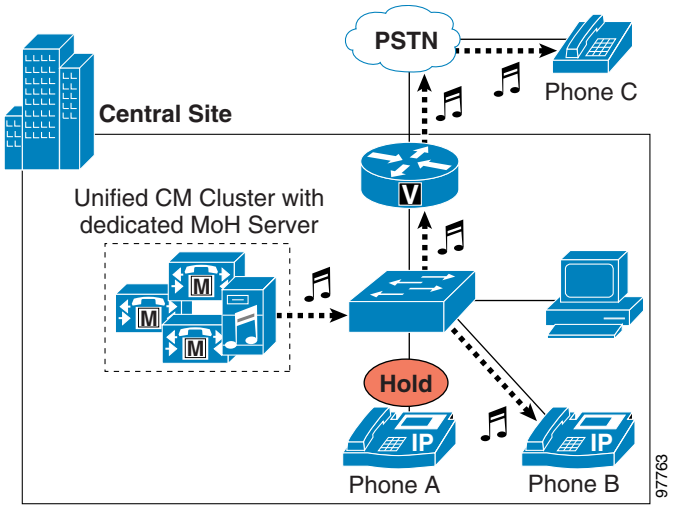
User and Network Hold

There are two basic types of user hold:

- User on hold at an IP phone or other endpoint device
- User on hold at the PSTN, where MoH is streamed to the gateway

Figure 7-2 shows these two types of call flows. If phone A is in a call with phone B and phone A (holder) pushes the Hold softkey, then a music stream is sent from the MoH server to phone B (holdee). The music stream can be sent to holdees within the IP network or holdees on the PSTN, as is the case if phone A places phone C on hold. In the case of phone C, the MoH stream is sent to the voice gateway interface and converted to the appropriate format for the PSTN phone. When phone A presses the Resume softkey, the holdee (phone B or C) disconnects from the music stream and reconnects to phone A.

Figure 7-2 Basic User Hold Example

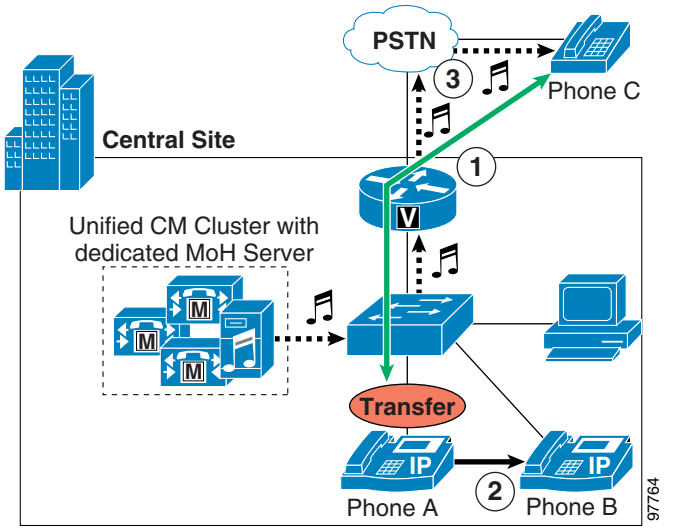


Network hold includes the following types:

- Call transfer
- Call Park
- Conference setup
- Application-based hold

Figure 7-3 shows the call transfer call flow. When phone A receives a call from PSTN phone C (step 1), phone A answers the call and then transfers it to phone B (step 2). During the transfer process, phone C receives an MoH stream from the MoH server via the gateway (step 3). After phone A completes the transfer action, phone C disconnects from the music stream and gets redirected to phone B, the transfer destination. This process is the same for other network hold operations such as call park and conference setup.

Figure 7-3 Basic Network Hold Example for Call Transfer



Unicast and Multicast MoH Call Flows

MoH operation is quite similar to normal phone call flows, with the MoH server acting as an endpoint device to which the holdee device can connect or disconnect as required. However, there are distinct differences between unicast and multicast MoH call flow behavior. A unicast MoH call flow is initiated by a message from Cisco Unified CallManager to the MoH server. This message tells the MoH server to send an audio stream to the holdee device's IP address. On the other hand, a multicast MoH call flow is initiated by a message from Cisco Unified CallManager to the holdee device. This message instructs the endpoint device to join the multicast group address of the configured multicast MoH audio stream.

For a detailed look at MoH call flows, see the section on Detailed Unicast and Multicast MoH Call Flows, page 7-20.

MoH Configuration Considerations and Best Practices

This section highlights some MoH configuration considerations and best practice to help you design a robust MoH solution.

Codec Selection

If you need multiple codecs for MoH deployment, configure them in the IP Voice Streaming Media App service parameter under Cisco CallManager Service Parameters Configuration. Select the desired codec types from the Supported MoH Codecs list under the Clusterwide Parameters section. By default, only G.711 mu-law is selected. To select another codec type, click on it in the scrollable list. For multiple selections, hold down the CTRL key and use the mouse to select multiple codecs from the scrollable list. After making your selection, click the Update button.

Note If you are using the G.729 codec for MoH audio streams, be aware that this codec is optimized for speech and it provides only marginal audio fidelity for music.

Multicast Addressing

Proper IP addressing is important for configuring multicast MoH. Addresses for IP multicast range from 224.0.1.0 to 239.255.255.255. The Internet Assigned Numbers Authority (IANA), however, assigns addresses in the range 224.0.1.0 to 238.255.255.255 for public multicast applications. Cisco strongly discourages using public multicast addresses for music on hold. Instead, Cisco recommends that you configure multicast MoH audio sources to use IP addresses in the range 239.1.1.1 to 239.255.255.255, which is reserved for administratively controlled applications on private networks.

Furthermore, you should configure multicast audio sources to increment on the IP address and not the port number, for the following reasons:

- IP phones placed on hold join multicast IP addresses, not port numbers.

  Cisco IP phones have no concept of multicast port numbers. Therefore, if all the configured codecs for a particular audio stream transmit to the same multicast IP address (even on different port numbers), all streams will be sent to the IP phone even though only one stream is needed. This has the potential of saturating the network with unnecessary traffic because the IP phone is capable of receiving only a single MoH stream.

- IP network routers route multicast based on IP addresses, not port numbers.

  Routers have no concept of multicast port numbers. Thus, when it encounters multiple streams sent to the same multicast group address (even on different port numbers), the router forwards all streams of the multicast group. Because only one stream is needed, network bandwidth is over-utilized and network congestion can eventually result.

MoH Audio Sources

Audio sources are shared among *all* MoH servers in the Cisco Unified CallManager cluster. You can configure up to 51 unique audio sources per cluster (50 audio file sources and one fixed/live source via a sound card). For exceptions to this limit, refer to the sections on Using Multiple Fixed or Live Audio Sources, page 7-9 and Multicast MoH from Branch Router Flash, page 7-17.

Using Multiple Fixed or Live Audio Sources

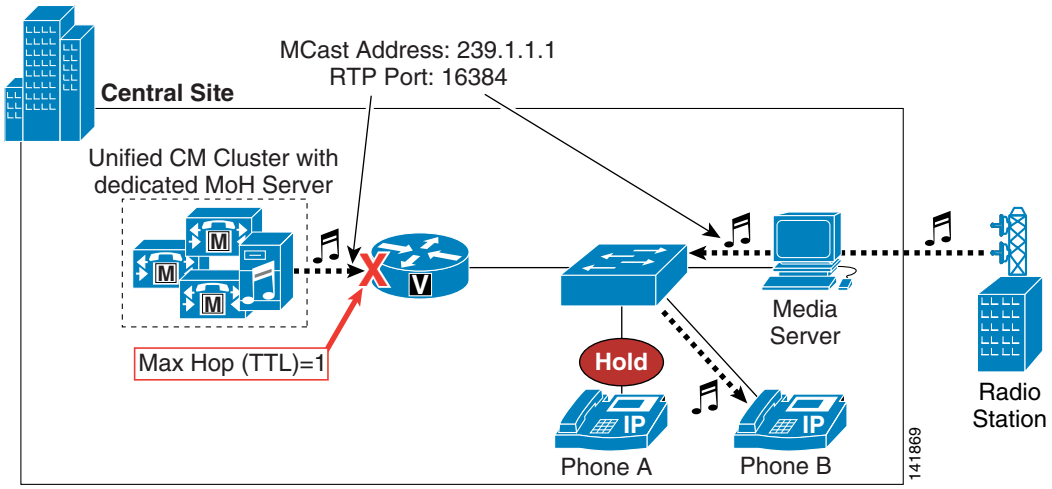
Each MoH server is capable of streaming only one fixed audio source. In most cases, if multiple fixed or live audio sources are needed, a separate MoH server is required for each source. However, it is possible to provide multiple fixed-source MoH audio streams by using external non-MoH servers or devices that are capable of streaming multicast from fixed or live sources.

For each external source, you must configure the MoH server with an audio source that has the same multicast IP address and port number as that of the audio source stream being multicast by the external source server or device. In addition, you should block this configured (non-external) audio source from traversing the WAN by setting its maximum hop count to one (1) or by using access control lists (ACLs) to prevent the packets from streaming further than the local subnet.

Figure 7-4 shows an example of an external live source being used as an MoH stream. In this figure, the MoH server is streaming a multicast audio source to 239.1.1.1 (on RTP port 16384), and this stream has been limited to a maximum hop of one, thus ensuring that it will not travel beyond the local MoH server's subnet. At the same time, the media server is multicasting an audio stream derived from a live radio station feed. This stream is also using 239.1.1.1 as its multicast address and 16384 as the RTP port number, but this stream has a hop count or Time to Live (TTL) greater than one to ensure that it is able to reach phone B when phone A presses the Hold softkey.

Note Multicast Time to Live (TTL) values are decremented (or they expire) only when packets traverse Layer 3 interfaces.

Figure 7-4 External Live Audio Source Example



Note Using live radio broadcasts as multicast audio sources can have legal ramifications. Consult your legal department for potential issues.

Numerous streams can be multicast from one or more external media servers, by configuring additional audio sources on multiple MoH servers and then sourcing audio streams from the external servers using the same multicast group addresses configured on the MoH servers. However, because a combination of the user/network hold audio source of the holder and the MRGL of the holdee determines the MoH stream that an endpoint device hears, it can become difficult to predict which particular stream an endpoint will receive in an environment with many overlapping multicast group addresses. For this reason, Cisco recommends that you configure only a single multicast audio source on each MoH server. This recommendation ensures that the audio source an endpoint receives is uniquely identifiable by a single combination of user/network hold audio source and MRGL.

Unicast and Multicast in the Same Cisco Unified CallManager Cluster

In some cases, administrators might want to configure a single Cisco Unified CallManager cluster to handle both unicast and multicast MoH streams. This configuration might be necessary because the telephony network contains devices or endpoint that do not support multicast or because some portions of the network are not enabled for multicast.

Use one of the following methods to enable a cluster to support both unicast and multicast MoH audio streams:

- Deploy separate MoH servers, with one server configured as a unicast MoH server and the second server configured as a multicast MoH server.
- Configure separate media resource groups (MRGs) for the same MoH server, with one MRG configured to use multicast for audio streams and the second MRG configured to use unicast.

In either case, you must configure at least two MRGs and at least two media resource group lists (MRGLs). Configure one unicast MRG and one unicast MRGL for those endpoints requiring unicast MoH. Likewise, configure one multicast MRG and one multicast MRGL for those endpoints requiring multicast MoH.

When deploying separate MoH servers, configure one server without multicast enabled (unicast-only) and configure a second MoH server with multicast enabled. Assign the unicast audio resource of the unicast-only MoH server and the multicast audio resource of the multicast MoH server to the unicast and multicast MRGs, respectively. Ensure that the **Use Multicast for MoH Audio** box is checked for the multicast MRG but not for the unicast MRG. Also assign these unicast and multicast MRGs to their respective MRGLs. In this case, an MoH stream is unicast or multicast based on the server from which it is served and on whether the MRG is configured to use multicast.

When deploying a single MoH server for both unicast and multicast MoH, configure the server and its audio source for multicast. Assign this same audio source to both the unicast MRG and the multicast MRG, and check the **Use Multicast for MoH Audio** box for the multicast MRG. In this case, an MoH stream is unicast or multicast based solely on whether the MRG is configured to use multicast.

Note When configuring the unicast MRG, do not be confused by the fact that the audio resource you are adding to this MRG has [Multicast] appended to the end of the resource name even though you are adding it to the unicast MRG. This label is simply an indication that the resource is capable of being multicast, but the **Use Multicast for MoH Audio** box determines whether the resource will be sent as unicast or multicast.

In addition, you must configure individual devices or device pools to use the appropriate MRGL. You can place all unicast devices in a device pool or pools and configure those device pools to use the unicast MRGL. Likewise, you can place all multicast devices in a device pool or pools and configure those device pools to use the multicast MRGL. Optionally, you can configure individual devices to use the appropriate unicast or multicast MRGL. Also configure a User Hold Audio Source and Network Hold Audio Source for each device pool, individual device, or (in the case of phone devices) individual lines or directory numbers to determine the appropriate audio source to stream.

When choosing a method for deploying both multicast and unicast MoH in the same cluster, an important factor to consider is the number of servers required. When using a single MoH server for both unicast and multicast, fewer MoH servers are required throughout the cluster. Deploying separate multicast and unicast MoH servers will obviously require more servers within the cluster.

Redundancy

Cisco recommends that you configure and deploy multiple MoH servers for completely redundant MoH operation. If the first MoH server fails or becomes unavailable because it no longer has the resources required to service requests, the second server can provide continued MoH functionality. For proper redundant configuration, assign resources from at least two MoH servers to each MRG in the cluster.

Resources in the MRG are used in the order listed. When a device requests an MoH audio resource, Cisco Unified CallManager attempts to stream the first MoH resource in the MRG to the device. If the first resource is unavailable (due to server failure or lack of resources), Cisco Unified CallManager then attempts to use the next MoH resource in the MRG.

In environments where both multicast and unicast MoH are required, be sure to provide redundancy for both transport types to ensure MoH redundancy for all endpoints in the network.

Quality of Service (QoS)

Convergence of data and voice on a single network requires adequate QoS to ensure that time-sensitive and critical real-time applications such as voice are not delayed or dropped. To ensure proper QoS for voice traffic, the streams must be marked, classified, and queued as they enter and traverse the network

to give the voice streams preferential treatment over less critical traffic. MoH servers automatically mark audio stream traffic the same as voice bearer traffic, with a Differentiated Services Code Point (DSCP) value of 46 or Per Hop Behavior (PHB) value of EF (ToS of 0xB8). Therefore, as long as QoS is properly configured on the network, MoH streams will receive the same classification and priority queueing treatment as voice RTP media traffic.

Call signaling traffic between MoH servers and Cisco Unified CallManager servers is marked with a DSCP value of 26 or PHB value of AF31 (ToS of 0x68) by default. However, in order to conform with Cisco QoS marking recommendations, this traffic should be marked with a DSCP value of 24 or PHB value of CS3 (ToS of 0x60) to ensure that it is properly classified and queued within the network. The default traffic marking can be changed to the appropriate marking by changing the *IP Type of Service to Cisco CallManager* service parameter value from 0x68 to 0x60 under the Cisco IP Voice Media Streaming App Service Parameter configuration page.

Hardware and Capacity Planning for MoH Resources

As with all media resources, capacity planning is crucial to make certain that the hardware, once deployed and configured, can support the anticipated call volume of the network. For this reason, it is important to be aware of the hardware capacity for MoH resources and to consider the implications of multicast and unicast MoH in relation to this capacity.

Server Platform Limits

Table 7-1 lists the server platforms and the maximum number of simultaneous MoH sessions each can support. Ensure that network call volumes do not exceed these limits because, once MoH sessions have reached these limits, additional load could result in poor MoH quality, erratic MoH operation, or even loss of MoH functionality.

Table 7-1 Maximum Number of MoH Sessions per Server Platform Type

Server Platform	Codecs Supported	MoH Sessions Supported
MCS 7815 MCS 782x (All models) MCS 7830 (All models) SPE-310 HP DL320 IBM xSeries 33x (All models)	G.711 (A-law and mu-law) G.729a Wideband audio	Co-resident server: 40 MoH sessions Standalone MoH server: 200 MoH sessions
MCS 7835 (All models) MCS 7845 (All models) HP DL380 IBM xSeries 34x (All models)	G.711 (A-law and mu-law) G.729a Wideband audio	Coresident server: 100 MoH sessions Standalone MoH server: 250 MoH sessions[1]

1. You can configure a maximum of 51 unique audio sources per Cisco Unified CallManager cluster.

The following MoH Server Configuration parameters affect MoH server capacity:

- **Maximum Half Duplex Streams**

  This parameter determines the number of devices that can be placed on unicast MoH. By default this value is set to 250.

  The Maximum Half Duplex Streams parameter should be set to the value derived from the following formula:

  (Server and deployment capacity) – ((Number of multicast MoH sources) * (Number of MoH codecs enabled))

  For example:

MCS-7835 standalone MoH server	Multicast MoH audio sources	MoH codecs enables (G.711 mu-law and G.729)	Maximum half-duplex streams
250	- (12	* 2)	= 226

  Therefore, in this example, the Maximum Half Duplex Streams parameter would be configured with a value of no more than 226.

  The value of this parameter should never be set higher than the capacities indicated in Table 7-1, based on the platform and deployment type (co-resident or standalone).

- **Maximum Multicast Connections**

  This parameter determines the number of devices that can be placed on multicast MoH. By default this value is set to 30.

  The Maximum Multicast Connections parameter should be set to a number that ensures that all devices can be placed on multicast MoH if necessary. Although the MoH server can generate only a finite number of multicast streams (maximum of 204), large numbers of held devices can join each multicast stream, thus it is necessary to set this parameter to an artificially large number. Typically multicast traffic is accounted for based on the number of streams being generated; however, Cisco Unified CallManager maintains a count of the actual number of devices placed on multicast MoH or joined to each multicast MoH stream. This method is different than the way multicast traffic is normally tracked.

Failure to configure these parameters properly could lead to under-utilization of MoH server resources or failure of the server to handle the network load.

Note The maximum session limits listed in Table 7-1 apply to unicast, multicast, or simultaneous unicast and multicast sessions. The limits represent the recommended maximum sessions a platform can support, irrespective of the transport mechanism.

Resource Provisioning and Capacity Planning

When provisioning for co-resident or standalone MoH server configurations, network administrators should consider the type of transport mechanism used for the MoH audio streams. If using unicast MoH, each device on hold will require a separate MoH stream. However, if using multicast MoH and only a single audio source, then only a single MoH stream is required for each configured codec type, no matter how many devices of that type are on hold.

For example, given a cluster with 30,000 phones and a 2% hold rate (only 2% of all endpoint devices are on hold at any given time), 600 MoH streams or sessions would be required. Given a unicast-only MoH environment, two standalone MoH servers (MCS 7835 or 7845) and one co-resident MoH server (MCS 7835 or 7845) would be required to handle this load, as shown by the following calculation:

[(250 sessions per MCS 7835 or 7845 standalone server) * (2 standalone servers)] + [(100 sessions per MCS 7835 or 7845 co-resident server) * (1 co-resident server)] = 600 sessions

By comparison, a multicast-only MoH environment with 36 unique MoH audio streams, for example, would require only one co-resident MoH server (MCS 7815, 782*x*, or 7830), as shown by the following calculation:

(40 sessions per MCS 7815, 782*x*, or 7830 co-resident server) * (1 co-resident server) > 36 sessions

These 36 unique multicast streams could be provisioned in any one of the following ways:

- 36 unique audio sources streamed using a single codec
- 18 unique audio sources streamed using only 2 codecs
- 12 unique audio sources streamed using only 3 codecs
- 9 unique audio source streamed using all 4 codecs

As these examples show, multicast MoH can provide a considerable savings in server resources over unicast MoH.

In the preceding examples, the 2% hold rate is based on 30,000 phones and does not take into account gateways or other endpoint devices in the network that are also capable of being placed on hold. You should consider these other devices when calculating a hold rate because they could potentially be placed on hold just as the phones can.

The preceding calculations also do not provide for MoH server redundancy. If an MoH server fails or if more than 2% of the users go on hold at the same time, there are no other MoH resources in this scenario to handle the overflow or additional load. Your MoH resource calculations should include enough extra capacity to provide for redundancy.

Note Because you can configure only 51 unique audio sources per Cisco Unified CallManager cluster and because there are only four possible codecs for MoH streams, the maximum number of multicast streams per MoH server is 204.

Implications for MoH With Regard to IP Telephony Deployment Models

The various IP Telephony deployment models introduce additional considerations for MoH configuration design. Which deployment model you choose can also affect your decisions about MoH transport mechanisms (unicast or multicast), resource provisioning, and codecs. This section discusses these issues in relation to the various deployment models.

For more detailed information about the deployment models, see the chapter on IP Telephony Deployment Models, page 2-1.

Single-Site Campus (Relevant to All Deployments)

Single-site campus deployments are typically based on a LAN infrastructure and provide sufficient bandwidth for large amounts of traffic. Because bandwidth is typically not limited in a LAN infrastructure, Cisco recommends the use of the G.711 (A-law or mu-law) codec for all MoH audio streams in a single-site deployment. G.711 provides the optimal voice and music streaming quality in an IP Telephony environment.

MoH server redundancy should also be considered. In the event that an MoH server becomes overloaded or is unavailable, configuring multiple MoH servers and assigning them in preferred order to MRGs ensures that another server can take over and provide the MoH streams.

With the increasing diversity of network technologies, in a large single-site campus it is likely that some endpoint devices will be unable to support multicast. For this reason, you might have to deploy both unicast and multicast MoH resources. For example, wireless IP phones do not support multicast due to the behavior of wireless technology. Thus, when deploying wireless IP phones, you have to configure both multicast and unicast MoH.

To ensure that off-net calls and application-handled calls receive expected MoH streams when placed on hold, configure all gateways and other devices with the appropriate MRGLs and audio sources, or assign them to appropriate device pools and common profiles.

Centralized Multisite Deployments

Multisite IP telephony deployments with centralized call processing typically contain WAN connections to multiple non-central sites. These WAN links usually cause bandwidth and throughput bottlenecks. To minimize bandwidth consumption on these links, Cisco recommends the use of the G.729 codec for all MoH audio streams traversing the WAN. Because the G.729 codec is optimized for voice and not music applications, you should use G.729 only across the WAN, where the bandwidth savings far outweighs the lower quality afforded by G.729 for MoH transport. Likewise, because multicast traffic provides significant bandwidth savings, you should always use multicast MoH when streaming audio to endpoints across the WAN.

If the sound quality of an MoH stream becomes an issue when using the G.729 codec across the WAN, you can use the G.711 codec for MoH audio streams across the WAN while still using G.729 for voice calls. In order to send MoH streams across the WAN with the G.711 codec but voice calls across the WAN with the G.729 codec, place all MoH servers in a Cisco Unified CallManager region by themselves, and configure that region to use G.711 between itself and all other regions. Thus, when a call is placed between two phones on either side of a WAN, the G.729 codec is used between their

respective regions. However, when the call is placed on hold by either party, the MoH audio stream is encoded using G.711 because G.711 is the configured codec to use between the MoH server's region and the region of the phone placed on hold.

Call Admission Control and MoH

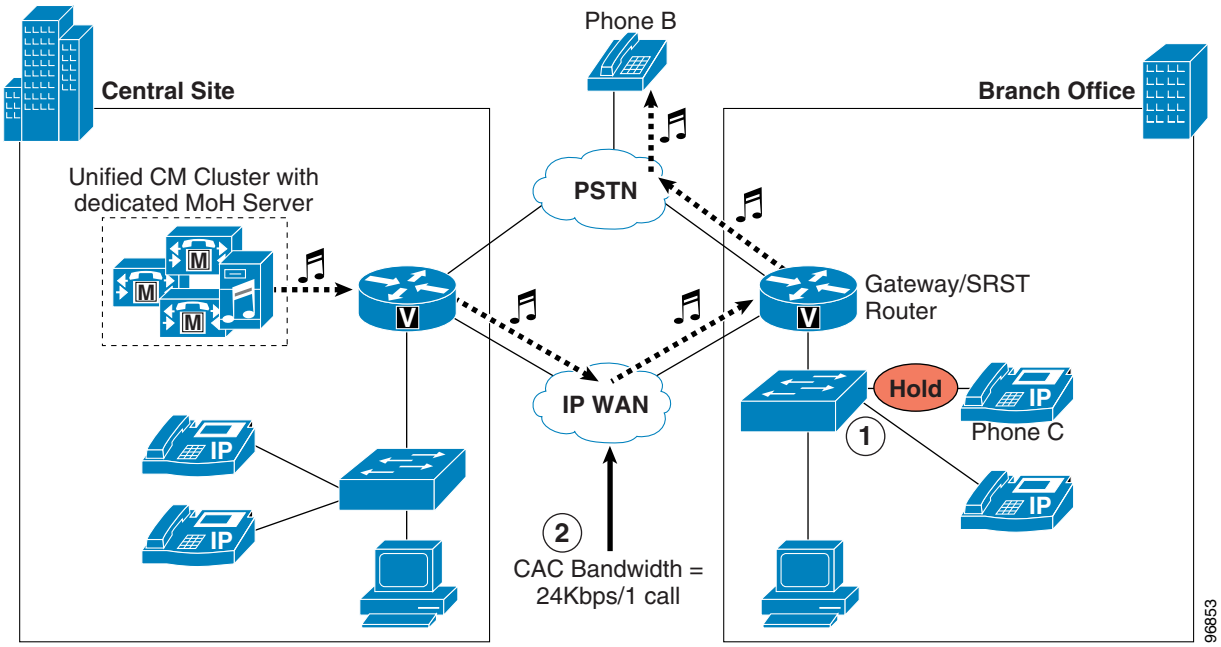
Call admission control (CAC) is required when IP telephony traffic is traveling across WAN links. Due to the limited bandwidth available on these links, it is highly possible that voice media traffic might get delayed or dropped. The Cisco Unified CallManager locations-based call admission control mechanism enables you to configure each location in the IP telephony environment to accept or allow only a certain number of calls across the WAN link to other locations, thus preventing oversubscription of WAN bandwidth and delayed or lost voice packets. By specifying a bandwidth value for the WAN link, you can limit the number of calls based on the speed of the link. If the limit is reached or exceeded, Cisco Unified CallManager rejects all other calls that are attempted across the link. For additional information, see Call Admission Control, page 9-1.

Cisco Unified CallManager locations-based call admission control is capable of tracking unicast MoH streams traversing the WAN but not multicast MoH streams. Thus, even if WAN bandwidth has been fully subscribed, a multicast MoH stream will not be denied access to the WAN by call admission control. Instead, the stream will be sent across the WAN, likely resulting in poor audio stream quality and poor quality on all other calls traversing the WAN. To ensure that multicast MoH streams do not cause this over-subscription situation, you should over-provision the QoS configuration on all downstream WAN interfaces by configuring the low-latency queuing (LLQ) voice priority queue with additional bandwidth. Because MoH streams are uni-directional, only the voice priority queues of the downstream interfaces (from the central site to remote sites) must be over-provisioned. Add enough bandwidth for every unique multicast MoH stream that might traverse the WAN link. For example, if there are four unique multicast audio streams that could potentially traverse the WAN, then add 96 kbps to the voice priority queue (4 * 24 kbps per G.729 audio stream = 96 kbps).

Figure 7-5 shows an example of call admission control and MoH in a centralized multisite deployment. For this example, assume that IP phone C is in a call with a PSTN phone (phone B). At this point, no bandwidth has been consumed on the WAN. When phone C pushes the Hold softkey (step 1), phone B receives an MoH stream from the central-site MoH server via the WAN, thereby consuming bandwidth on the link. Whether or not this bandwidth is taken into consideration by call admission control depends on the type of MoH stream. If multicast MoH is streamed, then call admission control will not consider the 24 kbps being consumed (therefore, QoS on the downstream WAN interfaces should be provisioned accordingly). However, if unicast MoH is streamed, call admission control will subtract 24 kbps from the available WAN bandwidth (step 2).

Note The preceding example might seem to imply that unicast MoH should be streamed across the WAN. However, this is merely an example used to illustrate locations-based call admission control with MoH and is not intended as a recommendation or endorsement of this configuration. As stated earlier, multicast MoH is the recommended transport mechanism for sending MoH audio streams across the WAN.

Figure 7-5 Locations-Based Call Admission Control and MoH



Multicast MoH from Branch Router Flash

Beginning with Cisco IOS Release 12.2(15)ZJ and SRST Release 3.0, MoH can be multicast in a remote or branch site via the branch router's flash. Multicast MoH from a Cisco IOS router's flash enhances the MoH feature for the following reasons:

- The branch gateway or router can provide multicast MoH when it is in SRST mode and the branch devices have lost connectivity to the central-site Cisco Unified CallManager.
- This configuration can eliminate the need to forward MoH across the WAN to remote branch sites by providing locally sourced MoH even when the WAN is up and the phones are controlled by Cisco Unified CallManager.

Example 7-1 illustrates the commands to use in the Cisco IOS router configuration (under the SRST section) to enable multicast MoH from the router flash.

Example 7-1 Enabling Multicast MoH from Branch Router Flash

```
SRST-router(config)#call-manager-fallback
SRST-router(config-cm-fallback)#ip source-address 10.1.1.1
SRST-router(config-cm-fallback)#moh music-on-hold.au
SRST-router(config-cm-fallback)#multicast moh 239.192.240.1 port 16384 route 10.1.1.254
```

In Example 7-1, the name of the audio file on the router flash is music-on-hold.au, and the configured multicast address and port number are 239.192.240.1 and 16384 respectively. The optional **route** command indicates a source interface address for the multicast stream. If no route option is specified, the multicast stream will be sourced from the configured SRST default address as specified by the **ip source-address** command under the SRST configuration. Note that you can stream only a single audio file from flash and that you can use only a single multicast address and port number per router.

When the branch router is operating in SRST mode, it can stream multicast MoH to all analog and digital ports within the chassis, thereby providing MoH to analog phones and PSTN callers. At this time, IP phones in SRST mode cannot receive multicast MoH from the SRST router's flash and will receive tone-on-hold instead.

Note An SRST license is require regardless of whether the SRST functionality will actually be used. The license is required because the configuration for streaming MoH from branch router flash is done under the SRST configuration mode and, even if SRST functionality will not be used, at least one **max-ephones** and one **max-dn** must be configured. These configuration commands are required in addition to the commands listed in Example 7-1.

Once configured, the router will continue to stream the MoH stream from flash even when not in SRST mode. When the branch router is not operating in SRST mode, it can multicast MoH from the flash to all local devices, including IP phones. The branch router's configuration for non-SRST multicast MoH from flash is the same as for the SRST configuration. (See Example 7-1.) However, which multicast address you configure on the router depends on the intended operation. If you want multicast MoH from flash only in SRST mode (for example, if MoH received by remote devices is sourced from the central MoH server when not in SRST mode), then the multicast address and port number configured on the router should not overlap with any of the central-site MoH server audio sources. Otherwise, remote devices might continue to receive MoH from the local router flash, depending on the configured user/network hold audio sources.

If you always want multicast MoH from the branch router flash, then you must configure the central-site server with an audio source that has the same multicast IP address and port number as configured on the branch router. In this scenario, because the multicast MoH audio stream is always coming from the router's flash, it is not necessary for the central site MoH server audio source to traverse the WAN.

To prevent the central site audio stream(s) from traversing the WAN, use one of the following methods:

- Configure a maximum hop count

  Configure the central-site MoH audio source with a maximum hop count (or TTL) low enough to ensure that it will not stream further than the central-site LAN.

- Configure an access control list (ACL) on the WAN interface

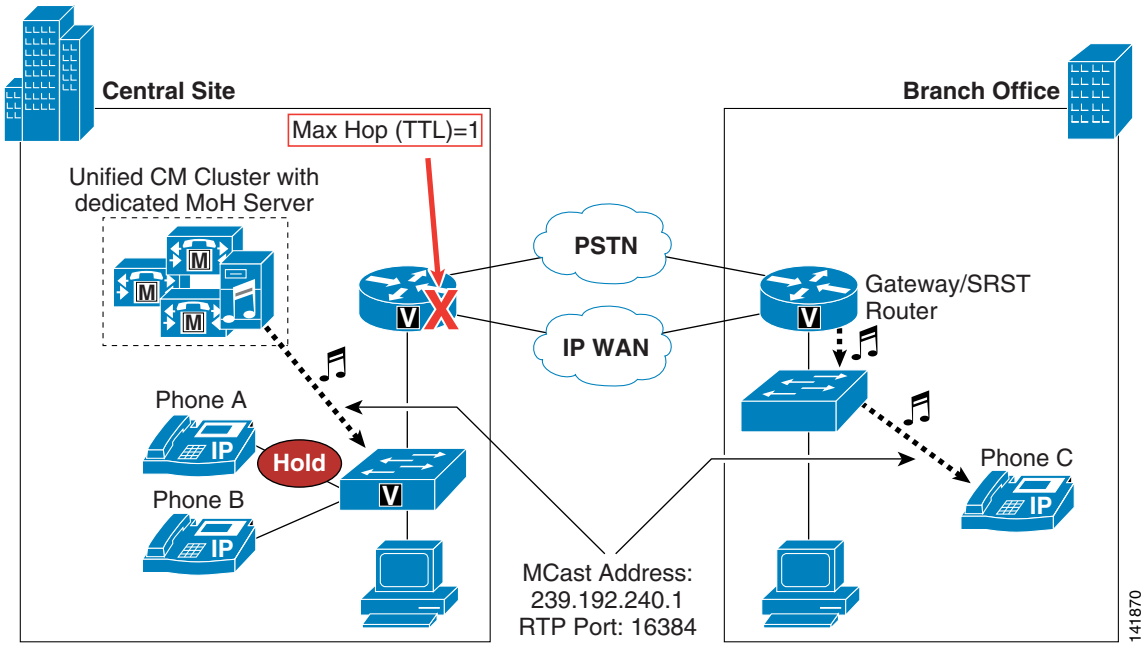
  Configure an ACL on the central-site WAN interface to disallow packets destined to the multicast group address(es) from being sent out the interface.

- Disable multicast routing on the WAN interface

  Do not configure multicast routing on the WAN interface, thus ensuring that multicast streams are not forwarded into the WAN.

Figure 7-6 illustrates streaming multicast MoH from the flash of a remote router when it is not in SRST mode. After phone A places phone C on hold, phone C receives multicast MoH from the local SRST router. In this figure, the MoH server is streaming a multicast audio source to 239.192.240.1 (on RTP port 16384), however this stream has been limited to a maximum hop of one (1) to ensure that it will not travel off the local MoH server's subnet and across the WAN. At the same time, the branch office SRST router/gateway is multicasting an audio stream from flash. This stream is also using 239.192.240.1 as its multicast address and 16384 as the RTP port number. When phone A presses the Hold softkey, phone C receives the MoH audio stream sourced by the SRST router.

Figure 7-6 Multicast MoH from Branch Router Flash



When using this method for delivering multicast MoH, configure all devices within the Cisco Unified CallManager cluster to use the same user hold and network hold audio source and configure all branch routers with the same multicast group address and port number. Because the user or network hold audio source of the holder is used to determine the audio source, if you configure more than one user or network hold audio source within the cluster, there is no way to guarantee that a remote holdee will always receive the local MoH stream. For example, suppose a central-site phone is configured with an audio source that uses group address 239.192.254.1 as its user and network hold audio source. If this phone places a remote device on hold, the remote device will attempt to join 239.192.254.1 even if the local router flash MoH stream is sending to multicast group address 239.192.240.1. If instead all devices in the network are configured to use the user/network hold audio source with multicast group address 239.192.240.1 and all branch routers are configured to multicast from flash on 239.192.240.1, then every remote device will receive the MoH from its local router's flash.

In networks with multiple branch routers configured to stream multicast MoH from flash, it is possible to have more than 51 unique MoH audio sources in a cluster. Each branch site router can multicast a unique audio file from flash, although all routers must multicast this audio on the same multicast group address. In addition, the central-site MoH server can multicast a MoH stream on this same multicast group address. Thus, if there are 100 branch sites each multicasting an audio file from flash, then the cluster can contain 101 unique MoH audio sources (100 branch streams and one central-site stream). If you want more than one unique audio stream in the central site, you can stream fixed/live sources from additional MoH servers or from external media servers (as described in Using Multiple Fixed or Live Audio Sources, page 7-9), but you should not configure more than one audio source per server.

Distributed Multisite Deployments

Multisite IP telephony deployments with distributed call processing typically contain WAN or MAN connections between the sites. These lower-speed links usually cause bandwidth and throughput bottlenecks. To minimize bandwidth consumption on these links, Cisco recommends use of the G.729

codec for all MoH audio streams traversing them. Because the G.729 codec is optimized for voice and not music applications, you should use G.729 only across the WAN/MAN links, where the bandwidth savings far outweighs the lower quality afforded by G.729 for MoH transport.

Unlike with centralized multisite deployments, in situations where G.711 might be required for MoH audio streams traveling across a WAN, MoH audio streams cannot be forced to G.711 in a distributed multisite deployment. Even when MoH servers are placed in a separate Cisco Unified CallManager region and the G.711 codec is configured between this region and the intercluster or SIP trunk's region, the codec of the original voice call is maintained when a call between the two clusters is placed on hold by either phone. Because these intercluster calls are typically encoded using G.729 for bandwidth savings, a MoH stream from either cluster will also be encoded using G.729.

Furthermore, multicast MoH is not supported for calls between Cisco Unified CallManager clusters (intercluster calls). Therefore, you must configure at least one unicast MoH resource in each Cisco Unified CallManager cluster if you want MoH on the intercluster or SIP trunk.

Proper multicast address management is another important design consideration in the distributed intercluster environment. All MoH audio source multicast addresses must be unique across all Cisco Unified CallManager clusters in the deployment to prevent possible overlap of streaming resources throughout the distributed network.

Clustering Over the WAN

As its name suggests, clustering-over-the-WAN deployments also contain the same type of lower-speed WAN links as other multisite deployments and therefore are subject to the same requirements for G.729 codec, multicast transport mechanism, and solid QoS for MoH traffic traversing these links.

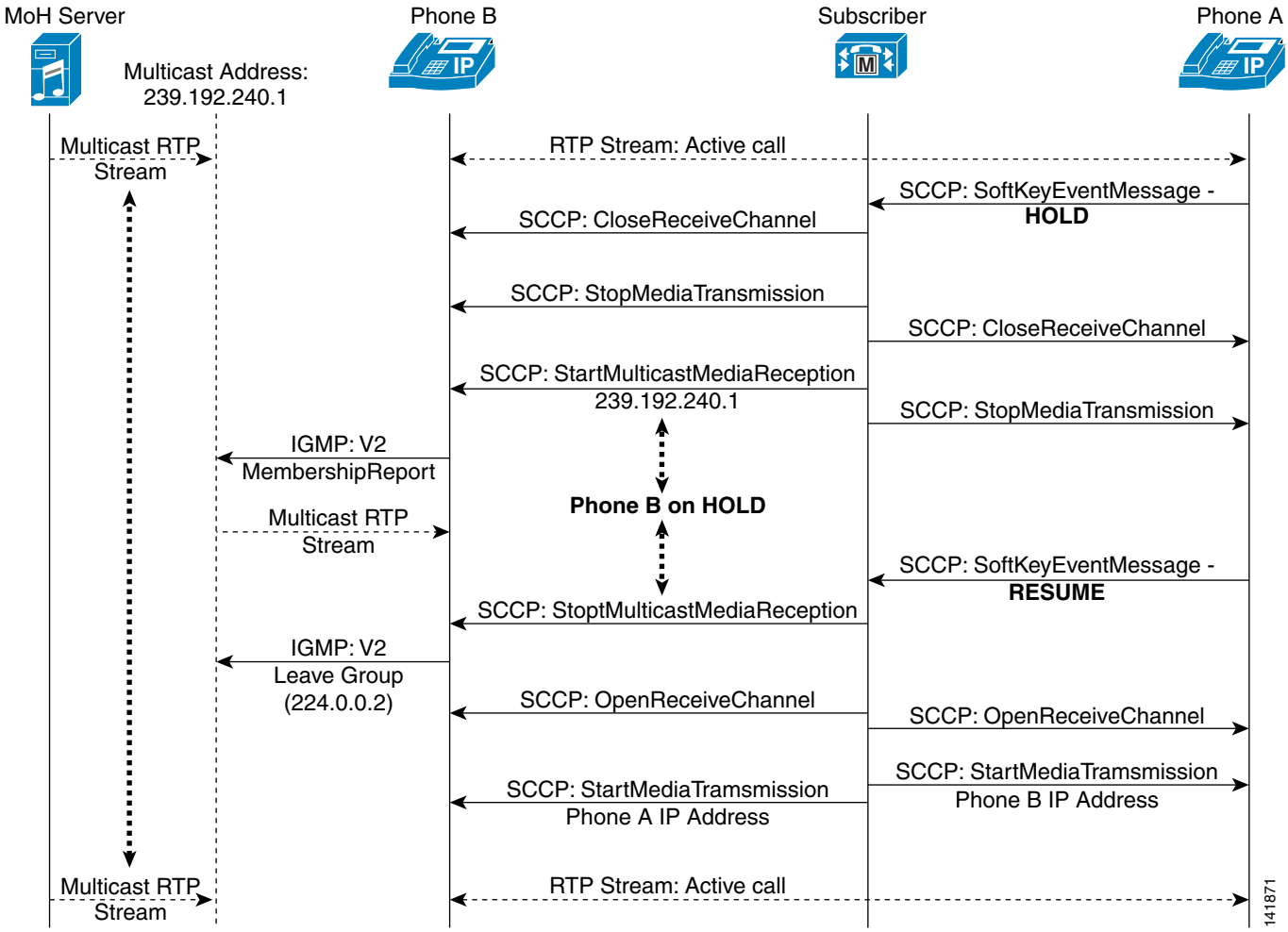
In addition, you should deploy MoH server resources at each side of the WAN in this type of configuration. In the event of a WAN failure, devices on each side of the WAN will be able to continue to receive MoH audio streams from their locally deployed MoH server. Furthermore, proper MoH redundancy configuration is extremely important. The devices on each side of the WAN should point to an MRGL whose MRG has a priority list of MoH resources with at least one local resource as the highest priority. Additional MoH resources should be configured for this MRG in the event that the primary server becomes unavailable or is unable to process requests. At least one other MoH resource in the list should point to an MoH resource on the remote side of the WAN in the event that resources at the local side of the WAN are unavailable.

Detailed Unicast and Multicast MoH Call Flows

The following sections provide detailed illustrations and explanations of unicast and multicast MoH call flows.

Figure 7-7 illustrates a typical multicast call flow. As shown in the diagram, when the Hold softkey is pressed at phone A, Cisco Unified CallManager instructs both phone A and phone B to Close Receive Channel and Stop Media Transmission. This action effectively stops the RTP two-way audio stream. Next, Cisco Unified CallManager tells phone B (the holdee) to Start Multicast Media Reception from multicast group address 239.192.240.1. The phone then issues an Internet Group Management Protocol (IGMP) V2 Membership Report message indicating that it is joining this group.

Figure 7-7 Detailed Multicast MoH Call Flow

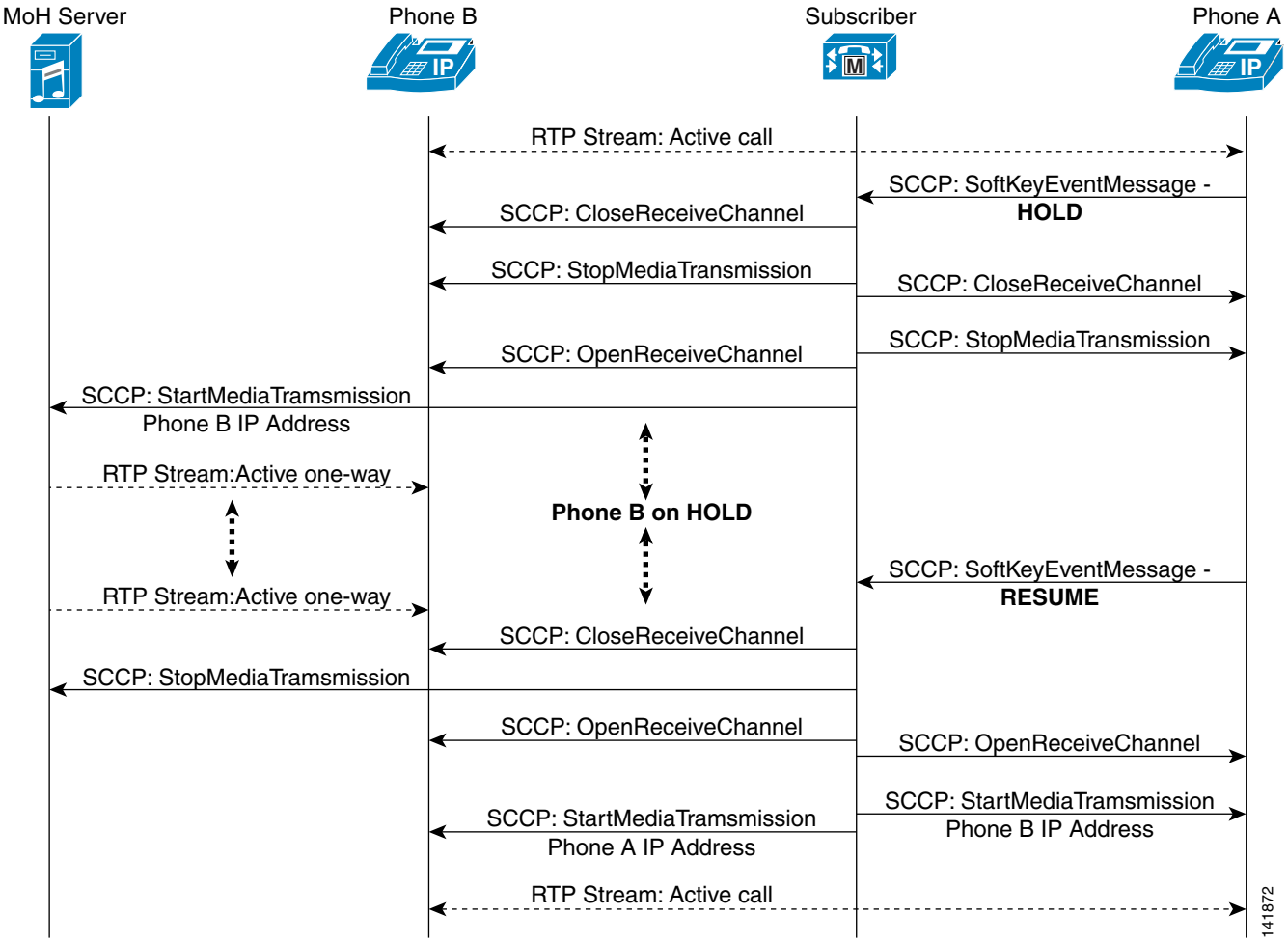


Meanwhile, the MoH server has been sourcing RTP audio to this multicast group address and, upon joining the multicast group, phone B begins receiving the MoH stream. Once phone A presses the Resume softkey, Cisco Unified CallManager instructs phone B to Stop Multicast Media Reception. Phone B then sends an IGMP V2 Leave Group message to 224.0.0.2 to indicate that the multicast stream is no longer needed. This effectively ends the MoH session. Next, Cisco Unified CallManager sends a series of Open Receive Channel messages to phones A and B, just as would be sent at the beginning of a phone call between the two phones. Soon afterwards, Cisco Unified CallManager instructs both phones to Start Media Transmission to each other's IP addresses. The phones are once again connected via an RTP two-way audio stream.

Note The call flow diagrams in Figure 7-7 and Figure 7-8 assume that an initial call is up between phones A and B, with a two-way RTP audio stream. These diagrams are representative of call flows and therefore include only the pertinent traffic required for proper MoH operation. Thus, keep-alives, acknowledgements, and other miscellaneous traffic have been eliminated to better illustrate the interaction. The initial event in each diagram is the Hold softkey action performed by phone A.

Figure 7-8 depicts a unicast MoH call flow. In this call flow diagram, when the Hold softkey is pressed at phone A, Cisco Unified CallManager instructs both phone A and phone B to Close Receive Channel and Stop Media Transmission. This action effectively stops the RTP two-way audio stream. Up to this point, unicast and multicast MoH call flows behave exactly the same.

Figure 7-8 Detailed Unicast MoH Call Flow



Next, Cisco Unified CallManager tells phone B (the holdee) to Open Receive Channel. (This is quite different from the multicast case, where Cisco Unified CallManager tells the holdee to Start Multicast Media Reception.) Then Cisco Unified CallManager tells the MoH server to Start Media Transmission to the IP address of phone B. (This too is quite different behavior from the multicast MoH call flow, where the phone is prompted to join a multicast group address.) At this point, the MoH server is sending a one-way unicast RTP music stream to phone B. When phone A presses the Resume softkey, Cisco Unified CallManager instructs the MoH server to Stop Media Transmission and instructs phone B to Close Receive Channel, effectively ending the MoH session. As with the multicast scenario, Cisco Unified CallManager sends a series of Open Receive Channel messages and Start Media Transmissions messages to phones A and B with each other's IP addresses. The phones are once again connected via an RTP two-way audio stream.