CHAPTER 17

Media Resources

Revised: April 30, 2012; OL-21733-18

A media resource is a software-based or hardware-based entity that performs media processing functions on the data streams to which it is connected. Media processing functions include mixing multiple streams to create one output stream (conferencing), passing the stream from one connection to another (media termination point), converting the data stream from one compression type to another (transcoding), streaming music to callers on hold (music on hold), echo cancellation, signaling, voice termination from a TDM circuit (coding/decoding), packetization of a stream, streaming audio (annunciation), and so forth. The software-based resources are provided by the Unified CM IP Voice Media Streaming Service (IP VMS). Digital signal processor (DSP) cards provide both software and hardware based resources.

This chapter explains the overall Media Resources Architecture and Cisco IP Voice Media Streaming Application service, and it focuses on the following media resources:

Use this chapter to gain an understanding of the function and capabilities of each media resource type and to determine which resource would be required for your deployment.

For proper DSP sizing of Cisco Integrated Service Router (ISR) gateways, you can use the Cisco Unified Communications Sizing Tool (Unified CST), available to Cisco employees and partners at http://tools.cisco.com/cucst. If you are not a Cisco partner or employee, you can use the DSP Calculator at http://www.cisco.com/go/dspcalculator. For other Cisco non-ISR gateway platforms (such as the Cisco 1700, 2600, 3700, and AS5000 Series) and/or Cisco IOS releases preceding and up to 12.4 mainline, you can access the legacy DSP calculator at http://www.cisco.com/pcgi-bin/Support/DSP/cisco_dsp_calc.pl.

What's New in This Chapter

Table 17-1 lists the topics that are new in this chapter or that have changed significantly from previous releases of this document.

Table 17-1 New or Changed Information Since the Previous Release of This Document

New or Revised Topic	Described in:	Revision Date
Numerous updates and general reorganization of the chapter	All sections of this chapter	April 30, 2012
Minor corrections and changes	Various sections throughout this chapter	June 30, 2011
Cisco Business Edition capability	Various sections throughout this chapter	February 28, 2011
DSP calculator tools for DSP sizing for Cisco ISR gateways and legacy platforms	Various sections throughout this chapter	February 28, 2011
Updates to DSP hardware and Cisco IOS release versions	Various sections throughout this chapter	February 28, 2011
Music on Hold (MoH) information has been integrated into this chapter, and the separate MoH chapter has been eliminated from this document.	Various sections throughout this chapter	November 15, 2010
Cisco Integrated Services Routers Generation 2 (ISR G2) platforms	Various sections throughout this chapter	April 2, 2010
PVDM3 DSPs	Various sections throughout this chapter	April 2, 2010

Media Resources Architecture

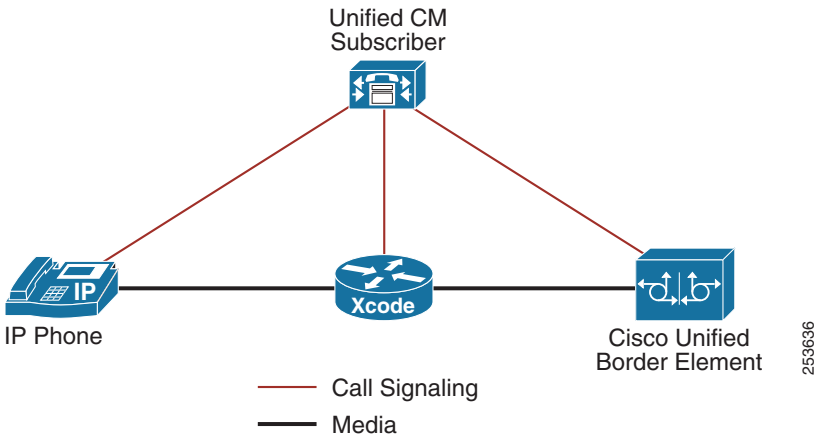
To properly design the media resource allocation strategy for an enterprise, it is critical to understand the Cisco Unified CM architecture for the various media resource components. The following sections highlight the important characteristics of media resource design with Unified CM.

Media Resource Manager

The Media Resource Manager (MRM), a software component in the Unified CM, determines whether a media resource needs to be allocated and inserted in the media path. This media resource may be provided by the Unified CM IP Voice Media Streaming Application service or by digital signal processor (DSP) cards. When the MRM decides and identifies the type of the media resource, it searches through the available resources according to the configuration settings of the media resource group list (MRGL) and media resource groups (MRGs) associated with the devices in question. MRGLs and MRGs are constructs that hold related groups of media resources together for allocation purposes and are described in detail in the section on Media Resource Groups and Lists, page 17-39.

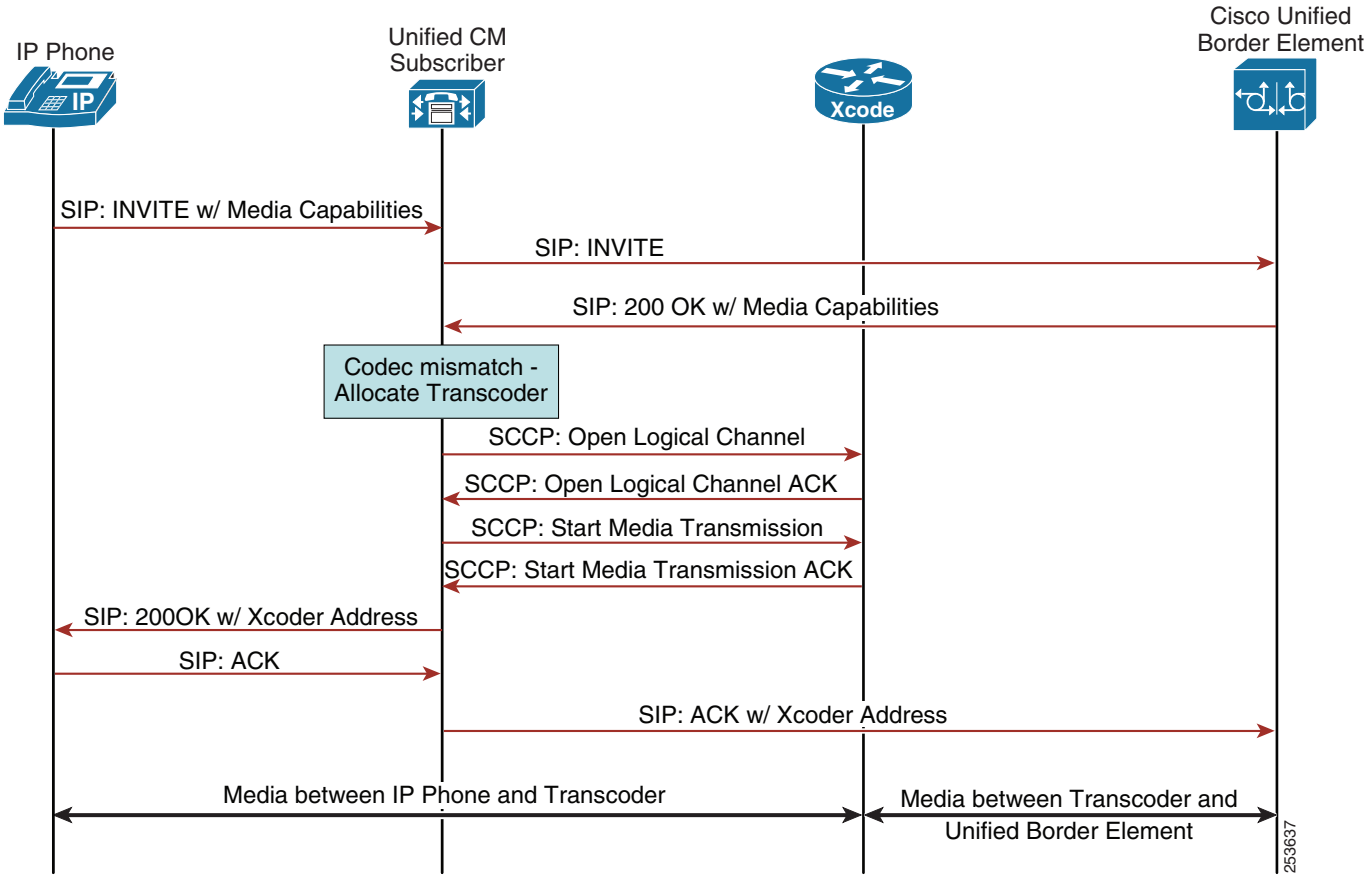
Figure 17-1 shows how a media resource such as a transcoder may be placed in the media path between an IP phone and a Cisco Unified Border Element when a common codec between the two is not available.

Figure 17-1 Use of a Transcoder Where a Common Codec Is Not Available



Unified CM communicates with media resources using Skinny Client Control Protocol (SCCP). This messaging is independent of the protocol that might be in use between Unified CM and the communicating entities. Figure 17-2 shows an example of the message flow, but it does not show all of the SCCP or SIP messages exchanged between the entities.

Figure 17-2 Message Flow Between Components



Cisco IP Voice Media Streaming Application

The Cisco IP Voice Media Streaming Application provides the following software-based media resources:

- Conference bridge
- Music on Hold (MoH)
- Annunciator
- Media termination point (MTP)

The details of these resources are covered in the respective sections below.

When the IP Voice Media Streaming Application is activated, one of each of the above resources is automatically configured. Conferencing, annunciator, and MTP services can be disabled if required. If these resources are not needed, Cisco recommends that you disable them by modifying the appropriate service parameter in the Unified CM configuration. The service parameters have default settings for the maximum number of connections that each service can handle. For details on how to modify the service parameters, refer to the appropriate version of the *Cisco Unified Communications Manager Administration Guide*, available at

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

Give careful consideration to situations that require multiple resources and to the load they place on the IP Voice Media Streaming Application. The media resources can reside on the same server as Unified CM or on a dedicated server not running the Unified CM call processing service. If your deployment requires more than the default number of any resource, Cisco recommends that you configure that resource to run on its own dedicated server. If heavy use of media resources is expected within a deployment, Cisco recommends deploying dedicated Unified CM media resource nodes (non-publisher nodes that do not perform call processing within the cluster) or relying on hardware-based media resources. Software-based media resources on Unified CM nodes are intended for small deployments or deployments where need for media resources is limited.

Note Cisco Business Edition 3000 provides only MoH and annunciator software-based media resources. No software-based media resources are available for conferencing and MTP. Hardware-based conferencing and MTP resources are provided by the DSP cards on board the Cisco MCS 7890 C2 platform.

Voice Termination

Voice termination applies to a call that has two call legs, one leg on a time-division multiplexing (TDM) interface and the second leg on a Voice over IP (VoIP) connection. The TDM leg must be terminated by hardware that performs encoding/decoding and packetization of the stream. This termination function is performed by a digital signal processor (DSP) resource residing in the same hardware module, blade, or platform.

All DSP hardware on Cisco TDM gateways is capable of terminating voice streams, and certain hardware is also capable of performing other media resource functions such as conferencing or transcoding (see Conferencing, page 17-6 and Transcoding, page 17-9). The DSP hardware has either fixed DSP resources that cannot be upgraded or changed, or modular DSP resources that can be upgraded.

The number of supported calls per DSP depends on the computational complexity of the codec used for a call and also on the complexity mode configured on the DSP. Cisco IOS enables you to configure a complexity mode on the hardware module. Hardware platforms such as the PVDM2 and PVDM3 DSPs support three complexity modes: medium, high and flex mode. Some of the other hardware platforms support only medium and high complexity modes.

Medium and High Complexity Mode

You can configure each DSP separately as either medium complexity, high complexity, or flex mode (PVDM3 DSPs and those based on C5510). The DSP treats all calls according to its configured complexity, regardless of the actual complexity of the codec of the call. A resource with configured complexity equal or higher than the actual complexity of the incoming call must be available, or the call will fail. For example, if a call requires a high-complexity codec but the DSP resource is configured for medium complexity mode, the call will fail. However, if a medium-complexity call is attempted on a DSP configured for high complexity mode, then the call will succeed and Cisco IOS will allocate a high-complexity mode resource.

Flex Mode

Flex mode, available on hardware platforms that use the C5510 chipset and on PVDM3 DSPs, eliminates the requirement to specify the codec complexity at configuration time. A DSP in flex mode accepts a call of any supported codec type, as long as it has available processing power.

For C5510-based DSPs, the overhead of each call is tracked dynamically via a calculation of processing power in millions of instructions per second (MIPS). Cisco IOS performs a MIPS calculation for each call received and subtracts MIPS credits from its budget whenever a new call is initiated. The number of MIPS consumed by a call depends on the codec of the call. The DSP will allow a new call as long as it has remaining MIPS credits greater than or equal to the MIPS required for the incoming call.

Similarly, PVDM3 DSP modules use a credit-based system. Each module is assigned a fixed number of "credits" that represent a measure of its capacity to process media streams. Each media operation, such as voice termination, transcoding, and so forth, is assigned a cost in terms of credits. As DSP resources are allocated for a media processing function, its cost value is subtracted from the available credits. A DSP module runs out of capacity when the available credits run out and are no longer sufficient for the requested operation. The credit allocation rules for PVDM3 DSPs are rather complex.

For proper DSP sizing of Cisco ISR gateways, you can use the Cisco Unified Communications Sizing Tool (Unified CST), available to Cisco employees and partners at http://tools.cisco.com/cucst. If you are not a Cisco partner, you can use the DSP Calculator at http://www.cisco.com/go/dspcalculator. For other Cisco non-ISR gateway platforms (such as the Cisco 1700, 2600, 3700, and AS5000 Series) and/or Cisco IOS releases preceding and up to 12.4 mainline, you can access the legacy DSP calculator at http://www.cisco.com/cgi-bin/Support/DSP/cisco_dsp_calc.pl.

Flex mode has an advantage when calls of multiple codecs must be supported on the same hardware because flex mode can support more calls than when the DSPs are configured as medium or high complexity. However, flex mode does allow oversubscription of the resources, which introduces the risk of call failure if all resources are used. With flex mode it is possible to have fewer DSP resources than with physical TDM interfaces.

Compared to medium or high complexity mode, flex mode has the advantage of supporting the most G.711 calls per DSP. For example, a PVDM2-16 DSP can support 8 G.711 calls in medium complexity mode or 16 G.711 calls in flex mode.

Conferencing

A conference bridge is a resource that joins multiple participants into a single call (audio or video). It can accept any number of connections for a given conference, up to the maximum number of streams allowed for a single conference on that device. There is a one-to-one correspondence between media streams connected to a conference and participants connected to the conference. The conference bridge mixes the streams together and creates a unique output stream for each connected party. The output stream for a given party is the composite of the streams from all connected parties minus their own input stream. Some conference bridges mix only the three loudest talkers on the conference and distribute that composite stream to each participant (minus their own input stream if they are one of the talkers).

Audio Conferencing

Audio conferencing can be performed by both software-based and hardware-based conferencing resources. A hardware conference bridge has all the capabilities of a software conference bridge. In addition, some hardware conference bridges can support multiple low bit-rate (LBR) stream types such as G.729 or G.723. This capability enables some hardware conference bridges to handle mixed-mode conferences. In a mixed-mode conference, the hardware conference bridge transcodes G.729 and G.723 streams into G.711 streams, mixes them, and then encodes the resulting stream into the appropriate stream type for transmission back to the user. Some hardware conference bridges support only G.711 conferences.

All conference bridges that are under the control of Cisco Unified Communications Manager (Unified CM) use Skinny Client Control Protocol (SCCP) to communicate with Unified CM.

Unified CM allocates a conference bridge from a conferencing resource that is registered with the Unified CM cluster. Both hardware and software conferencing resources can register with Unified CM at the same time, and Unified CM can allocate and use conference bridges from either resource. Unified CM does not distinguish between these types of conference bridges when it processes a conference allocation request.

The number of individual conferences that may be supported by the resource varies, and the maximum number of participants in a single conference varies, depending on the resource.

The following types of conference bridge resources may be used on a Unified CM system:

- Software Audio Conference Bridge (Cisco IP Voice Media Streaming Application), page 17-6
- Hardware Audio Conference Bridge (Cisco NM-HDV2, NM-HD-1V/2V/2VE, PVDM2, and PVDM3 DSPs), page 17-7
- Hardware Audio Conference Bridge (Cisco WS-SVC-CMM-ACT), page 17-7
- Hardware Audio Conference Bridge (Cisco NM-HDV and 1700 Series Routers), page 17-7

Software Audio Conference Bridge (Cisco IP Voice Media Streaming Application)

A software unicast conference bridge is a standard conference mixer that is capable of mixing G.711 audio streams and Cisco Wideband audio streams. Any combination of Wideband or G.711 a-law and mu-law streams may be connected to the same conference. The number of conferences that can be supported on a given configuration depends on the server where the conference bridge software is running and on what other functionality has been enabled for the application. The Cisco IP Voice Media Streaming Application is a resource that can also be used for several functions, and the design must consider all functions together (see Cisco IP Voice Media Streaming Application, page 17-4).

Hardware Audio Conference Bridge (Cisco NM-HDV2, NM-HD-1V/2V/2VE, PVDM2, and PVDM3 DSPs)

DSPs that are configured through Cisco IOS as conference resources will load firmware into the DSPs that are specific to conferencing functionality only, and these DSPs cannot be used for any other media feature. Any PVDM2 or PVDM3 based hardware, such as the NM-HDV2, may be used simultaneously in a single chassis for voice termination but may not be used simultaneously for other media resource functionality. The DSPs based on PVDM-256K and PVDM2 have different DSP farm configurations, and only one may be configured in a router at a time. DSPs on PVDM2 hardware are configured individually as voice termination, conferencing, media termination, or transcoding, so that DSPs on a single PVDM may be used as different resource types. Allocate DSPs to voice termination first, then to other functionality as needed.

Starting with Cisco IOS Release 12.4(15)T, the limit on the maximum number of participants has been increased to 32. A conference based on these DSPs can be configured to have a maximum of 8, 16, or 32 participants. The DSP resources for a conference are reserved during configuration, based on the profile attributes and irrespective of how many participants actually join. Refer to the following module data sheets for accurate information on module capacity and capabilities:

- For capacity information on PVDM2 modules, refer to the *High-Density Packet Voice Digital Signal Processor Module for Cisco Unified Communications Solutions* data sheet, available at

  http://www.cisco.com/en/US/prod/collateral/routers/ps5854/product_data_sheet0900aecd8016e845_ps3115_Products_Data_Sheet.html

- For capacity information on PVDM3 modules, refer to the *High-Density Packet Voice Video Digital Signal Processor Module for Cisco Unified Communications Solutions* data sheet, available at

  http://www.cisco.com/en/US/prod/collateral/modules/ps3115/data_sheet_c78-553971.html

Note The integrated gateway on the Cisco MCS 7890 C2 platform for Cisco Business Edition 3000 supports up to 24 conference streams.

Hardware Audio Conference Bridge (Cisco WS-SVC-CMM-ACT)

The following guidelines and considerations apply to this DSP resource:

- DSPs on this hardware are configured individually as voice termination, conferencing, media termination, or transcoding, so that DSPs on a single module may be used as different resource types. Allocate DSPs to voice termination first.
- Each ACT Port Adaptor contains 4 DSPs that are individually configurable. Each DSP can support 32 conference participants. You can configure up to 4 ACT Port Adaptors per CMM Module.
- This Cisco Catalyst-based hardware provides DSP resources that can provide conference bridges of up to 128 participants per bridge. A conference bridge may span multiple DSPs on a single ACT Port Adaptor; but conference bridges cannot span across multiple ACT Port Adaptors.
- The G.711 and G.729 codecs are supported on these conference bridges without extra transcoder resources. However, transcoder resources would be necessary if other codecs are used.

Hardware Audio Conference Bridge (Cisco NM-HDV and 1700 Series Routers)

The following guidelines and considerations apply to these DSP resources:

- This hardware utilizes the PVDM-256K type modules that are based on the C549 DSP chipset.
- Conferences using this hardware provide bridges that allow up to 6 participants in a single bridge.
- The resources are configured on a per-DSP basis as conference bridges.

- The NM-HDV may have up to 5 PVDM-256K modules, while the Cisco 1700 Series Routers may have 1 or 2 PVDM-256K modules.
- Each DSP provides a single conference bridge that can accept G.711 or G.729 calls.
- The Cisco 1751 is limited to 5 conference calls per chassis, and the Cisco 1760 can support 20 conference calls per chassis.

Note Any PVDM2-based hardware, such as the NM-HDV2, may be used simultaneously in a single chassis for voice termination but may not be used simultaneously for other media resource functionality. The DSPs based on PVDM-256K and PVDM2 have different DSP farm configurations, and only one may be configured in a router at a time.

Video Conferencing

Video-capable endpoints provide the capability to conduct video conferences that function similar to audio conferences. Video conferences can be invoked as ad-hoc conferences from a Skinny Client Control Protocol (SCCP) device through the use of Conf, Join, or cBarge softkeys.

The video portion of the conference can operate in either of two modes:

- Voice activation

  In this mode, the video endpoints display the dominant participant (the one speaking most recently or speaking the loudest). In this way, the video portion follows or tracks the audio portion. This mode is optimal when one participant speaks most of the time, as with an instructor teaching or training a group.

- Continuous presence

  In this mode, input from all (or selected) video endpoints is displayed simultaneously and continuously. The audio portion of the conference follows or tracks the dominant speaker. Continuous presence is more popular, and it is optimal for conferences or discussions between speakers at various sites.

Videoconferencing resources are of two types:

- Software videoconferencing bridges

  Software videoconferencing bridges process video and audio for the conference using just software. Cisco Unified MeetingPlace Express Media Server is a software videoconferencing bridge that can support ad-hoc video conferences. Cisco Unified MeetingPlace Express Media Server supports only voice activation mode for video conferences.

- Hardware videoconferencing bridges

  Hardware videoconferencing bridges have hardware DSPs that are used for the video conferences. The Cisco 3500 Series Multipoint Control Units (MCUs) and, starting with Cisco IOS Release 15.1.4M, the PVDM3 DSPs provide this type of videoconferencing bridge. Most hardware videoconferencing bridges can also be used as audio-only conference bridges. Hardware videoconferencing bridges provide the advantages of video transrating, higher video resolution, and scalability.

Videoconferencing bridges can be configured in a manner similar to audio conferencing resources, with similar characteristics for media resource groups (MRGs) and media resource group lists (MRGLs) for the device pools or endpoints.

Cisco Unified CM includes the Intelligent Bridge Selection feature, which provides a method for selecting conference resources based on the capabilities of the endpoints in the conference. For additional details on this functionality, see Intelligent Bridge Selection, page 12-20.

Secure Conferencing

Secure conferencing is a way to use regular conferencing to ensure that the media for the conference is secure and cannot be compromised. There are various security levels that a conference can have, such as authenticated or encrypted. With secure conferencing, the devices and conferencing resource can be authenticated to be trusted devices, and the conference media can then be encrypted so that every authenticated participant sends and received encrypted media for that conference. In most cases the security level of the conference will depend on the lowest security level of the participants in the conference. For example, if there is one participant who is not using a secure endpoint, then the entire conference will be non-secure. As another example, if one of the endpoints is authenticated but does not do encryption, then the conference will be in authenticated mode.

Secure conferencing provides conferencing functionality at an enhanced security level and prevents unauthorized capture and decryption of conference calls.

Consider the following factors when designing secure conferencing:

- Security levels of devices (phones and conferencing resources)
- Security overhead for call signaling and secure (SRTP) media
- Bandwidth utilization impact if secure participants are across the WAN
- Any intermediate devices such as NAT and firewalls that might not support secure calls across them

Secure conferencing is subject to the following restrictions and limitations:

- Secure conferencing is supported only for audio conferencing; video conferencing is not supported.
- With secure conferencing, Cisco IOS DSPs support a maximum of 8 participants in a conference.
- Secure conferencing may also use more DSP resources than non-secure conferencing, so DSPs must be provisioned according to the DSP Calculator.
- Some protocols may rely on IPSec to secure the call signaling.
- Secure conferencing cannot be cascaded between Unified CM and Unified CM Express.
- MTPs and transcoders do not support secure calls. Therefore, a conference might no longer be secure if any call into that conference invokes an MTP or a transcoder.
- An elaborate security policy might be needed.
- Secure conferencing might not be available for all codecs.

Transcoding

A transcoder is a device that converts an input stream from one codec into an output stream that uses a different codec. Starting with Cisco IOS Release 15.0.1M, a transcoder also supports transrating, whereby it connects two streams that utilize the same codec but with a different packet size.

Transcoding from G.711 to any other codec is referred to as traditional transcoding. Transcoding between any two non-G.711 codecs is called universal transcoding and requires Universal Cisco IOS transcoders. Universal transcoding is supported starting with Cisco IOS Release 12.4.20T. Universal transcoding has a lower DSP density than traditional transcoding.

In a Unified CM system, the typical use of a transcoder is to convert between a G.711 voice stream and the low bit-rate compressed voice stream G729a. The following cases determine when transcoder resources are needed:

- Single codec for the entire system

  A single codec is generally used in a single-site deployment that usually has no need for conserving bandwidth. When a single codec is configured for all calls in the system, then no transcoder resources are required. In this scenario, G.711 is the most common choice that is supported by all vendors.

- Multiple codecs in use in the system, with all endpoints capable of all codec types

  The most common reason for multiple codecs is to use G.711 for LAN calls to maximize the call quality and to use a low-bandwidth codec to maximize bandwidth efficiency for calls that traverse a WAN with limited bandwidth. Cisco recommends using G.729a as the low-bandwidth codec because it is supported on all Cisco Unified IP Phone models as well as most other Cisco Unified Communications devices, therefore it can eliminate the need for transcoding. Although Unified CM allows configuration of other low-bandwidth codecs between regions, some phone models do not support those codecs and therefore would require transcoders. They would require one transcoder for a call to a gateway and two transcoders if the call is to another IP phone. The use of transcoders is avoided if all devices support and are configured for both G.711 and G.729 because the devices will use the appropriate codec on a call-by-call basis.

- Multiple codecs in use in the system, and some endpoints support or are configured for G.711 only

  This condition exists when G.729a is used in the system but there are devices that do not support this codec, or a device with G.729a support may be configured to not use it. In this case, a transcoder is also required. Devices from some third-party vendors may not support G.729.

Note Cisco Unified MeetingPlace Express prior to release 2.0 supported G.711 only. In an environment where G.729 is configured for a call into earlier versions of Cisco Unified MeetingPlace Express, transcoder resources are required.

A transcoder is also capable of performing the same functionality as a media termination point (MTP). In cases where transcoder functionality and MTP functionality are both needed, a transcoder is allocated by the system. If MTP functionality is required, Unified CM will allocate either a transcoder or an MTP from the resource pool, and the choice of resource will be determined by the media resource groups, as described in the section on Media Resource Groups and Lists, page 17-39.

To finalize the design, it is necessary to know how many transcoders are needed and where they will be placed. For a multi-site deployment, Cisco recommends placing a transcoder local at each site where it might be required. If multiple codecs are needed, it is necessary to know how many endpoints do not support all codecs, where those endpoints are located, what other groups will be accessing those resources, how many maximum simultaneous calls these device must support, and where those resources are located in the network.

Transcoding Resources

DSP resources are required to perform transcoding. Those DSP resources can be located in the voice modules and the hardware platforms for transcoding that are listed in the following sections.

Hardware Transcoder (Cisco NM-HDV2, NM-HD-1V/2V/2VE, and PVDM2 DSPs)

The number of sessions supported on each DSP is determined by the codecs used in universal transcoding mode. The following guidelines and considerations apply to these DSP resources:

- Transcoding is available between G.711 mu-law or a-law and G.729a, G.729ab, G.722, and iLBC. A single PVDM2-16 can support 8 sessions for transcoding between low and medium complexity codecs (such as G.711 and G.729a or G.722) or 6 sessions for transcoding between low and high complexity codecs (such as G.711 and G.729 or iLBC).

Note If transcoding is not required between G.711 and G.722, Cisco recommends that you do not include G.722 in the Cisco IOS configuration of the dspfarm profile. This is to preclude Unified CM from selecting G.722 as the codec for a call in which transcoding is required. DSP resources configured as Universal Transcoders are required for transcoding between G.722 and other codecs.

- Cisco Unified IP Phones use only the G.729a variants of the G.729 codecs. The default for a new DSP farm profile is G.729a/G.729ab/G.711u/G.711a. Because a single DSP can provide only one function at a time, the maximum sessions configured on the profile should be specified in multiples of 8 to prevent wasted resources.

For capacity information on PVDM2 modules, refer to the *High-Density Packet Voice Digital Signal Processor Module for Cisco Unified Communications Solutions* data sheet, available at

http://www.cisco.com/en/US/prod/collateral/routers/ps5854/product_data_sheet0900aecd8016e845_ps3115_Products_Data_Sheet.html

Hardware Transcoder (Cisco WS-SVC-CMM-ACT)

The following guidelines and considerations apply to this DSP resource:

- Transcoding is available between G.711 mu-law or a-law and G.729, G.729b, or G.723.
- There are 4 DSPs per ACT that may be allocated individually to DSP pools.
- The CCM-ACT can have 16 transcoded calls per DSP or 64 per ACT. The ACT reports resources as streams rather than calls, and a single transcoded call consists of two streams.

Hardware Transcoder (Cisco NM-HDV and 1700 Series Routers)

The following guidelines and considerations apply to these DSP resources:

This hardware utilizes the PVDM-256K type modules, and each DSP provides 2 transcoding sessions.

- The NM-HDV may have up to 4 PVDM-256K modules, and the Cisco 1700 Series Routers may have 1 or 2 PVDM-256K modules. The Cisco 1751 Router has a chassis limit of 16 sessions, and the Cisco 1760 Router has a chassis limit of 20 sessions.
- NM-HDV and NM-HDV2 modules may be used simultaneously in a single chassis for voice termination but may not be used simultaneously for other media resource functionality. Only one type of DSP farm configuration may be active at one time (either the NM-HDV or the HM-HDV2) for conferencing, MTP, or transcoding.

- Transcoding is supported from G.711 mu-law or a-law to any of G.729, G.729a, G.729b, or G.729ab codecs.

Hardware Transcoder (PVDM3 DSP)

PVDM3 DSPs are hosted by Cisco 2900 Series and 3900 Series Integrated Services Routers, and they support both secure and non-secure transcoding from any and to any codec. As with voice termination and conferencing, each transcoding session debits the available credits for each type of PVDM3 DSPs. The available credits determine the total capacity of the DSP.

For example, a PVDM3-16 can support 12 sessions for transcoding between low and medium complexity codecs (such as G.711 and G.729a or G.722) or 10 sessions for transcoding between low and high complexity codecs (such as G.711 and G.729 or iLBC).

Note For Cisco Business Edition 3000, the default gateway configuration will support only 10 transcoding sessions per Cisco MCS 7890 appliance.

For capacity information on PVDM3 modules, refer to the *High-Density Packet Voice Video Digital Signal Processor Module for Cisco Unified Communications Solutions* data sheet, available at

http://www.cisco.com/en/US/prod/collateral/modules/ps3115/data_sheet_c78-553971.html

Media Termination Point (MTP)

A media termination point (MTP) is an entity that accepts two full-duplex media streams. It bridges the streams together and allows them to be set up and torn down independently. The streaming data received from the input stream on one connection is passed to the output stream on the other connection, and vice versa. MTPs have many possible uses, such as:

- Re-Packetization of a Stream, page 17-12
- DTMF Conversion, page 17-12
- Protocol-specific usage
  - DTMF Relay over SIP Trunks, page 17-14
  - H.323 Supplementary Services, page 17-17
  - H.323 Outbound Fast Connect, page 17-17

Re-Packetization of a Stream

An MTP can be used to transcode G.711 a-law audio packets to G.711 mu-law packets and vice versa, or it can be used to bridge two connections that utilize different packetization periods (different sample sizes). Note that re-packetization requires DSP resources in a Cisco IOS MTP.

DTMF Conversion

DTMF tones are used during a call to signal to a far-end device for purposes of navigating a menu system, entering data, or other manipulation. They are processed differently than DTMF tones sent during a call setup as part of the call control. There are several methods for sending DTMF over IP, and two communicating endpoints might not support a common procedure. In these cases, Unified CM may

dynamically insert an MTP in the media path to convert DTMF signals from one endpoint to the other. Unfortunately, this method does not scale because one MTP resource is required for each such call. The following sections help determine the optimum amount of MTP resources required, based on the combination of endpoints, trunks, and gateways in the system.

If Unified CM determines that an MTP needs to be inserted but no MTP resources are available, it uses the setting of the service parameter **Fail call if MTP allocation fails** to decide whether or not to allow the call to proceed.

DTMF Relay Between Endpoints

The following methods are used to relay DTMF from one endpoint to another.

Named Telephony Events (RFC 2833)

Named Telephony Events (NTEs) defined by RFC 2833 are a method of sending DTMF from one endpoint to another after the call media has been established. The tones are sent as packet data using the already established RTP stream and are distinguished from the audio by the RTP payload type field. For example, the audio of a call can be sent on a session with an RTP payload type that identifies it as G.711 data, and the DTMF packets are sent with an RTP payload type that identifies them as NTEs. The consumer of the stream utilizes the G.711 packets and the NTE packets separately.

Key Press Markup Language (RFC 4730)

The Key Press Markup Language (KPML) is defined in RFC 4730. Unlike NTEs, which is an in-band method of sending DTMF, KPML uses the signaling channel (out-of-band, or OOB) to send SIP messages containing the DTMF digits.

KPML procedures use a SIP SUBSCRIBE message to register for DTMF digits. The digits themselves are delivered in NOTIFY messages containing an XML encoded body.

Unsolicited Notify (UN)

Unsolicited Notify procedures are used primarily by Cisco IOS SIP Gateways to transport DTMF digits using SIP NOTIFY messages. Unlike KPML, these NOTIFY messages are unsolicited, and there is no prior registration to receive these messages using a SIP SUBSCRIBE message. But like KPML, Unsolicited Notify messages are out-of-band.

Also unlike KPML, which has an XML encoded body, the message body in these NOTIFY messages is a 10-character encoded digit, volume, and duration, describing the DTMF event.

H.245 Signal, H.245 Alphanumeric

H.245 is the media control protocol used in H.323 networks. In addition to its use in negotiating media characteristics, H.245 also provides a channel for DTMF transport. H.245 utilizes the signaling channel and, hence, provides an out-of-band (OOB) way to send DTMF digits. The Signal method carries more information about the DTMF event (such as its actual duration) than does Alphanumeric.

Cisco Proprietary RTP

This method sends DTMF digits in-band, that is, in the same stream as RTP packets. However, the DTMF packets are encoded differently than the media packets and use a different payload type. This method is not supported by Unified CM but is supported on Cisco IOS Gateways.

Skinny Client Control Protocol (SCCP)

The Skinny Client Control Protocol is used by Unified CM for controlling the various SCCP-based devices registered to it. SCCP defines out-of-band messages that transport DTMF digits between Unified CM and the controlled device.

DTMF Relay Between Endpoints in the Same Unified CM Cluster

The following rules apply to endpoints registered to Unified CM servers in the same cluster:

- Calls between two non-SIP endpoints do not require MTPs.

  All Cisco Unified Communications endpoints other than SIP send DTMF to Unified CM via various signaling paths, and Unified CM forwards the DTMF between dissimilar endpoints. For example, an IP phone may use SCCP messages to Unified CM to send DTMF, which then gets sent to an H.323 gateway via H.245 signaling events. Unified CM provides the DTMF forwarding between different signaling types.

- Calls between two Cisco SIP endpoints do not require MTPs.

  All Cisco SIP endpoints support NTE, so DTMF is sent directly between endpoints and no conversion is required.

- A combination of a SIP endpoint and a non-SIP endpoint might require MTPs.

  To determine the support for NTE in your devices, refer to the product documentation for those devices. Support of NTE is not limited to SIP and can be supported in devices with other call control protocols. Unified CM has the ability to allocate MTPs dynamically on a call-by-call basis, based on the capabilities of the pair of endpoints.

DTMF Relay over SIP Trunks

A SIP trunk configuration is used to set up communication with a SIP User Agent such as another Cisco Unified CM cluster or a SIP gateway.

SIP negotiates media exchange via Session Description Protocol (SDP), where one side offers a set of capabilities to which the other side answers, thus converging on a set of media characteristics. SIP allows the initial offer to be sent either by the caller in the initial INVITE message (Early Offer) or, if the caller chooses not to, the called party can send the initial offer in the first reliable response (Delayed Offer).

By default, Unified CM SIP trunks send the INVITE without an initial offer (Delayed Offer). Unified CM has two configurable options to enable a SIP trunk to send the offer in the INVITE (Early Offer):

- Media Termination Point Required

  Checking this option on the SIP trunk assigns an MTP for every outbound call. This statically assigned MTP supports only the G.711 codec or the G.729 codec, thus limiting media to voice calls only.

- Early Offer support for voice and video calls (insert MTP if needed)

  Checking this option on the SIP Profile associated with the SIP Trunk inserts an MTP only if the calling device cannot provide Unified CM with the media characteristics required to create the Early Offer (for example, where an inbound call to Unified CM is received on a Delayed Offer SIP trunk or a Slow Start H.323 trunk). This option is available only with Unified CM 8.5 and later releases.

In general, Cisco recommends **Early Offer support for voice and video calls (insert MTP if needed)** because this configuration option reduces MTP usage. Calls from older SCCP phones registered to Unified CM over SIP Early Offer trunks use an MTP to create the Offer SDP. These calls support voice, video, and encryption. Inbound calls to Unified CM from SIP Early Offer trunks or H.323 Slow Start

trunks that are extended over a SIP Early Offer trunk use an MTP to create the Offer SDP. However, these calls support audio only in the initial call setup, but they can be escalated to support video mid-call if the called or calling device invokes it.

Also note that MTP resources are not required for incoming INVITE messages, whether or not they contain an initial offer.

Whether or not an MTP will be allocated by Unified CM depends on the capabilities of the communicating endpoints and the configuration on the intermediary device, if any. For example, the SIP trunk may be configured to handle DTMF exchange in one of several ways: a SIP trunk can carry DTMF using KPML or it can instruct the communicating endpoints to use NTE.

Note As described in this section, SIP Early Offer can also be enabled by checking the **Media Termination Point Required** option on the SIP trunk. However, this option increases MTP usage because an MTP is assigned for every outbound call rather than on an as-needed basis.

SIP Trunk MTP Requirements

By default, the SIP trunk parameter **Media Termination Point Required** and the SIP Profile parameter **Early Offer support for voice and video calls (insert MTP if needed)** are not selected.

Use the following steps to determine whether MTP resources are required for your SIP trunks.

1. Is the far-end SIP device defined by this SIP trunk capable of accepting an inbound call without a SIP Early Offer?

   If not, then on the SIP Profile associated with this trunk, check the box to enable **Early Offer support for voice and video calls (insert MTP if needed)**. For outbound SIP trunk calls, an MTP will be inserted only if the calling device cannot provide Unified CM with the media characteristics required to create the Early Offer, or if DTMF conversion is needed.

   If yes, then do not check the **Early Offer support for voice and video calls (insert MTP if needed)** box, and use Step 2. to determine whether an MTP is inserted dynamically for DTMF conversion. Note that DTMF conversion can be performed by the MTP regardless of the codec in use.

   Note The option for **Early Offer support for voice and video calls (insert MTP if needed)** is available only with Unified CM 8.5 and later releases.

2. Select a Trunk DTMF Signaling Method, which controls the behavior of DTMF selection on that trunk. Available MTPs will be allocated based on the requirements for matching DTMF methods for all calls.

   a. DTMF Signaling Method: No Preference

      In this mode, Unified CM attempts to minimize the usage of MTP by selecting the most appropriate DTMF signaling method.

      If both endpoints support NTE, then no MTP is required.

      If both devices support any out-of-band DTMF mechanism, then Unified CM will use KPML or Unsolicited Notify over the SIP trunk. For example, this is the case if a Cisco Unified IP Phone 7936 using SCCP (which supports DTMF using only SCCP messaging) communicates with a Cisco Unified IP Phone 7970 using SIP (which supports DTMF using NTE and KPML) over a SIP trunk configured as described above. The only case where MTP is required is when

one of the endpoints supports out-of-band only and the other supports NTE only (for example, an SCCP Cisco Unified IP Phone 7936 communicating with a SIP Cisco Unified IP Phone 7970).

b. DTMF Signaling Method: RFC 2833

By placing a restriction on the DTMF signaling method across the trunk, Unified CM is forced to allocate an MTP if any one or both the endpoints do not support NTE. In this configuration, the only time an MTP will not be allocated is when both endpoints support NTE.

c. DTMF Signaling Method: OOB and RFC 2833

In this mode, the SIP trunk signals both KPML (or Unsolicited Notify) and NTE-based DTMF across the trunk, and it is the most intensive MTP usage mode. The only cases where MTP resources will not be required is when both endpoints support both NTE and any OOB DTMF method (KPML or SCCP).

Note Cisco Unified IP Phones play DTMF to the end user when DTMF is received via SCCP, but they do not play tones received by NTE. However, there is no requirement to send DTMF to another end user. It is necessary only to consider the endpoints that originate calls combined with endpoints that might need DTMF, such as PSTN gateways, application servers, and so forth.

DTMF Relay on SIP Gateways and Cisco Unified Border Element

Cisco SIP Gateways support KPML, NTE, or Unsolicited Notify as the DTMF mechanism, depending on the configuration. Because there may be a mix of endpoints in the system, multiple methods may be configured on the gateway simultaneously in order to minimize MTP requirements.

On Cisco SIP Gateways, configure both **sip-kpml** and **rtp-nte** as DTMF relay methods under SIP dial peers. This configuration will enable DTMF exchange with all types of endpoints, including those that support only NTE and those that support only OOB methods, without the need for MTP resources. With this configuration, the gateway will negotiate both NTE and KPML with Unified CM. If NTE is not supported by the Unified CM endpoint, then KPML will be used for DTMF exchange. If both methods are negotiated successfully, the gateway will rely on NTE to receive digits and will not subscribe to KPML.

Cisco SIP gateways also have the ability to use proprietary Unsolicited Notify (UN) method for DTMF. The UN method sends a SIP Notify message with a body that contains text describing the DTMF tone. This method is also supported on Unified CM and may be used if **sip-kpml** is not available. Configure **sip-notify** as the DTMF relay method. Note that this method is Cisco proprietary.

SIP gateways that support only NTE require MTP resources to be allocated when communicating with endpoints that do not support NTE.

H.323 Trunks and Gateways

For the H.323 gateways and trunks there are three reasons for invoking an MTP:

- H.323 Supplementary Services, page 17-17
- H.323 Outbound Fast Connect, page 17-17
- DTMF Conversion, page 17-17

H.323 Supplementary Services

MTPs can be used to extend supplementary services to H.323 endpoints that do not support the H.323v2 OpenLogicalChannel and CloseLogicalChannel request features of the Empty Capabilities Set (ECS). This requirement occurs infrequently. All Cisco H.323 endpoints support ECS, and most third-party endpoints have support as well. When needed, an MTP is allocated and connected into a call on behalf of an H.323 endpoint. When an MTP is required on an H.323 call and none is available, the call will proceed but will not be able to invoke supplementary services.

H.323 Outbound Fast Connect

H.323 defines a procedure called Fast Connect, which reduces the number of packets exchanged during a call setup, thereby reducing the amount of time for media to be established. This procedure uses Fast Start elements for control channel signaling, and it is useful when two devices that are utilizing H.323 have high network latency between them because the time to establish media depends on that latency. Unified CM distinguishes between inbound and outbound Fast Start based on the direction of the call setup, and the distinction is important because the MTP requirements are not equal. For inbound Fast Start, no MTP is required. Outbound calls on an H.323 trunk do require an MTP when Fast Start is enabled. Frequently, it is only inbound calls that are problematic, and it is possible to use inbound Fast Start to solve the issue without also enabling outbound Fast Start.

DTMF Conversion

An H.323 trunk supports the signaling of DTMF by means of H.245 out-of-band methods. H.323 intercluster trunks also support DTMF by means of NTE. There are no DTMF configuration options for H.323 trunks; Unified CM dynamically chooses the DTMF transport method.

The following scenarios can occur when two endpoints on different clusters are connected with an H.323 trunk:

- When both endpoints are SIP, then NTE is used. No MTP is required for DTMF.
- When one endpoint is SIP and supports both KPML and NTE, but the other endpoint is not SIP, then DTMF is sent as KPML from the SIP endpoint to Unified CM, and H.245 is used on the trunk. No MTP is required for DTMF.
- If one endpoint is SIP and supports only NTE but the other is not SIP, then H.245 is used on the trunk. An available MTP is allocated for the call. The MTP will be allocated on the Unified CM cluster where the SIP endpoint is located.

For example: A Cisco Unified IP Phone 7970 using SIP to communicate with a Cisco Unified IP Phone 7970 running SCCP, will use NTE when connected via a SIP trunk but will use OOB methods when communicating over an H.323 trunk (with the trunk using the H.245 method).

When a call is inbound from one H.323 trunk and is routed to another H.323 trunk, NTE will be used for DTMF when both endpoints are SIP. H.245 will be used if either endpoint is not SIP. An MTP will be allocated if one side is a SIP endpoint that supports only NTE and the other side is non-SIP.

DTMF Relay on H.323 Gateways and Cisco Unified Border Element

H.323 gateways support DTMF relay via H.245 Alphanumeric, H.245 Signal, NTE, and audio in the media stream. The NTE option must not be used because it is not supported on Unified CM for H.323 gateways at this time. The preferred option is H.245 Signal. MTPs are required for establishing calls to an H.323 gateway if the other endpoint does not have signaling capability in common with Unified CM. For example, a Cisco Unified IP Phone 7960 running the SIP stack supports only NTEs, so an MTP is needed with an H.323 gateway.

CTI Route Points

A CTI Route Point uses CTI events to communicate with CTI applications. For DTMF purposes, the CTI Route Point can be considered as an endpoint that supports all OOB methods and does not support RFC 2833. For such endpoints, the only instance where an MTP will be required for DTMF conversion would be when it is communicating with another endpoint that supports only RFC 2833.

CTI Route Points that have first-party control of a phone call will participate in the media stream of the call and require an MTP to be inserted. When the CTI has third-party control of a call so that the media passes through a device that is controlled by the CTI, then the requirement for an MTP is dependent on the capabilities of the controlled device.

Example 17-1 Call Flow that Requires an MTP for NTE Conversion

Assume the example system has CTI route points with first-party control (the CTI port terminates the media), which integrate to a system that uses DTMF to navigate an IVR menu. If all phones in the system are running SCCP, then no MTP is required. In this case Unified CM controls the CTI port and receives DTMF from the IP phones via SCCP. Unified CM provides DTMF conversion.

However, if there are phones running a SIP stack (that support only NTE and not KPML), an MTP is required. NTEs are part of the media stream; therefore Unified CM does not receive them. An MTP is invoked into the media stream and has one call leg that uses SCCP, and the second call leg uses NTEs. The MTP is under SCCP control by Unified CM and performs the NTE-to-SCCP conversion. Note that the newer phones that do support KPML will not need an MTP.

MTP Usage with a Conference Bridge

MTPs are utilized in a conference call when one or more participant devices in the conference use RFC 2833. When the conference feature is invoked, Unified CM allocates MTP resources for every conference participant device in the call that supports only RFC 2833. This is regardless of the DTMF capabilities of the conference bridge used.

MTP Resources

The following types of devices are available for use as an MTP:

Software MTP (Cisco IP Voice Media Streaming Application)

A software MTP is a device that is implemented by enabling the Cisco IP Voice Media Streaming Application on a Unified CM server. When the installed application is configured as an MTP application, it registers with a Unified CM node and informs Unified CM of how many MTP resources it supports. A software MTP device supports only G.711 streams. The IP Voice Media Streaming Application is a resource that may also be used for several functions, and the design guidance must consider all functions together (see Cisco IP Voice Media Streaming Application, page 17-4).

Software MTP (Based on Cisco IOS)

- The capability to provide a software-based MTP on the router is available beginning with Cisco IOS Release 12.3(11)T for the Cisco 3800 Series Routers; Release 15.0(1)M for the Cisco 2900 Series and 3900 Series Routers; Release IOS-XE for ASR1002, 1004, and 1006 Routers; Release IOS-XE 3.2 for ASR1001 Routers; and Release 12.3(8)T4 for other router models.
- This MTP allows configuration of any of the following codecs, but only one may be configured at a given time: G.711 mu-law and a-law, G.729a, G.729, G.729ab, G.729b, and passthrough. Some of these are not pertinent to a Unified CM implementation.
- Router configurations permit up to 1,000 individual streams, which support 500 transcoded sessions. This number of G.711 streams generates 10 Mbytes of traffic. The Cisco ISR G2s and ASR routers can support significantly higher numbers than this.

Hardware MTP (PVDM2, Cisco NM-HDV2 and NM-HD-1V/2V/2VE)

- This hardware uses the PVDM-2 modules for providing DSPs.
- Each DSP can provide 16 G.711 mu-law or a-law, 8 G.729a or G.722, or 6 G.729 or G.729b MTP sessions.

Hardware MTP (Cisco 2900 and 3900 Series Routers with PVDM3)

- These routers use the PVDM3 DSPs natively on the motherboards or PVDM2 with an adaptor on the motherboard or on service modules.
- The capacity of each of the DSP type varies from 16 G.711 a-law or mu-law sessions for the PVDM3-16 to 256 G.711 sessions for the PVDM3-256.

Note You cannot configure G.729 or G.729b codecs when configuring hardware MTP resources in Cisco IOS. However, Unified CM can use hardware transcoding resources as MTPs if all other MTP resources are exhausted or otherwise unavailable.

Trusted Relay Point

A Trusted Relay Point (TRP) is a device that can be inserted into a media stream to act as a control point for that stream. It may be used to provide further processing on that stream or as a method to ensure that the stream follows a specific desired path. There are two components to the TRP functionality, the logic utilized by Unified CM to invoke the TRP and the actual device that is invoked as the anchor point of the call. The TRP functionality can invoke an MTP device to act as that anchor point.

Unified CM provides a new configuration parameter for individual phone devices, which invokes a TRP for any call to or from that phone. The system utilizes the media resource pool mechanisms to manage the TRP resources. The media resource pool of that device must have an available device that will be invoked as a TRP.

See the chapter on Network Infrastructure, page 3-1, for an example of a use case for the TRP as a QoS enforcement mechanism, and see the chapter on Unified Communications Security, page 4-1, for an example of utilizing the TRP as an anchor point for media streams in a redundant data center with firewall redundancy.

Annunciator

An annunciator is a software function of the Cisco IP Voice Media Streaming Application that provides the ability to stream spoken messages or various call progress tones from the system to a user. It uses SCCP messages to establish RTP streams, and it can send multiple one-way RTP streams to devices such as Cisco IP phones or gateways. The device must be capable of SCCP to utilize this feature. SIP phones and devices are still able to receive all the various messages provided by the annunciator. For SIP devices, all these messages and tones are downloaded (pushed) to the device at registration so that they can be invoked as needed by SIP signaling messages from Unified CM.

Tones and announcements are predefined by the system. The announcements support localization and may also be customized by replacing the appropriate .wav file. The annunciator is capable of supporting G.711 a-law and mu-law, G.729, and Wideband codecs without any transcoding resources.

The following features require an annunciator resource:

- Cisco Multilevel Precedence Preemption (MLPP)

  This feature has streaming messages that it plays in response to the following call failure conditions.

  - Unable to preempt due to an existing higher-precedence call.
  - A precedence access limitation was reached.
  - The attempted precedence level was unauthorized.
  - The called number is not equipped for preemption or call waiting.

- Integration via SIP trunk

  SIP endpoints have the ability to generate and send tones in-band in the RTP stream. Because SCCP devices do not have this ability, an annunciator is used in conjunction with an MTP to generate or accept DTMF tones when integrating with a SIP endpoint. The following types of tones are supported:

  - Call progress tones (busy, alerting, and ringback)
  - DTMF tones

- Cisco IOS gateways and intercluster trunks

  These devices require support for call progress tone (ringback tone).

- System messages

  During the following call failure conditions, the system plays a streaming message to the end user:

  - A dialed number that the system cannot recognize
  - A call that is not routed due to a service disruption
  - A number that is busy and not configured for preemption or call waiting

- Conferencing

  During a conference call, the system plays a barge-in tone to announce that a participant has joined or left the bridge.

An annunciator is automatically created in the system when the Cisco IP Voice Media Streaming Application is activated on a server. If the Media Streaming Application is deactivated, then the annunciator is also deleted. A single annunciator instance can service the entire Unified CM cluster if it meets the performance requirements (see Annunciator Performance, page 17-21); otherwise, you must configure additional annunciators for the cluster. Additional annunciators can be added by activating the Cisco IP Voice Media Streaming Application on other servers within the cluster.

The annunciator registers with a single Unified CM at a time, as defined by its device pool. It will automatically fail over to a secondary Unified CM if a secondary is configured for the device pool. Any announcement that is playing at the time of an outage will not be maintained.

An annunciator is considered a media device, and it can be included in media resource groups (MRGs) to control which annunciator is selected for use by phones and gateways.

Annunciator Performance

By default, the annunciator is configured to support 48 simultaneous streams, which is the maximum recommended for an annunciator running on the same server (co-resident) with the Unified CM service. If the server has only 10 Mbps connectivity, lower the setting to 24 simultaneous streams.

A standalone server without the Cisco CallManager Service can support up to 255 simultaneous announcement streams, and a high-performance server with dual CPUs and a high-performance disk system can support up to 400 streams. You can add multiple standalone servers to support the required number of streams.

Cisco RSVP Agent

In order to provide topology-aware call admission control, Unified CM invokes one or two RSVP Agents during the call setup to perform an RSVP reservation across the IP WAN. These agents are MTP or transcoder resources that have been configured to provide RSVP functionality. RSVP resources are treated the same way as regular MTPs or transcoders from the perspective of allocation of an MTP or transcoder resource by Unified CM.

The Cisco RSVP Agent feature was first introduced in Cisco IOS Release 12.4(6)T. For details on RSVP and Cisco RSVP Agents, refer to the chapter on Call Admission Control, page 11-1.

Music on Hold

The Music on Hold (MoH) feature requires that each MoH server must be part of a Unified CM cluster and participate in the data replication schema. Specifically, the MoH server must share the following information with the Unified CM cluster through the database replication process:

- 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

To configure a MoH server, enable the Cisco IP Voice Media Streaming Application Service on one or more Unified CM nodes. An MoH server can be deployed along with Unified CM on the same server or in standalone mode.

Unicast and Multicast MoH

Unified CM supports unicast and multicast MoH transport mechanisms.

A unicast MoH stream is a point-to-point, one-way audio Real-Time Transport Protocol (RTP) stream from the MoH server to the endpoint requesting MoH. It uses a separate source stream for each user or connection. Thus, if twenty devices are on hold, then twenty streams are generated over the network between the server and these endpoint devices. 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. However, these additional MoH streams can potentially have a negative effect on network throughput and bandwidth.

A multicast MoH stream is a point-to-multipoint, one-way audio RTP stream between the MoH server and the multicast group IP address. The endpoints requesting an MoH audio stream can join the multicast group as needed. This mode of MoH conserves system resources and bandwidth because it enables multiple users to use the same audio source stream to provide music on hold. For this reason, multicast is an extremely attractive transport mechanism 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.

There are distinct differences between unicast and multicast MoH in terms of call flow behavior. A unicast MoH call flow is initiated by a message from Unified CM 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 Unified CM 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 MoH Call Flows, page 17-27.

Supported Unicast and Multicast Gateways

The following gateways support both unicast and multicast MoH:

- Cisco 2900 Series and Cisco 3900/3900E Series ISR G2 Routers with Cisco IOS 15.0.1M or later release
- Cisco 2800 Series and 3800 Series Routers with MGCP or H.323 and Cisco IOS 12.3.14T or later release
- Cisco 2800 Series and 3800 Series Routers with SIP and Cisco IOS 12.4(24)T or later release
- Cisco VG224 Analog Voice Gateways with MGCP and Cisco IOS 12.3.14T or later release
- Cisco VG204 and VG202 Analog Voice Gateways with MGCP or SCCP and Cisco IOS 12.4(22)T or later release
- Cisco VG248 Analog Phone Gateways
- Cisco ASR 1000 Series Aggregation Services Routers

Note Cisco 2800 Series, 3800 Series, and VG248 gateways are End of Sale (EoS). There are other legacy gateways that also support unicast and multicast MoH.

Note The Cisco Unified Border Element on Cisco ASR 1000 Series Aggregation Services Routers might not support one-way streaming of music or announcements by the Cisco Unified Communications Manager Music on Hold (MoH) feature. For more information, refer to the release notes for your version of Cisco Unified Communications Manager, available at http://www.cisco.com/en/US/products/sw/voicesw/ps556/prod_release_notes_list.html.

MoH Selection Process

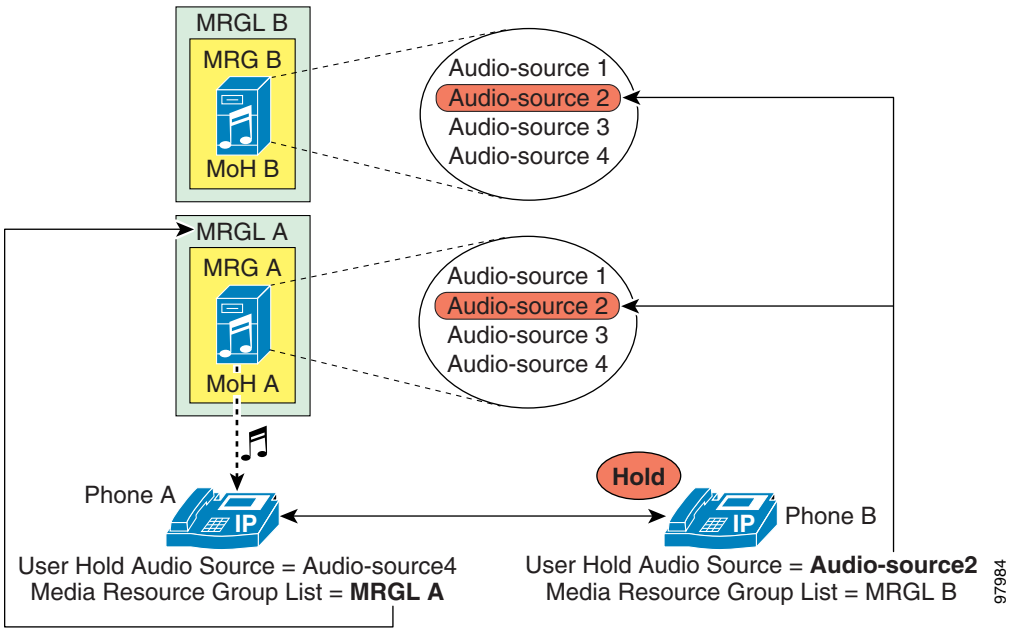
This section describes the MoH selection process as implemented in Unified CM.

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.

As illustrated by the example in Figure 17-3, 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 17-3 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 containing a MoH server configured for multicast.

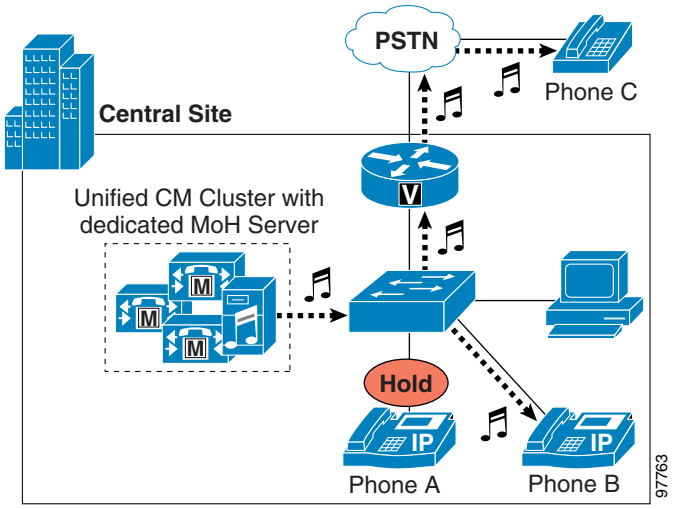
User and Network Hold

User hold includes the following types:

- 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 17-4 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 17-4 Basic User Hold Example



Network hold can occur in following scenarios:

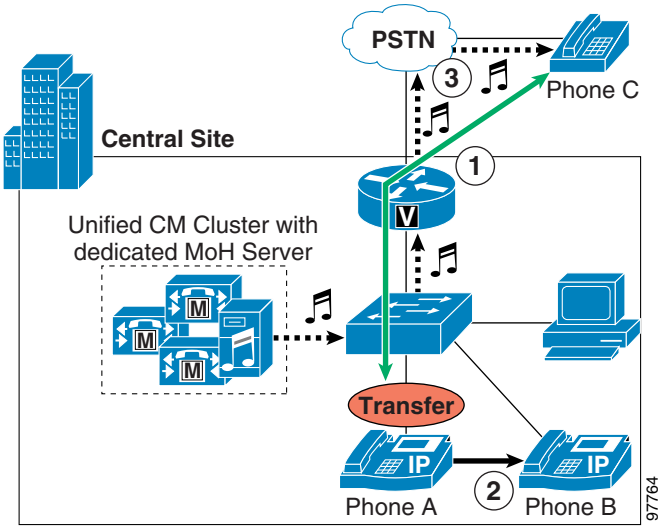
- Call transfer
- Call Park
- Conference setup
- Application-based hold

Figure 17-5 illustrates an example of network hold during a call transfer. The call flow involves the following steps:

1. Phone A receives a call from PSTN phone C.
2. Phone A answers the call and then transfers it to phone B. During the transfer process, phone C is put on network hold.
3. Phone C receives an MoH stream from the MoH server via the gateway. After phone A completes the transfer action, phone C disconnects from the music stream and gets redirected to phone B.

This process is the same for other network hold operations such as call park and conference setup.

Figure 17-5 Basic Network Hold Example for Call Transfer



MoH Sources

A Unified CM MoH server can generate a MoH stream from two types of sources, audio file and fixed source, either of which can be transmitted as unicast or multicast. You can configure a maximum of 51 MoH audio sources per Unified CM cluster, of which up to 50 can be audio files but only one can be a fixed source.

Audio File

Audio files (.wav format) can be uploaded to Unified CM, which then automatically generates MoH audio files for the specified codecs. Unified CM supports G711 (a-law and mu-law), G.729 Annex A, and Wideband codecs for MoH streams.

Note Before configuring a MoH audio source, you must upload the .wav formatted audio source file to every MoH server within the cluster using the upload file function in the Unified CM Administration interface. Cisco recommends that you first upload the audio source file onto each MoH server in the cluster, then upload it onto the publisher (even if not an MoH server), and finally assign an MoH Audio Stream Number and configure the MoH audio source in the Unified CM Administration interface on the publisher.

Fixed Source

If recorded or live audio is needed, MoH can be generated from a fixed source connected to the audio input of the local sound card. The Cisco MoH USB audio sound card (MOH-USB-AUDIO=) must be used for connecting a fixed or live audio source to the MoH server. This USB sound card is compatible with all Cisco MCS platforms that support Cisco Unified CM.

This mechanism enables you to use radios, CD players, or any other compatible sound source to stream MoH. The stream from the fixed audio source is transcoded in real-time to support the codec that was configured through Unified CM 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.

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 Selection

To determine which User and Network Audio Source configuration setting to apply in a particular case, Unified CM 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 Device Configuration setting
4. Cluster-wide default setting

Unified CM 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

Note that system default MoH resources are resources that are not assigned to any MRG and they are always unicast.

MoH Call Flows

The following sections provide detailed illustrations and explanations of unicast and multicast MoH call flows for both SCCP and SIP endpoints.

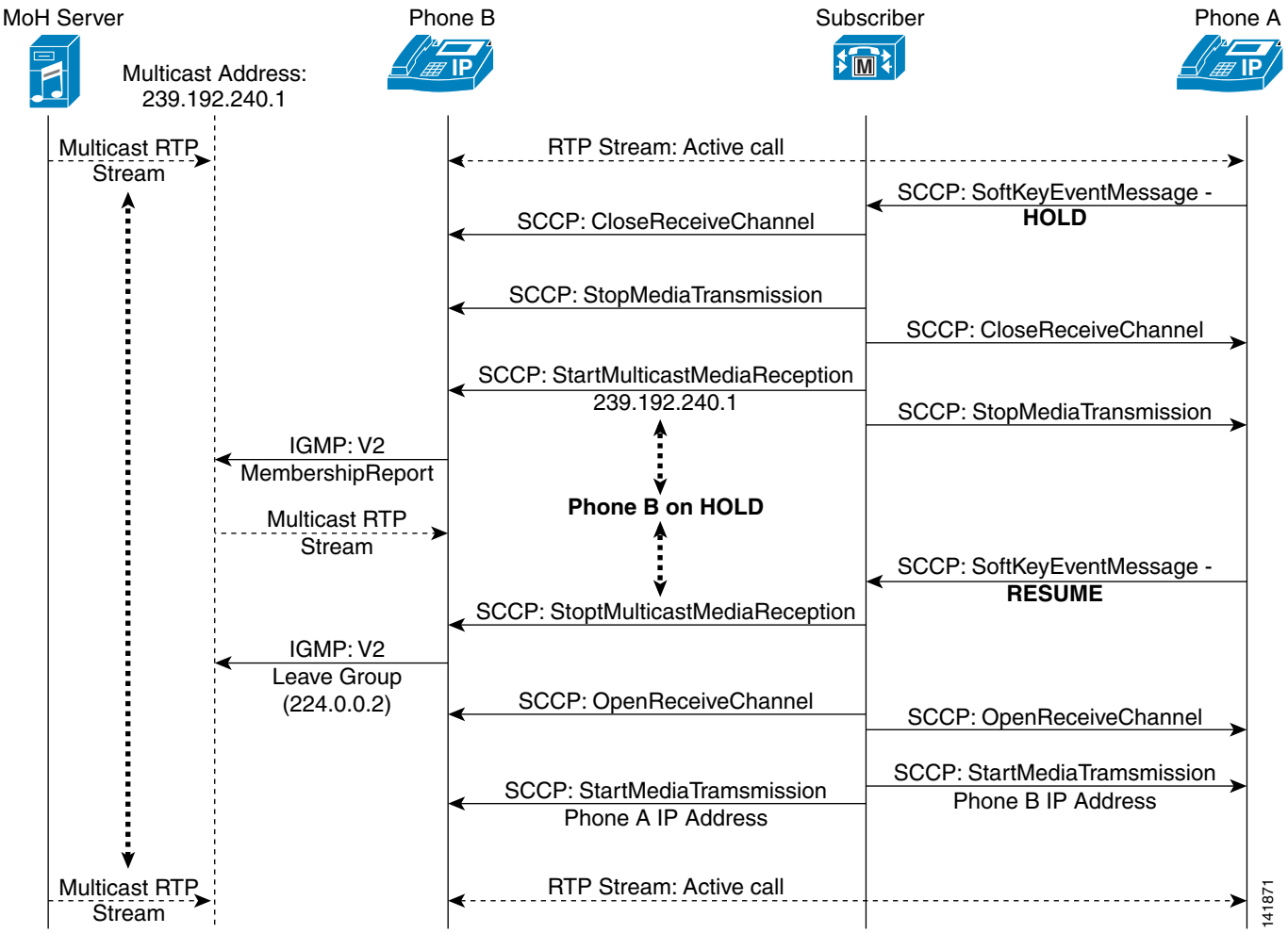
SCCP Call Flows

This section describes the multicast and unicast call flows for music on hold with Skinny Client Control Protocol (SCCP) endpoints.

SCCP Multicast Call Flow

Figure 17-6 illustrates a typical SCCP multicast call flow. As shown in the diagram, when the Hold softkey is pressed at phone A, Unified CM 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, Unified CM 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 17-6 Detailed SCCP Multicast MoH Call Flow



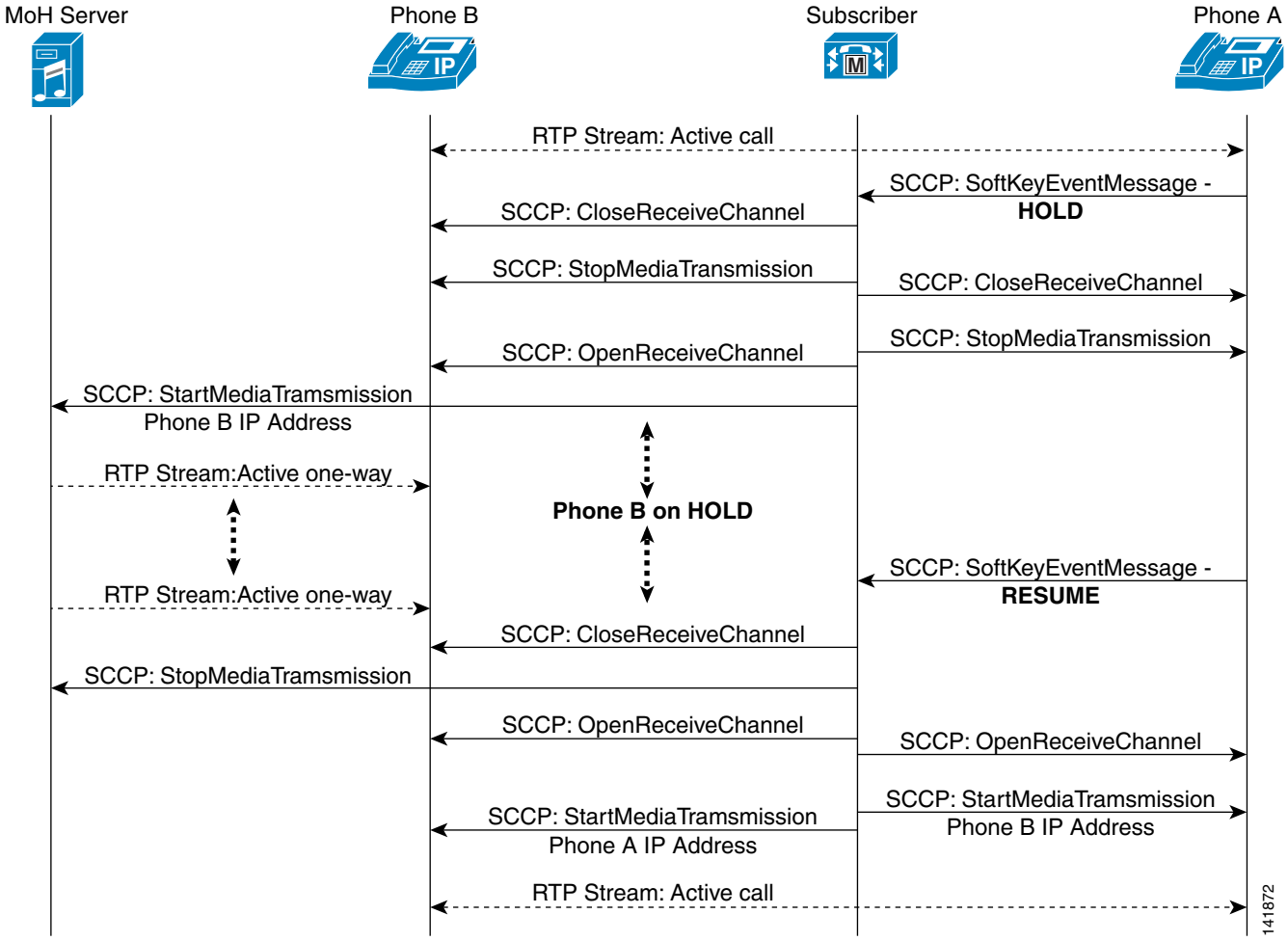
Meanwhile, the MoH server has been sourcing RTP audio to this multicast group address and, upon joining the multicast group, phone B begins receiving the MoH stream. Once phone A presses the Resume softkey, Unified CM instructs phone B to Stop Multicast Media Reception. Phone B then sends an IGMP V2 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, Unified CM 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, Unified CM instructs both phones to Start Media Transmission to each other's IP addresses. The phones are once again connected by means of an RTP two-way audio stream.

Note The call flow diagrams in Figure 17-6 and Figure 17-7 assume that an initial call exists 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.

SCCP Unicast Call Flow

Figure 17-7 depicts an SCCP unicast MoH call flow. In this call flow diagram, when the Hold softkey is pressed at phone A, Unified CM 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 way.

Figure 17-7 Detailed SCCP Unicast MoH Call Flow



Next, Unified CM tells phone B (the holdee) to Open Receive Channel. (This is quite different from the multicast case, where Unified CM tells the holdee to Start Multicast Media Reception.) Then Unified CM 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, Unified CM 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, Unified CM 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 by means of an RTP two-way audio stream.

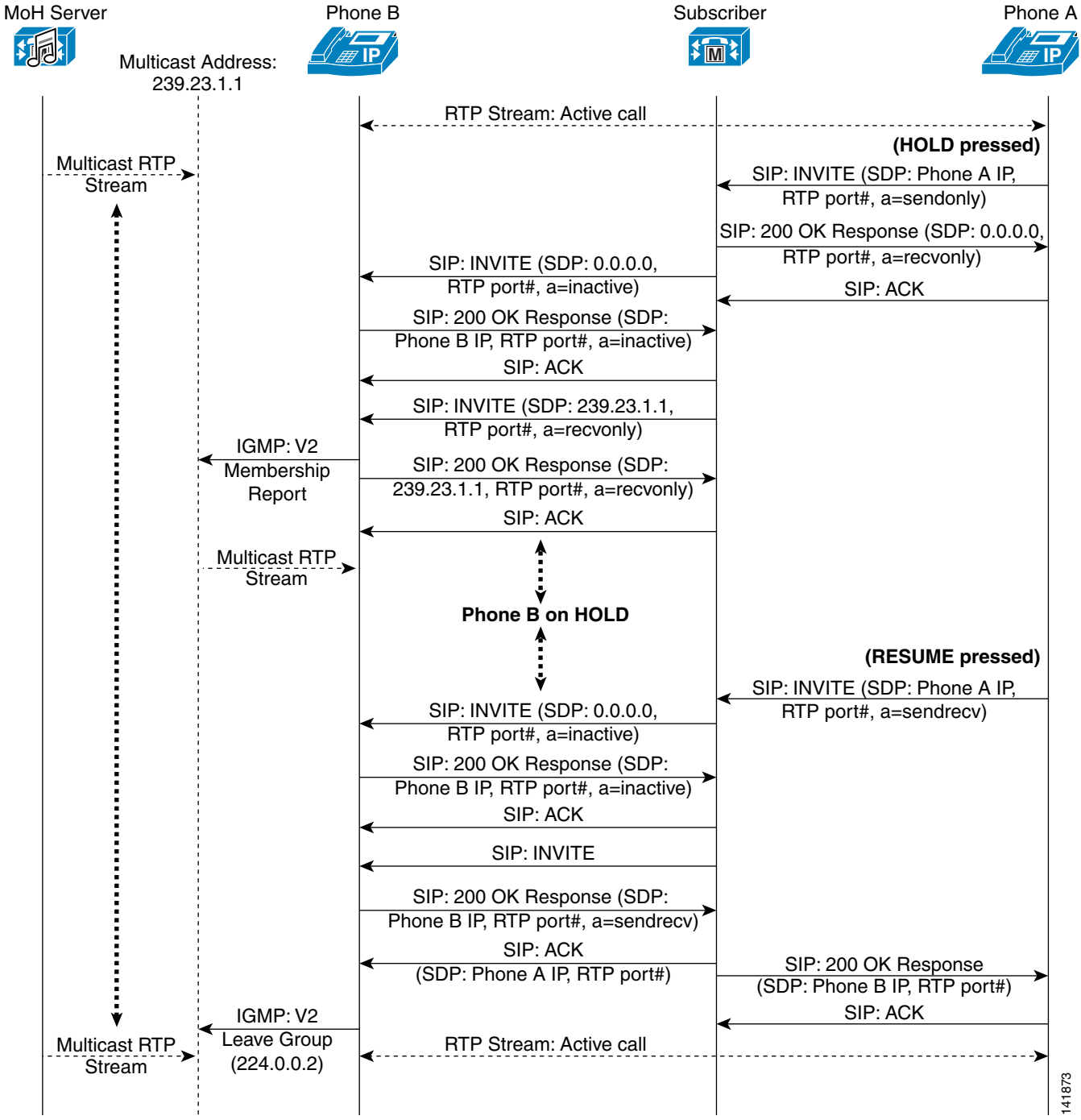
SIP Call Flows

This section describes the multicast and unicast call flows for music on hold with Session Initiation Protocol (SIP) endpoints.

SIP Multicast Call Flow

Figure 17-8 illustrates a typical SIP multicast call flow. As shown in the diagram, when the Hold softkey is pressed at phone A, phone A sends a SIP INVITE with a Session Description Protocol (SDP) connection information indication of phone A's IP address and a media attribute indication of sendonly. Unified CM then instructs phone A to disconnect the RTP stream by means of a SIP 200 OK Response with an SDP connection information indication of 0.0.0.0 and a media attribute indication of recvonly. Phone B is then told to disconnect the RTP stream by means of a SIP INVITE from Unified CM with an SDP connection information indication of 0.0.0.0 and a media attribute of inactive. After a SIP 200 OK Response is sent back from phone B to Unified CM indicating an SDP media attribute of inactive, Unified CM then sends a SIP INVITE to phone B with an SDP connection information indication of the MoH multicast group address (in this case 239.23.1.1) and a media attribute of recvonly.

Figure 17-8 Detailed SIP Multicast MoH Call Flow



Next, phone B in Figure 17-8 issues an IGMP V2 Membership Report message indicating that it is joining this multicast group. In addition, phone B sends a SIP 200 OK Response back to Unified CM indicating an SDP media attribute of recvonly in response to the previous SIP INVITE. Meanwhile, the MoH server has been sourcing RTP audio to this MoH multicast group address and, upon joining the multicast group, phone B begins receiving the one-way MoH stream.

When the user at phone A presses the Resume softkey, phone A sends a SIP INVITE with an SDP connection information indication of phone A's IP address and media attribute indications of phone A's receiving RTP port and sendrecv. Unified CM then instructs phone B to disconnect from the multicast MoH stream by means of a SIP INVITE with an SDP connection information indication of 0.0.0.0 and a media attribute indication of inactive. A SIP 200 OK Response is sent back from phone B to Unified CM, indicating an SDP media attribute of inactive.

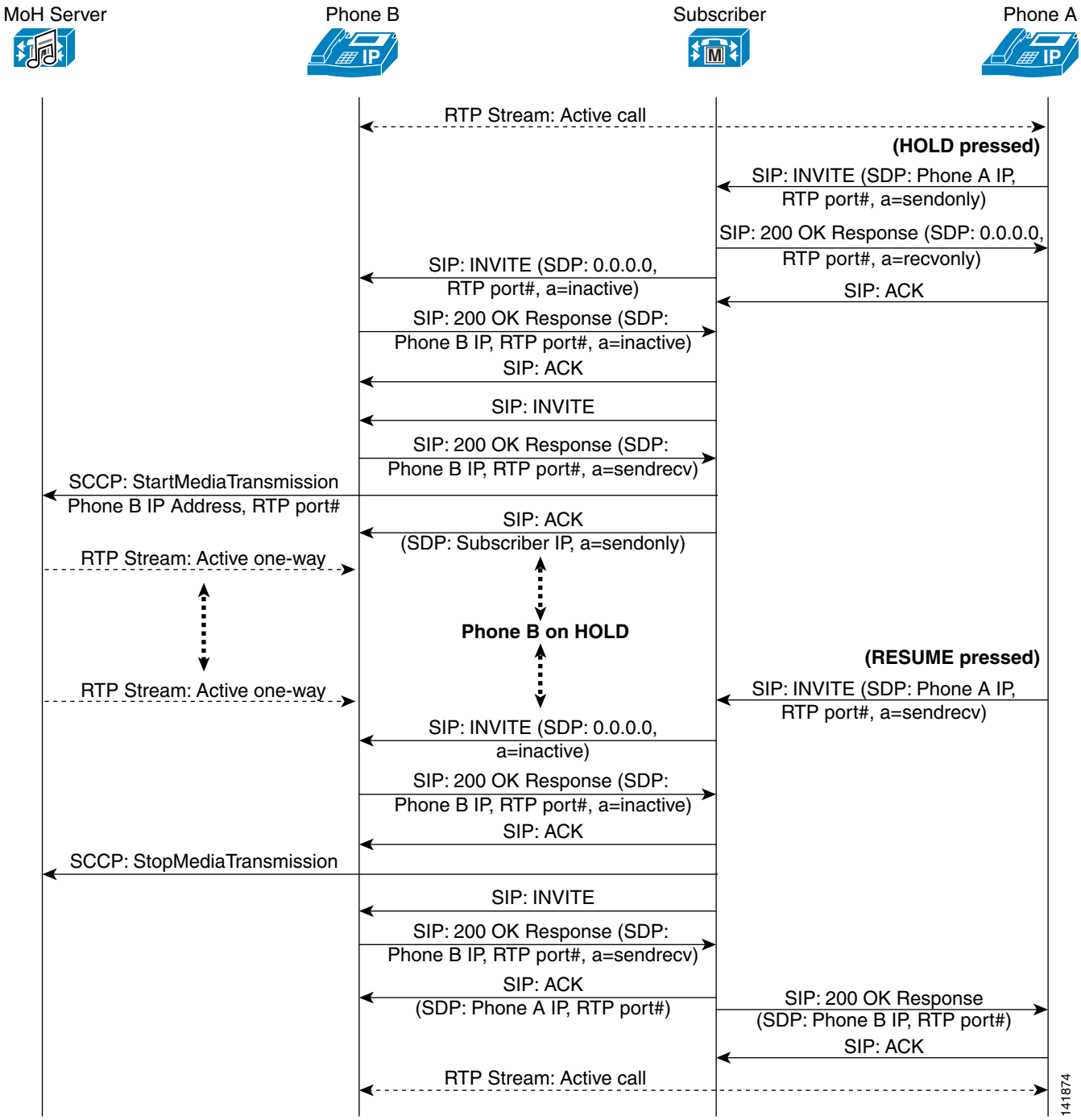
Next Unified CM sends a SIP INVITE to phone B, and phone B responds with a SIP 200 OK Response with an SDP connection information indication of phone B's IP address and media attribute indications of phone B's receiving RTP port and sendrecv. Unified CM responds by sending a SIP ACK to phone B with an SDP connection information indication of phone A's IP address and a media attribute of phone A's receiving RTP port number. Likewise, Unified CM forwards a SIP 200 OK Response to phone A's original resuming SIP INVITE, with an SDP connection information indication of phone B's IP address and a media attribute of phone B's receiving RTP port number. 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. Finally, the RTP two-way audio stream between phones A and B is reestablished.

Note The call flow diagrams in Figure 17-8 and Figure 17-9 assume that an initial call exists 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, some acknowledgements, progression indications, 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.

SIP Unicast Call Flow

Figure 17-9 depicts a SIP unicast MoH call flow. As shown in the diagram, when the Hold softkey is pressed at phone A, phone A sends a SIP INVITE with an SDP connection information indication of phone A's IP address and a media attribute indication of sendonly. Unified CM then instructs phone A to disconnect the RTP stream by means of a SIP 200 OK Response with an SDP connection information indication of 0.0.0.0 and a media attribute indication of recvonly. Phone B is then told to disconnect the RTP stream by means of a SIP INVITE from Unified CM, with an SDP connection information indication of 0.0.0.0 and a media attribute of inactive. Next a SIP 200 OK Response is sent back from phone B to Unified CM, indicating an SDP media attribute of inactive. Up to this point, unicast and multicast MoH call flows are exactly the same.

Figure 17-9 Detailed SIP Unicast MoH Call Flow



Unified CM then sends a SIP INVITE to phone B, and phone B responds back with a SIP 200 OK Response indicating SDP connection information with phone B's IP address and media attribute indications of phone B's receiving RTP port number and sendrecv. Unified CM then sends a SCCP StartMediaTransmission message to the MoH server, with phone B's address and receiving RTP port

number. This is followed by a SIP ACK from Unified CM to phone B indicating SDP connection information of the Unified CM IP address and a media attribute of sendonly. Meanwhile, the MoH server begins sourcing RTP audio to phone B, and phone B begins receiving the one-way MoH stream.

When the user at phone A presses the Resume softkey, phone A sends a SIP INVITE with an SDP connection information indication of phone A's IP address and media attribute indications of phone A's receiving RTP port and sendrecv. Unified CM then instructs phone B to disconnect from the multicast MoH stream by means of a SIP INVITE with an SDP connection information indication of 0.0.0.0 and a media attribute indication of inactive. A SIP 200 OK Response is sent back from phone B to Unified CM, indicating an SDP media attribute of inactive. Then Unified CM sends an SCCP StopMediaTransmission message to the MoH server, causing the MoH server to stop forwarding the MoH stream to phone B.

Next Unified CM sends a SIP INVITE to phone B, and phone B responds with a SIP 200 OK Response with an SDP connection information indication of phone B's IP address and media attribute indications of phone B's receiving RTP port and sendrecv. Unified CM responds by sending a SIP ACK to phone B, with an SDP connection information indication of phone A's IP address and a media attribute of phone A's receiving RTP port number. Likewise, Unified CM forwards a SIP 200 OK Response to phone A's original resuming SIP INVITE with an SDP connection information indication of phone B's IP address and a media attribute of phone B's receiving RTP port. Finally, the RTP two-way audio stream between phones A and B is reestablished.

Capacity Planning for Media Resources

This section provides information on the capacities of various network modules and chassis that carry DSPs, the capacities of the chassis to carry network modules, and software dependencies of the hardware.

For all Cisco ISR G1 and G2 capacity planning, use the DSP Calculator available at http://www.cisco.com/go/dspcalculator. For other platforms (such as the Cisco 1700, 2600, and 3700 Series Routers), use the legacy DSP calculator at http://www.cisco.com/cgi-bin/Support/DSP/cisco_dsp_calc.pl.

The DSP resources for Unified Communications solutions are provided by NM-HD, NM-HDV, and PVDM modules. NM-HD and NM-HDV2 modules are supported on Cisco ISR G1 and G2 Series platforms. Refer to the respective product data sheets for capacity information for these modules.

PVDM modules are available in three models: PVDM-256K, PVDM2, and the newer PVDM3. Each of the models has several modules with different density support. For example, a PVDM-256K-4 and a PVDM2-16 are single DSP modules in their respective category. PVDM2 modules are supported on Cisco ISR G1 and ISR G2 platforms (minimum of Cisco IOS Release 15.0(1)M is required for the Cisco ISR G2 Series). The PVDM3 DSP modules are supported on the Cisco 2900 Series and 3900 Series platforms and require a minimum Cisco IOS Release of 15.0(1) M. PVDM3 modules provide DSP resources for both voice and video. The PVDM3 modules are newer than the PVDM2 and PVDM-256K modules, and the three types are not interchangeable.

Some things to consider when doing capacity planning for hardware-based media resources include the density of the module, the underlying platform (Cisco ISR G1 or G2), and the minimum Cisco IOS version required.

For capacity information on PVDM2 modules, refer to the *High-Density Packet Voice Digital Signal Processor Module for Cisco Unified Communications Solutions* data sheet, available at

http://www.cisco.com/en/US/prod/collateral/routers/ps5854/product_data_sheet0900aecd8016e845_ps3115_Products_Data_Sheet.html

For capacity information on PVDM3 modules, refer to the *High-Density Packet Voice Video Digital Signal Processor Module for Cisco Unified Communications Solutions* data sheet, available at

http://www.cisco.com/en/US/prod/collateral/modules/ps3115/data_sheet_c78-553971.html

Considerations for Cisco 2900 and 3900 Series Platforms

The following guidelines and considerations apply to the DSP resources hosted by these platforms:

- The Cisco 2900 and 3900 Series Routers support only the PVDM3 DSPs in the on-board (motherboard) DSP slots. PVDM2 DSPs may be used in those slots by using an adaptor card. NM-HD and NM-HDV2 cards can be used in Service Module slots with an adaptor card.
- PVDM2 and PVDM3 modules cannot be used at the same time on the same motherboard.
- DSP sharing can be done only between the same DSP types. For example, if the motherboard is populated with PVDM3 DSPs and the Service Modules are populated with PVDM2 DSPs, then the DSPs in the Service Modules may be shared with each other but DSPs on the motherboard may not be shared with those in the Service Modules.
- PVDM3 DSPs support all the functions that the PVDM2 DSPs support except for Cisco Fax Relay.

Unlike the PVDM2, the PVDM3 DSPs have a single software image for all media functions.

Cisco 2800 and 3800 Series Platforms

The following guidelines and considerations apply to the DSP resources hosted by these platforms:

- Although the Cisco 2800 and 3800 Series Routers all have two AIM slots, they do not support the AIM-VOICE-30 or AIM-ATM-VOICE-30 cards because PVDM2 modules that are installed on the motherboard provide that functionality.

You can install the NM-HDV2, NM-HD-*xx*, and NM-HDV modules in the Cisco IOS platforms as indicated in the product data sheets.

All three families of modules may be installed in a single chassis. However, the conferencing and transcoding features cannot be used simultaneously on both the NM-HDV family and either of the other two families (NM-HD-*xx* or NM-HDV2). In addition, the NM-HDV (TI-549), NM-HD-*xx*, and NM-HDV2 (TI-5510) cannot be used simultaneously for conferencing and transcoding within a single chassis.

You can mix NM-HDV and NM-HDV-FARM modules in the same chassis, but not all chassis can be completely populated by these modules.

Capacity Planning for Music On Hold

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 when doing capacity planning for MoH resources. The capacity of the MoH server depends on several factors such as deployment model (co-resident or standalone), underlying server platform, and so forth.

Co-resident and Standalone MoH

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

- Co-resident deployment

  The term *co-resident* refers to two or more services or applications running on the same server. In a co-resident deployment, the MoH feature runs on any server (either publisher or subscriber) in the cluster that is also running the Unified CM software.

- Standalone deployment

  A standalone deployment, places the MoH feature on a dedicated media resource server node within the Unified CM cluster. This server acts as neither a publisher or a subscriber. That is, the Cisco IP Voice Media Streaming Application service is the only service enabled on the server. The only function of this dedicated server is to send MoH streams to devices within the network

Server Platform Limits

Table 17-2 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 17-2 Maximum Number of MoH Sessions per Server Platform or OVA Template

Server Platform	Codecs Supported	MoH Sessions Supported
MCS 7816 MCS 7825 (or OVA equivalent) MCS 7828	G.711 (A-law and mu-law) G.729a Wideband audio	Co-resident or standalone server: 250 MoH sessions[1]
MCS 7835 (or OVA equivalent) MCS 7845 (or OVA equivalent)	G.711 (A-law and mu-law) G.729a Wideband audio	Co-resident or standalone server: 500 MoH sessions

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

The following two 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 (or OVA equivalent)	Multicast MoH audio sources	MoH codecs enabled (G.711 mu-law and G.729)	**Maximum half-duplex streams**
500	- (12	* 2)	= **476**

  Therefore, in this example, the Maximum Half Duplex Streams parameter would be configured with a value of no more than 476.

  The value of this parameter should never be set higher than the capacities indicated in Table 17-2, 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.

  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, a large number of held devices can join each multicast stream. This parameter should be set to a number that is greater than or equal to the number of devices that might be placed on multicast MoH at any given time. Typically multicast traffic is accounted for based on the number of streams being generated; however, Unified CM maintains a count of the actual number of devices placed on multicast MoH or joined to each multicast MoH stream. Although this method is different than the way multicast traffic is normally tracked, it is important to configure this parameter appropriately.

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

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. For details on how to configure the service parameters, refer to the *Cisco Unified Communications Manager Administration Guide*, available at

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

Note The maximum session limits listed in Table 17-2 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

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 requires 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 co-resident (or standalone) MoH servers would be required to handle this load, as shown by the following calculation:

[(500 sessions per MCS 7835 or 7845 co-resident server) * (1 co-resident server)] + [(250 sessions per MCS 7816, 7825, or 7878 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 7816, 7825, or 7878), as shown by the following calculation:

(250 sessions per MCS 7816, 7825, or 7878 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.

High Availability for Media Resources

The Unified CM constructs of media resource groups (MRGs) and media resource group lists (MRGLs) are used to control how the resources described in this chapter are organized and accessed. This section discusses considerations for how to utilize these constructs effectively.

Media Resource Groups and Lists

Media resource groups (MRGs) and media resource lists (MRGLs) provide a method to control how resources are allocated that could include rights to resources, location of resources, or resource type for specific applications. This section assumes you have an understanding of media resource groups and lists, and it highlights the following design considerations:

- The system defines a default media resource group that is not visible in the user interface. All resources are members of this default MRG when they are created. When using MRGs to control access to resources, it is necessary to move the resources out of the default MRG by explicitly configuring them in some other MRG. If the desired effect is for resources to be available only as a last resort for all calls, then the resources may remain in the default group. Also, if no control over resources is necessary, they may remain in the default group.
- Consumers of media resources use resources first from any media resource group (MRG) or media resource group list (MRGL) that their configuration specifies. If the required resource is not available, the default MRG is searched for the resource. For simple deployments, the default MRG alone may be used.
- Use media resource groups (MRGs) and media resource group lists (MRGLs) to provide sharing of resources across multiple Unified CMs. If you do not use MRGs and MRGLs, the resources are available to a single Unified CM only.
- MRGLs will use MRGs in the order that they are listed in the configuration. If one MRG does not have the needed resource, the next MRG is searched. If all MRGs are searched and no resource is found, the search terminates.
- Within an MRG, resources are allocated based on their order in their configuration even though Unified CM Administration displays the devices in an MRG in alphabetical order. If you want media resources to be allocated in a specific order, Cisco recommends that you create a separate MRG for each individual resource and use MRGLs to specify the order of allocation.
- When there are multiple devices providing the same type of resource within an MRG, the algorithm for allocating that resource load-balances across all those devices. The load balancing depends on the capacity of each device providing similar resources, so it frequently might not be a round-robin behavior. For example, if an MRG has more than one device providing MTP resources, the system will load-balance across each device for MTP requirements. When a resource has been used, a pointer for that MRG is incremented to the next device. A device may be present in more than one MRG, which will affect the pointers of all groups of which the device is a member.
- An MRG may contain multiple types of resources, and the appropriate resource will be allocated from the group based on the feature needed. MTPs and transcoders are a special case because a transcoder may also be used as an MTP. For example, when both MTPs and transcoders exist in the same MRG and an MTP is required, the allocation is done based on the order in which the resources appear in the MRG. If transcoder devices appear earlier than MTPs in the MRG, transcoder resources will be allocated for the MTP requirement until the transcoder resources are exhausted and then the system will start allocating MTPs. For this reason, it is important to consider the order of resources when creating MRGs and MRGLs.

- MRGs can also be used to group resources of similar types. As explained in the example above, because a transcoder is a more expensive resource, Cisco recommend grouping transcoders and MTPs into separate MRGs and invoking the right resource by assigning MRGs appropriately. Another example involves conference bridges. Conference bridge resources vary in the number of participants they support, and different MRGs could be used to group the conference resources by conference bridge size.
- You can also use MRGs and MRGLs to separate resources based on geographical location, thereby conserving WAN bandwidth whenever possible.
- Ensure that the media resources themselves have configurations that prevent further invocation of other media resources. For example, if an MTP is inserted into a call and the codec configured on that MTP does not match the one needed by Unified CM for the call, then a transcoder may also be invoked. A frequent mistake is to configure an MTP for G.729 or G.729b when Unified CM needs G.729a.

Redundancy and Failover Considerations for Cisco IOS-Based Media Resources

A high availability design with media resources must include redundant media resources. When these resources are Cisco IOS-based, they can be distributed on more than one Cisco IOS platform to guard against failure of a single platform and they can be registered to different primary Unified CM servers.

Cisco IOS supports two modes of failover capability: graceful and immediate. The default failover method is graceful, in which the resources register to a backup Unified CM server only after all media activity has ceased. The immediate method, on the other hand, makes the resources register to the backup Unified CM server as soon as failure of the primary is detected. In situations where there is only one set of media resources with no redundancy, Cisco recommends use of the immediate failover method.

High Availability for Music On Hold

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.

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.

Design Considerations for Media Resources

This section discusses specific considerations for deploying media resources for use with the various Unified CM deployment models. It also highlights the configuration considerations and best practices to help you design a robust solution for media resource allocation in your Unified CM implementation.

Deployment Models

This section examines where and when the MTP and transcoding resources are used within the following three enterprise IP Telephony deployment models:

- Single-Site Deployments, page 17-41
- Multisite Deployments with Centralized Call Processing, page 17-41
- Multisite Deployments with Distributed Call Processing, page 17-42

Single-Site Deployments

In a single-site deployment, there is no need for transcoding because there are no low-speed links to justify the use of a low bit-rate (LBR) codec. Some MTP resources might be required in the presence of a significant number of devices that are not compliant with H.323v2, such as older versions of Microsoft NetMeeting or certain video devices. MTP resources may be required for DTMF conversion if SIP endpoints are present (see Named Telephony Events (RFC 2833), page 17-13.)

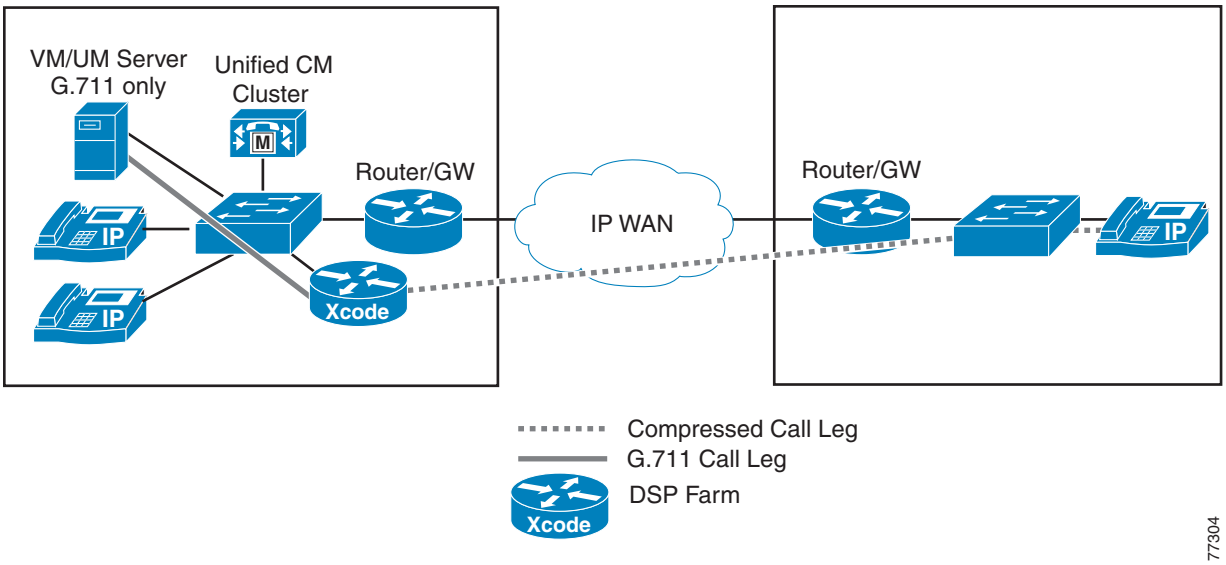
Multisite Deployments with Centralized Call Processing

In a centralized call processing deployment, the Unified CM cluster and the applications (such as voice mail and IVR) are located at the central site, while several remote sites are connected through an IP WAN. The remote sites rely on the centralized Unified CMs to handle their call processing.

Because WAN bandwidth is typically limited, calls are configured to use a low bit-rate codec such as G.729 when traversing the WAN. (See Figure 17-10.)

Voice compression between IP phones is easily configured through the use of *regions* and *locations* in Unified CM. A region defines the type of compression (for example, G.711 or G.729) used by the devices in that region, and a location specifies the total amount of bandwidth available for calls to and from devices at that location.

Figure 17-10 Transcoding for the WAN with Centralized Call Processing



Unified CM uses media resource groups (MRGs) to enable sharing of MTP and transcoding resources among the Unified CM servers within a cluster. In addition, when using an LBR codec (for example, G.729a) for calls that traverse different regions, the transcoding resources are used only if one (or both) of the endpoints is unable to use the LBR codec.

In Figure 17-10, Unified CM knows that a transcoder is required and allocates one based on the MRGL and/or MRG of the device that is using the higher-bandwidth codec. In this case it is the VM/UM server that determines which transcoder device is used. This behavior of Unified CM is based on the assumption that the transcoder resources are actually located close to the higher-bandwidth device. If this system was designed so that the transcoder for the VM/UM server was located at the remote site, then G.711 would be sent across the WAN, which would defeat the intended design. As a result, if there are multiple sites with G.711-only devices, then each of these sites would need transcoder resources when an LBR is run on the WAN.

The placement of other resources is also important. For example, if a conference occurs with three phones at a remote site and the conference resource is located in the central (call processing) site, then three media streams are carried over the WAN. If the conference resource were local, then the calls would not traverse the WAN. It is necessary to consider this factor when designing the bandwidth and call admission control for your WAN.

Multisite Deployments with Distributed Call Processing

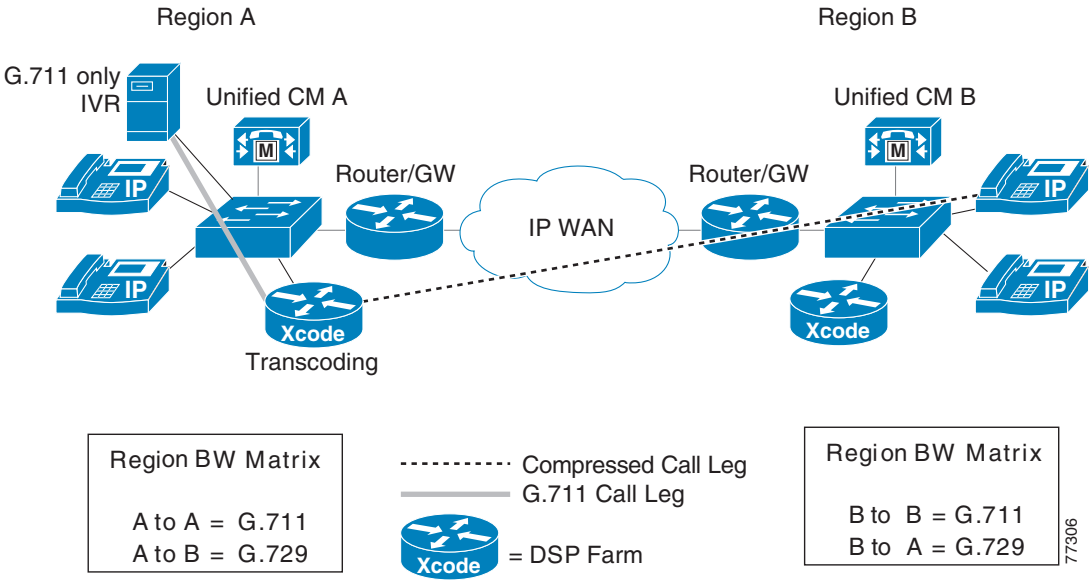
In distributed call processing deployments, several sites are connected through an IP WAN. Each site contains a Unified CM cluster that can, in turn, follow the single-site model or the centralized call processing model. A gatekeeper may be used for call admission control between sites.

Because WAN bandwidth is typically limited, calls between sites may be configured to use an LBR codec (such as G.729a) when traversing the WAN. H.323v2 intercluster trunks are used to connect Unified CM clusters. Unified CM also supports compressed voice call connections through the MTP service if a hardware MTP is used. (See Figure 17-11.)

A distributed call processing deployment might need transcoding and MTP services in the following situations:

- With current versions of Cisco applications, it is possible and recommended to avoid the use of transcoding resources. There might be specific instances where G.711 on a specific device cannot be avoided.
- Some endpoints (for example, video endpoints) do not support the H.323v2 features.

Figure 17-11 Intercluster Call Flow with Transcoding



Unified CM uses media resource groups (MRGs) to enable sharing of MTP and transcoding resources among the Unified CM servers within a cluster. In addition, for calls across intercluster trunks, MTP and transcoding resources are used only when needed, thus eliminating the need to configure the MTP service for applications that do not support LBR codecs.

The following characteristics apply to distributed call processing deployments:

- Only the intercluster calls that require transcoding will use the MTP service. For example, if both endpoints of a call are capable of using a G.729 codec, no transcoding resources will be used.
- Sharing MTP resources among servers within a cluster provides more efficient resource utilization.

Media Functions and Voice Quality

Any process that manipulates media can degrade the quality of the media. For example, encoding a voice stream for transmission across any network (IP or TDM) and decoding it at the other end will result in a loss of information, and the resulting voice stream will not be an exact reproduction of the original. If there are media traversal paths through the network that involve multiple encoding and decoding steps of the same voice stream, then each successive encoding/decoding operation will further degrade the voice quality. In general, such paths should be avoided. This is especially true for low-bandwidth codecs (LBC) such as G.729.

If such paths cannot be avoided, voice quality can generally be improved by using a higher bandwidth, low-compression codec, such as the G.711 or G.722 codecs, which are recommended wherever such paths are anticipated. Use of lower bandwidth, higher compression codecs in such scenarios is not recommended.

Music on Hold Design Considerations

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 Media Streaming Application service parameter **Supported MoH Codecs** under the Clusterwide Unified CM Service Parameters Configuration. From the Supported MoH Codecs list under the Clusterwide Parameters, select all the desired codec types that should be allowed for MoH streams. 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. The actual codec used for a MoH event is determined by the Region settings of the MoH server and the device being put on hold (IP phone, gateway, and so forth). Therefore, assign the proper Region setting to your MoH servers and configure the desired Region Relationships to control the codec selection of MoH interactions.

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

Configured audio sources are shared among *all* MoH servers in the Unified CM cluster, requiring each audio source file to be uploaded to every MoH server within the 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 methods of providing additional sources, refer to the sections on Using Multiple Fixed (Live) Audio Sources, page 17-45, and Multicast MoH from Branch Routers, page 17-50.

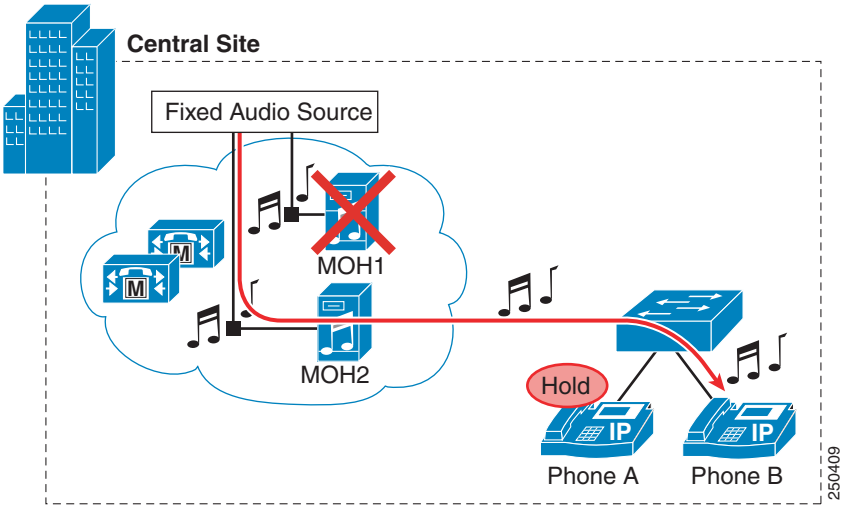
For those audio sources that will be used for multicast streaming, ensure that **Allow Multicasting** and **Play continuously (repeat)** are enabled. If continuous play of an audio source is not specified, only the first party placed on hold, not additional parties, will receive the MoH audio source.

Using Multiple Fixed (Live) Audio Sources

It is important to remember that only a single fixed audio source can be configured within Unified CM. However, each MoH server in the Unified CM cluster is capable of streaming a single fixed audio source by means of a Cisco MoH USB audio sound card (MOH-USB-AUDIO). When multiple fixed audio sources are needed, additional MoH servers can be added to provide these multiple sources. The audio supplied to each MoH server sound card can be the same or different, and the administrator can determine which MoH server is selected based on MRG and MRGL selections. When multiple audio sources are done in this manner, the holder's **User/Network Hold MoH Audio Source** should be configured for the fixed audio source (the single fixed audio source that is configured in Unified CM), and the MoH server to stream that fixed audio source to the device is then determined by the MRGL of the holdee.

In the case where the audio source is the same, this method also allows for redundancy of the fixed audio source. For example, in Figure 17-12 there are two MoH servers, each with an MOH-USB-AUDIO sound card connected to an audio source streaming audio derived from a live radio station feed. Phone B's MRGL contains first an MRG that contains the MOH1 server and second an MRG that contains the MOH2 server. Assuming the User/Network Hold Audio Source at Phone A has been set to the fixed audio source, after a call is established between Phone A and Phone B, and Phone B is placed on hold by Phone A, Phone B will receive the live feed audio source from the MOH1 server. In the case where the MOH1 server is down (or has no available capacity) when Phone A puts Phone B on hold, Phone B will receive the live feed audio source from the MOH2 server.

Figure 17-12 Fixed Audio Source Redundancy Example



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

Unicast and Multicast in the Same Unified CM Cluster

In some cases, administrators might want to configure a single Unified CM 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.
- Deploy a single MoH server with two media resource groups (MRGs), each containing 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-only MoH media resource and the multicast-enabled MoH media resource 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 whether the MRG is configured to use multicast and then on the server from which it is served.

When deploying a single MoH server for both unicast and multicast MoH, configure the server for multicast. Assign this same MoH media resource 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 MoH media 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 use 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. Lastly, configure a User Hold Audio Source and Network Hold Audio Source for each individual device or (in the case of phone devices) individual lines or directory numbers to assign 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.

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 a 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 Unified CM servers is automatically marked with a DSCP value of 24 or a PHB value of CS3 (ToS of 0x60) by default. Therefore, as long as QoS is properly configured on the network, this call signalling traffic will be properly classified and queued within the network along with all other call signalling traffic.

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 probable that voice media traffic might get delayed or dropped without appropriate call admission control. For additional information, see Call Admission Control, page 11-1.

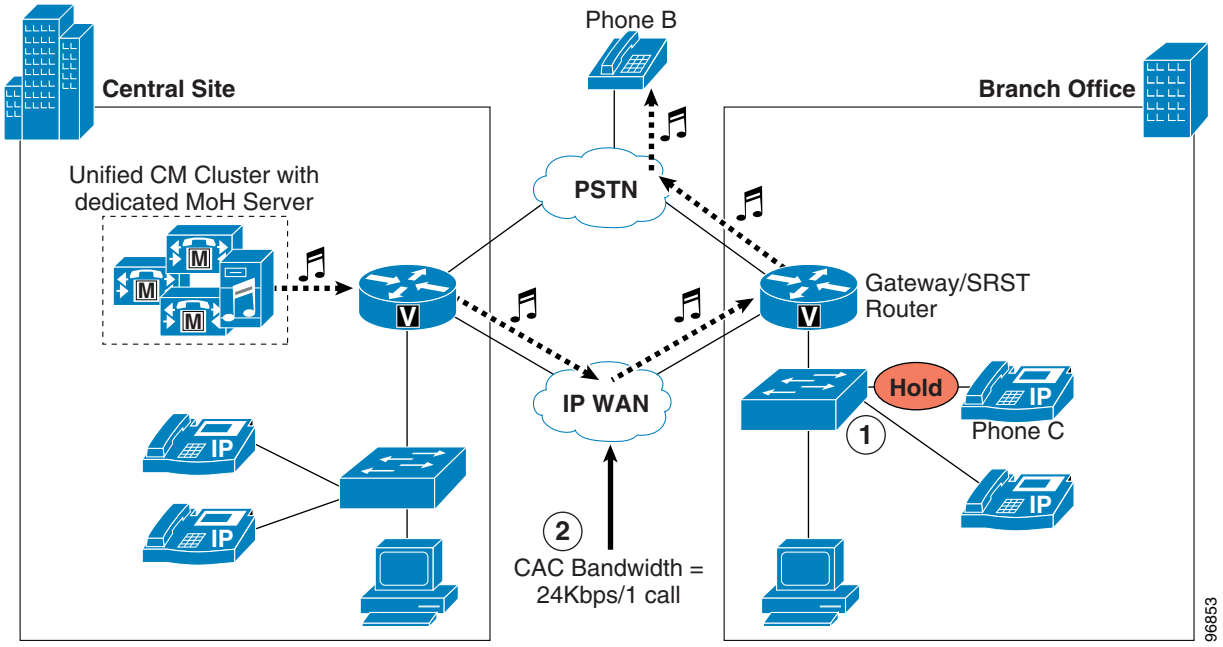
Call admission control for Unified CM (based on either static locations or RSVP-enabled locations) is capable of tracking unicast MoH streams traversing the WAN but not multicast MoH streams. Thus, even if WAN bandwidth has been fully subscribed, a multicast MoH stream will not be denied access to the 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 17-13 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 by way of 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 17-13 Locations-Based Call Admission Control and MoH



Deployment Models for Music on Hold

The various Unified Communications call processing 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 Unified Communications Deployment Models, page 5-1.

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 or areas of the network will be unable to support multicast. For this reason, you might have to deploy both unicast and multicast MoH resources. For more information, see the section on Unicast and Multicast in the Same Unified CM Cluster, page 17-46.

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.

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 Unified CM 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.

Multicast MoH from Branch Routers

Branch routers deployed with the Cisco Unified Survivable Remote Site Telephony (SRST) feature can provide multicast MoH in a remote or branch site, with the MoH streaming from the branch SRST router's flash or from a live feed connected to an analog port. Multicast MoH from a branch router via these two methods enhances the Cisco Unified Communications MoH feature in both of the following scenarios:

- Non-Fallback Mode

  When the WAN is up and the phones are controlled by Unified CM, this configuration can eliminate the need to forward MoH across the WAN to remote branch sites by providing locally sourced MoH.

- Fallback Mode

  When SRST is active and the branch devices have lost connectivity to the central-site Unified CM, the branch router can continue to provide multicast MoH.

When using the live feed option in either scenario, the SRST router provides redundancy by monitoring the live feed input, and it will revert to streaming MoH from a file in flash if the live feed is disconnected. You can use only a single multicast address and port number per SRST router to provide multicast MoH; therefore, the SRST router does not support streaming from both the live feed and the flash file at the same time. In addition, the SRST router can stream only a single audio file from flash.

Note An SRST license is required 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.

Non-Fallback Mode

During non-fallback mode (when the WAN is up and SRST is not active), the branch SRST router can provide multicast MoH to all local Cisco Unified Communications devices. To accomplish this, you must configure a Unified CM MoH 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 SRST router, 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.

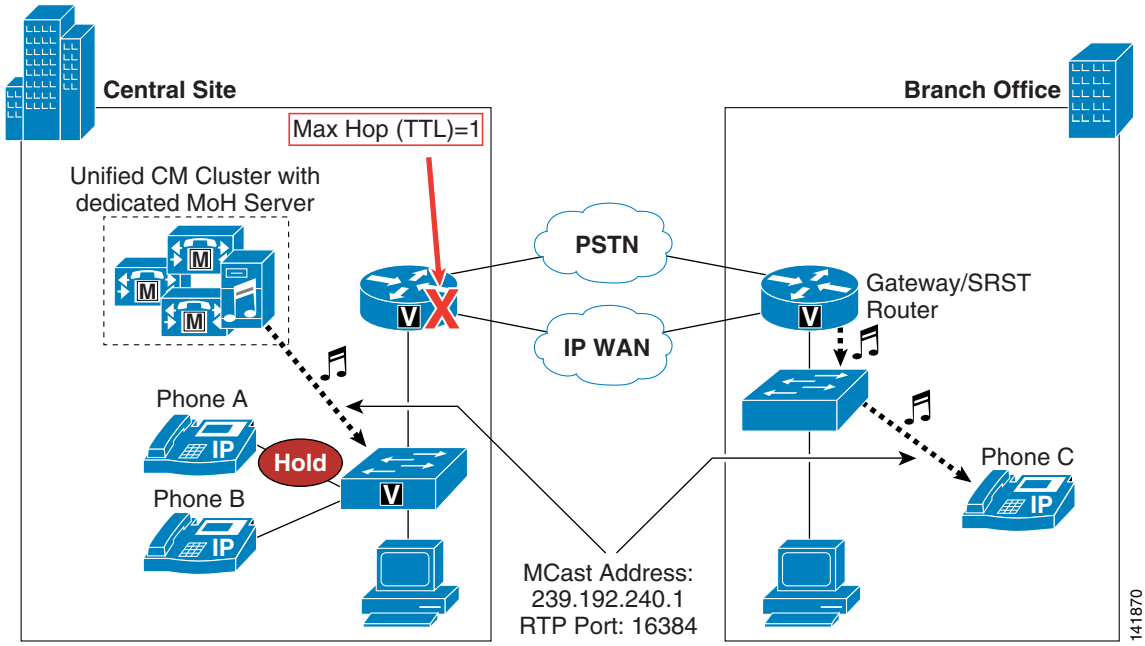
- Disable multicast routing on the WAN interface

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

Figure 17-14 illustrates streaming multicast MoH from a branch router when it is not in fallback mode. After phone A places phone C on hold, phone C receives multicast MoH from the local SRST router. In this example, 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 either flash or a live feed. 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 17-14 Multicast MoH from Branch Router



When using this method for delivering multicast MoH, configure all devices within the Unified CM 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.

In networks with multiple branch routers configured to stream multicast MoH, this allows for more than 51 unique MoH audio sources within the Unified CM cluster. Each branch site router can multicast a unique audio stream, 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 audio, then the cluster actually contains 101 unique MoH audio sources (100 branch streams and one central-site stream). If you want more than 51 unique audio streams in the central site, see the methods described in Using Multiple Fixed (Live) Audio Sources, page 17-45.

Fallback Mode

During fallback mode (when the WAN is down and SRST is active), the branch SRST router can stream multicast MoH to all analog and digital ports within the chassis, thereby providing MoH to analog phones and PSTN callers.

The branch router's configuration for fallback mode multicast MoH is the same as the normal operation configuration. However, which multicast address you configure on the router depends on the intended operation. If you want the branch router to provide multicast MoH to devices only in fallback mode (for example, if MoH received by remote devices is to be sourced from the central-site MoH server during non-fallback mode), then the multicast address and port number configured on the SRST 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.

Note that, once the branch SRST/gateway router is configured to provide multicast MoH, the router will continue to multicast the MoH stream even when not in fallback mode.

It is also possible to configure the fallback mode to use Cisco Unified Communications Manager Express (Unified CME) in SRST mode. Fallback mode behavior is still the same, but the configuration commands are slightly different. SRST commands are entered under the Cisco IOS **call-manager-fallback** construct, while the commands for Unified CME in SRST mode are entered under **telephony-service**.

There are four methods of providing multicast MoH via SRST:

- SRST multicast MoH from branch router flash
- SRST multicast MoH from a live feed
- Unified CME in SRST mode with multicast MoH from branch router flash
- Unified CME in SRST mode with multicast MoH from a live feed

For more details on configuration of Cisco Unified SRST and Unified CME, refer to the following documentation:

- *Cisco Unified SRST System Administrator Guide*, available at

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

- *Cisco Unified Communications Manager Express System Administrator Guide*, available at

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

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 Unified CM 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.

Another option is to provision multicast MoH for intercluster calls across an intercluster trunk (ICT) or SIP trunk. This allows endpoints in one Unified CM cluster to hear multicast MoH streamed from another Unified CM cluster, while making more efficient use of intercluster bandwidth. A properly designed IP Multicast environment is required to take advantage of this feature. For more information on IP Multicast, refer to the documentation available at

http://www.cisco.com/en/US/products/ps6552/products_ios_technology_home.html

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 Unified CM 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.