CHAPTER **13**

Call Admission Control

Revised: January 15, 2015; OL-30952-03

The call admission control function is an important component of any IP telephony system, especially those that involve multiple sites connected through an IP WAN. In order to better understand what call admission control does and why it is needed, consider the example in Figure 13-1.

Figure 13-1   Why Call Admission Control is Needed
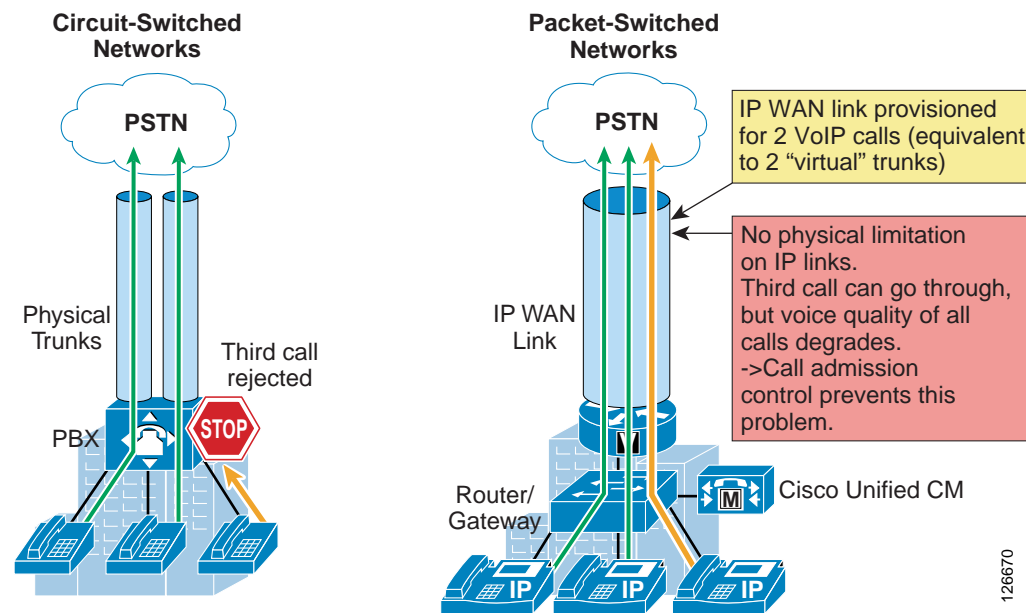
As shown on the left side of Figure 13-1, traditional TDM-based PBXs operate within circuit-switched networks, where a circuit is established each time a call is set up. As a consequence, when a legacy PBX is connected to the PSTN or to another PBX, a certain number of physical trunks must be provisioned. When calls have to be set up to the PSTN or to another PBX, the PBX selects a trunk from those that are available. If no trunks are available, the call is rejected by the PBX and the caller hears a network-busy signal.

Now consider the IP telephony system shown on the right side of Figure 13-1. Because it is based on a packet-switched network (the IP network), no circuits are established to set up an IP telephony call. Instead, the IP packets containing the voice samples are simply routed across the IP network together

with other types of data packets. Quality of Service (QoS) is used to differentiate the voice packets from the data packets, but bandwidth resources, especially on IP WAN links, are not infinite. Therefore, network administrators dedicate a certain amount of "priority" bandwidth to voice traffic on each IP WAN link. However, once the provisioned bandwidth has been fully utilized, the IP telephony system must reject subsequent calls to avoid oversubscription of the priority queue on the IP WAN link, which would cause quality degradation for all voice calls. This function is known as call admission control, and it is essential to guarantee good voice quality in a multisite deployment involving an IP WAN.

To preserve a satisfactory end-user experience, the call admission control function should always be performed during the call setup phase so that, if there are no network resources available, a message can be presented to the end-user or the call can be rerouted across a different network (such as the PSTN).

This chapter discusses the following main topics:

- Call Admission Control Architecture, page 13-3

  This section describes the call admission control mechanism available through Cisco Unified Communications Manager called Enhanced Location Call Admission Control. For information regarding Cisco IOS gatekeeper, RSVP, and RSVP SIP Preconditions, refer to the *Call Admission Control* chapter of the *Cisco Unified Communications System 9.0 SRND*, available at

    http://www.cisco.com/en/US/docs/voice_ip_comm/cucm/srnd/9x/cac.html

- Design Considerations for Call Admission Control, page 13-36

  This section shows how to apply Enhanced Location Call Admission Control based on the IP WAN topology.

What's New in This Chapter

Table 13-1 lists the topics that are new in this chapter or that have changed significantly from previous releases of this document.

Table 13-1   New or Changed Information Since the Previous Release of This Document

New or Revised Topic	Described in:	Revision Date
Location Bandwidth Manager (LBM)	Location Bandwidth Manager, page 13-11 Enhanced Location CAC Design and Deployment Recommendations and Considerations, page 13-13	January 15, 2015
Call admission control for video deployments	Call Admission Control Design Recommendations for Video Deployments, page 13-42	January 15, 2015
Extension Mobility Cross Cluster (EMCC)	Extension Mobility Cross Cluster with Enhanced Location CAC, page 13-36	November 19, 2013
Cisco Expressway	Design Recommendations for Cisco Expressway Deployments with Enhanced Location CAC, page 13-49	November 19, 2013

Call Admission Control Architecture

There are several mechanisms that perform the call admission control function in a Cisco Collaboration system. This section provides design and configuration guidelines for Enhanced Location Call Admission Control based on Cisco Unified CM. For all other call admission control mechanisms, refer to the following information:

- Topology-unaware mechanisms
  - Cisco IOS Gatekeeper Zones, page 13-3
- Topology-aware mechanisms
  - Unified Communications Architectures Using Resource Reservation Protocol (RSVP), page 13-3

Cisco IOS Gatekeeper Zones

Gatekeeper is a legacy technology used for dial-plan resolution and admission control in H.323 networks. With most networks today supporting SIP or mixed H.323 and SIP trunking, gatekeeper is not an effective technology for admission control supporting these environments. For networks supporting H.323 networking or for internetworking call processing agents over H.323 and not mixed SIP networks, refer to the Cisco IOS Gatekeeper information in the *Call Admission Control* chapter of the *Cisco Unified Communications System 9.0 SRND*, available at

http://www.cisco.com/en/US/docs/voice_ip_comm/cucm/srnd/9x/cac.html

Unified Communications Architectures Using Resource Reservation Protocol (RSVP)

Resource Reservation Protocol (RSVP) is a powerful technology useful in achieving endpoint QoS trust and admission control in Unified Communications voice and video networks. RSVP-enabled networks are possible although not recommended in Unified Communications and Collaboration deployments where TelePresence is pervasively deployed. The reason is that TelePresence requires a QoS overlay network and does not support RSVP call admission control. This means that deployments where RSVP call admission control for desktop video as well as telepresence-to-desktop video interoperation are desired, require a more complex design to ensure that telepresence-to-telepresence calls are provisioned without RSVP while all other calls are provisioned for RSVP. Due to this requirement, Enhanced Location Call Admission Control is preferred for pervasively deployed TelePresence and mixed Collaboration video and TelePresence deployments.

For further information on RSVP deployments, refer to the Resource Reservation Protocol (RSVP) information in the *Call Admission Control* chapter of the *Cisco Unified Communications System 9.0 SRND*, available at

http://www.cisco.com/en/US/docs/voice_ip_comm/cucm/srnd/9x/cac.html

Unified CM Enhanced Location Call Admission Control

Cisco Unified CM provides Enhanced Location Call Admission Control (ELCAC) to support complex WAN topologies as well as distributed deployments of Unified CM for call admission control where multiple clusters manage devices in the same physical sites using the same WAN uplinks. The Enhanced Location CAC feature also supports immersive video, allowing the administrator to control call admissions for immersive video calls such as TelePresence separately from other video calls.

To support more complex WAN topologies Unified CM has implemented a location-based network modeling functionality. This provides Unified CM with the ability to support multi-hop WAN connections between calling and called parties. This network modeling functionality has also been incrementally enhanced to support multi-cluster distributed Unified CM deployments. This allows the administrator to effectively "share" locations between clusters by enabling the clusters to communicate with one another to reserve, release, and adjust allocated bandwidth for the same locations across clusters. In addition, an administrator has the ability to provision bandwidth separately for immersive video calls such as TelePresence by allocating a new field to the Location configuration called **immersive video bandwidth**.

There are also tools to administer and troubleshoot Enhanced Location CAC. The CAC enhancements and design are discussed in detail in this chapter, but the troubleshooting and serviceability tools are discussed in separate product documentation.

Network Modeling with Locations, Links, and Weights

Enhanced Location CAC is a model-based static CAC mechanism. Enhanced Location CAC involves using the administration interface in Unified CM to configure Locations and Links to model the "Routed WAN Network" in an attempt to represent how the WAN network topology routes media between groups of endpoints for end-to-end audio, video, and immersive calls. Although Unified CM provides configuration and serviceability interfaces in order to model the network, it is still a "static" CAC mechanism that does not take into account network failures and network protocol rerouting. Therefore, the model needs to be updated when the WAN network topology changes. Enhanced Location CAC is also call oriented, and bandwidth deductions are per-call not per-stream, so asymmetric media flows where the bit-rate is higher in one direction than in the other will always deduct for the highest bit rate. In addition, unidirectional media flows will be deducted as if they were bidirectional media flows.

Enhanced Location CAC incorporates the following configuration components to allow the administrator to build the network model using Locations and Links:

- Locations — A Location represents a LAN. It could contain endpoints or simply serve as a transit location between links for WAN network modeling. For example, an MPLS provider could be represented by a Location.

- Links — Links interconnect locations and are used to define bandwidth available between locations. Links logically represent the WAN link and are configured in the Location user interface (UI).

- Weights — A weight provides the relative priority of a link in forming the *effective path* between any pair of locations. The effective path is the path used by Unified CM for the bandwidth calculations, and it has the least cumulative weight of all possible paths. Weights are used on links to provide a "cost" for the "effective path" and are pertinent only when there is more than one path between any two locations.

- Path — A path is a sequence of links and intermediate locations connecting a pair of locations. Unified CM calculates least-cost paths (lowest cumulative weight) from each location to all other locations and builds a map of the various paths. Only one "effective path" is used between any pair of locations.

- Effective Path — The effective path is the path with the least cumulative weight.

- Bandwidth Allocation — The amount of bandwidth allocated in the model for each type of traffic: audio, video, and immersive video (TelePresence).

- Location Bandwidth Manager (LBM) — The active service in Unified CM that assembles a network model from configured location and link data in one or more clusters, determines the effective paths between pairs of locations, determines whether to admit calls between a pair of locations based on the availability of bandwidth for each type of call, and deducts (reserves) bandwidth for the duration of each call that is admitted.

- Location Bandwidth Manager Hub — A Location Bandwidth Manager (LBM) service that has been designated to participate directly in intercluster replication of fixed locations, links data, and dynamic bandwidth allocation data. LBMs assigned to an LBM hub group discover each other through their common connections and form a fully-meshed intercluster replication network. Other LBM services in a cluster with an LBM hub participate indirectly in intercluster replication through the LBM hubs in their cluster.

Locations and Links

Unified CM uses the concept of locations to represent a physical site and to create an association with media devices such as endpoints, voice messaging ports, trunks, gateways, and so forth, through direct configuration on the device itself, through a device pool, or even through device mobility. Unified CM also uses a new location configuration parameter called *links*. Links interconnect locations and are used to define bandwidth available between locations. Links logically represent the WAN links. This section describes locations and links and how they are used.

The location configuration itself consists of three main parts: links, intra-location bandwidth parameters, and RSVP locations settings. The RSVP locations settings are not considered here for Enhanced Location CAC because they apply only to RSVP implementations. In the configuration, the link bandwidth parameters are displayed first while the intra-location bandwidth parameters are hidden and displayed by selecting the **Show advanced** link.

The intra-location bandwidth parameters allow the administrator to configure bandwidth allocations for three call types: audio, video, and immersive. They limit the amount of traffic within, as well as to or from, any given location. When any device makes or receives a call, bandwidth is deducted from the applicable bandwidth allocation for that call type. This feature allows administrators to limit the amount of bandwidth used on the LAN or transit location. In most networks today that consist of Gigabit LANs,

there is little or no reason to limit bandwidth on those LANs.

The link bandwidth parameters allow the administrator to characterize the provisioned bandwidth for audio, video, and immersive calls between "adjacent locations" (that is, locations that have a link configured between them). This feature offers the administrator the ability to create a string of location pairings in order to model a multi-hop WAN network. To illustrate this, consider a simple three-hop WAN topology connecting four physical sites, as shown in Figure 13-2. In this topology we want to create links between San Jose and Boulder, between Boulder and Richardson, and between Richardson and RTP. Note that when we create a link from San Jose to Boulder, for example, the inverse link (Boulder to San Jose) also exists. Therefore, the administrator needs to create the link pairing only once from either location configuration page. In the example in Figure 13-2, each of the three links has the same settings: a weight of 50, 240 kbps of audio bandwidth, 500 kbps of video bandwidth, and 5000 kbps (or 5 MB) of immersive bandwidth.

*Figure 13-2*   *Simple Link Example with Three WAN Hops*
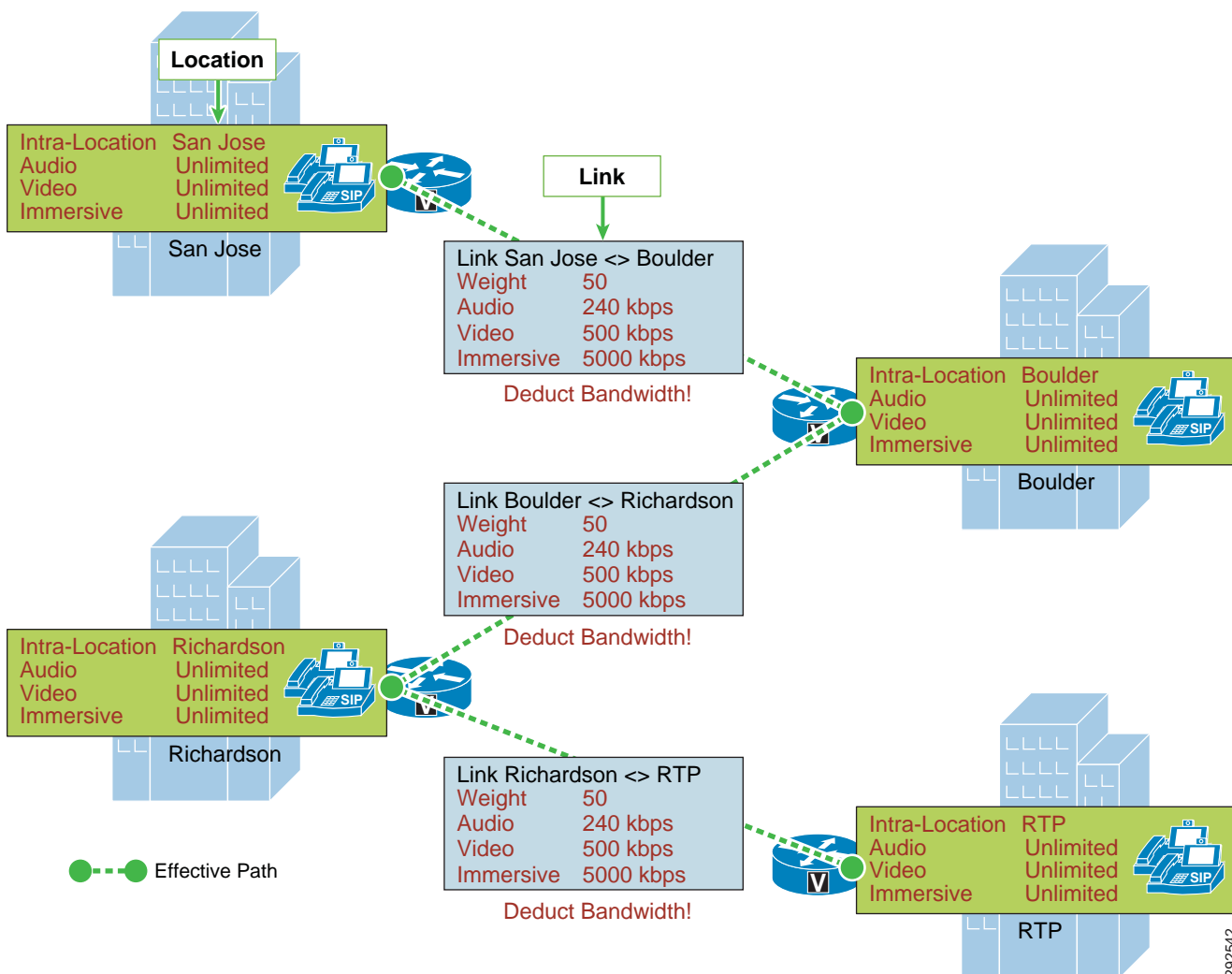
When a call is made between San Jose and RTP, Unified CM calculates the bandwidth of the requested call, which is determined by the region pair between the two devices (see Locations, Links, and Region Settings, page 13-10) and verifies the effective path between the two locations. That is to say,

Unified CM verifies the locations and links that make up the path between the two locations and accordingly deducts bandwidth from each link and (if applicable) from each location in the path. The intra-location bandwidth also is deducted along the path if any of the locations has configured a bandwidth value other than unlimited.

Weight is configurable on the link only and provides the ability to force a specific path choice when multiple paths between two locations are available. When multiple paths are configured, only one will be selected based on the cumulative weight, and this path is referred to as the *effective path*. This weight is static and the effective path does not change dynamically. Figure 13-3 illustrates weight configured on links between three locations: San Jose, Boulder, and Seattle.

Figure 13-3   Cumulative Path Weights
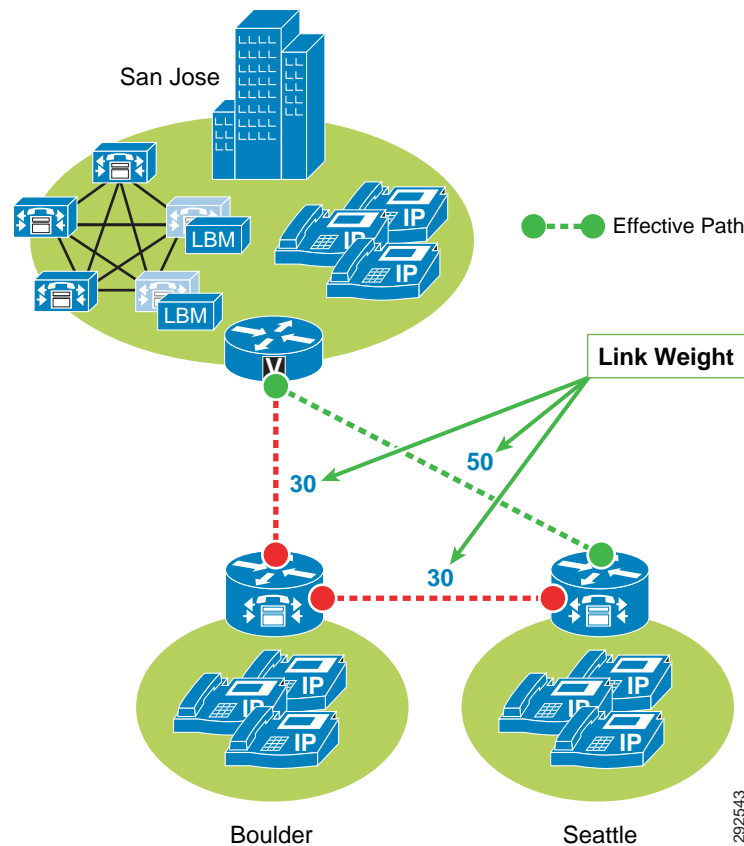
San Jose to Seattle has two paths, one direct link between the locations and another path through the Boulder location (link San Jose/Boulder and link Boulder/Seattle). The weight configured on the direct link between San Jose and Seattle is 50 and is less than the cumulative weight of links San Jose/Boulder and Boulder/Seattle which is 60 (30+30). Thus, the direct link is chosen as the effective path because the cumulative link weight is 50.

When you configure a device in Unified CM, the device can be assigned to a location. A location can be configured with links to other locations in order to build a topology. The locations configured in Unified CM are virtual locations and not real, physical locations. As mentioned, Unified CM has no knowledge of the actual physical topology of the network. Therefore, any changes to the physical network must be made manually in Unified CM to map the real underlying network topology with the Unified CM locations model. If a device is moved from one physical location to another, the system

administrator must either perform a manual update on its location configuration or else implement the device mobility feature so that Unified CM can correctly calculate bandwidth allocations for calls to and from that device. Each device is in location **Hub_None** by default. Location Hub_None is an example location that typically serves as a hub linking two or more locations, and it is configured by default with unlimited intra-location bandwidth allocations for audio, video, and immersive bandwidth.

Unified CM allows you to define separate voice, video, and immersive video bandwidth pools for each location and link between locations. Typically the locations intra-location bandwidth configuration is left as a default of **Unlimited** while the link between locations is set to a finite number of kilobits per second (kbps) to match the capacity of a WAN links between physical sites. If the location's intra-location audio, video, and immersive bandwidths are configured as **Unlimited**, there will be unlimited bandwidth available for all calls (audio, video, and immersive) within that location and transiting that location. On the other hand, if the bandwidth values are set to a finite number of kilobits per second (kbps), Unified CM will track all calls within the location and all calls that use the location as a transit location (a location that is in the calculation path but is not the originating or terminating location in the path).

For video calls, the video location bandwidth takes into account both the audio and the video portions of the video call. Therefore, for a video call, no bandwidth is deducted from the audio bandwidth pool. The same applies to immersive video calls.

The devices that can specify membership in a location include:

- IP phones
- CTI ports
- H.323 clients
- CTI route points
- Conference bridges
- Music on hold (MoH) servers
- Gateways
- Trunks
- Media termination point (via device pool)
- Trusted relay point (via device pool)
- Annunciator (via device pool)

The Enhanced Location Call Admission Control mechanism also takes into account the mid-call changes in call type. For example, if an inter-site video call is established, Unified CM will subtract the appropriate amount of video bandwidth from the respective locations and links in the path. If this video call changes to an audio-only call as the result of a transfer to a device that is not capable of video, Unified CM will return the allocated bandwidth to the video pool and allocate the appropriate amount of bandwidth from the audio pool along the same path. Calls that change from audio to video will cause the opposite change of bandwidth allocation.

Table 13-2 lists the amount of bandwidth requested by the static locations algorithm for various call speeds. For an audio call, Unified CM counts the media bit rates plus the IP and UDP overhead. For example, a G.711 audio call consumes 80 kbps (64k bit rate + 16k IP/UDP headers) deducted from the location's and link's audio bandwidth allocation. For a video call, Unified CM counts only the media bit rates for both the audio and video streams. For example, for a video call at a bit rate of 384 kbps, Unified CM will allocate 384 kbps from the video bandwidth allocation.

Table 13-2   *Amount of Bandwidth Requested by the Locations and Links Bandwidth Deduction Algorithm*

Call Speed	Static Location and Link Bandwidth Value
G.711 audio call (64 kbps)	80 kbps
G.729 audio call (8 kbps)	24 kbps
128 kbps video call	128 kbps
384 kbps video call	384 kbps
512 kbps video call	512 kbps
768 kbps video call	768 kbps

For a complete list of codecs and location and link bandwidth values, refer to the bandwidth calculations information in the *Call Admission Control* section of the *Cisco Unified Communications Manager System Guide*, available at

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

For example, assume that the link configuration for the location Branch 1 to Hub_None allocates 256 kbps of available audio bandwidth and 384 kbps of available video bandwidth. In this case the path from Branch 1 to Hub_None can support up to three G.711 audio calls (at 80 kbps per call) or ten G.729 audio calls (at 24 kbps per call), or any combination of both that does not exceed 256 kbps. The link between locations can also support different numbers of video calls depending on the video and audio codecs being used (for example, one video call requesting 384 kbps of bandwidth or three video calls with each requesting 128 kbps of bandwidth).

When a call is placed from one location to the other, Unified CM deducts the appropriate amount of bandwidth from the effective path of locations and links from one location to another. Using Figure 13-2 as an example, a G.729 call between San Jose and RTP locations causes Unified CM to deduct 24 kbps from the available bandwidth at the links between San Jose and Boulder, between Boulder and Richardson, and between Richardson and RTP. When the call has completed, Unified CM returns the bandwidth to those same links over the effective path. If there is not enough bandwidth at any one of the links over the path, the call is denied by Unified CM and the caller receives the network busy tone. If the calling device is an IP phone with a display, that device also displays the message "Not Enough Bandwidth."

When an inter-location call is denied by call admission control, Unified CM can automatically reroute the call to the destination through the PSTN connection by means of the Automated Alternate Routing (AAR) feature. For detailed information on the AAR feature, see Automated Alternate Routing, page 14-78.

Note   AAR is invoked only when Enhanced Location Call Admission Control denies the call due to a lack of network bandwidth along the effective path. AAR is not invoked when the IP WAN is unavailable or other connectivity issues cause the called device to become unregistered with Unified CM. In such cases, the calls are redirected to the target specified in the Call Forward No Answer field of the called device.

It is also worth noting that video devices can be enabled to **Retry Video Call as Audio** if a video call between devices fails CAC. This option is configured on the video endpoint configuration page in Unified CM and is applicable to video endpoints or trunks receiving calls. It should also be noted that for some video endpoints **Retry Video Call as Audio** is enabled by default and not configurable on the endpoint.

Locations, Links, and Region Settings

Locations work in conjunction with regions to define the characteristics of a call over the effective path of locations and links. Regions define the type of compression or bit rate (8 kbps or G.729, 64 kbps or G.722/G.711, and so forth) that is used between devices, and location links define the amount of available bandwidth for the effective path between devices. You assign each device in the system to both a region (by means of a device pool) and a location (by means of a device pool or by direct configuration on the device itself).

You can configure locations in Unified CM to define:

- Physical sites (for example, a branch office) or transit sites (for example, an MPLS cloud) — A location represents a LAN. It could contain endpoints or simply serve as a transit location between links for WAN network modeling.

- Link bandwidth between adjacent locations — Links interconnect locations and are used to define bandwidth available between locations. Links logically represent the WAN link between physical sites.

  – Audio Bandwidth — The amount of bandwidth that is available in the WAN link for voice and fax calls being made from devices in the location to the configured adjacent location. This bandwidth value is used by Unified CM for Enhanced Location Call Admission Control.

  – Video Bandwidth — The amount of video bandwidth that is available in the WAN link for video calls being made from devices in the location to the configured adjacent location. This bandwidth value is used by Unified CM for Enhanced Location Call Admission Control.

  – Immersive Video Bandwidth — The amount of immersive bandwidth that is available in the WAN link for TelePresence calls being made from devices in the location to the configured adjacent location. This bandwidth value is used by Unified CM for Enhanced Location Call Admission Control.

- Intra-location bandwidth

  – Audio Bandwidth — The amount of bandwidth that is available in the LAN for voice and fax calls being made from devices within the location. This bandwidth value is used by Unified CM for Enhanced Location Call Admission Control.

  – Video Bandwidth — The amount of video bandwidth that is available in the LAN for video calls being made from devices within the location. This bandwidth value is used by Unified CM for Enhanced Location Call Admission Control.

  – Immersive Video Bandwidth — The amount of immersive bandwidth that is available in the LAN for TelePresence calls being made from devices within the location. This bandwidth value is used by Unified CM for Enhanced Location Call Admission Control.

You can configure regions in Unified CM to define:

- The Maximum Audio Bit Rate used for intraregion and interregion calls

- The Maximum Session Bit Rate for Video Calls (Includes Audio) used for intraregion and interregion calls

- The Maximum Session Bit Rate for Immersive Video Calls (Includes Audio) used for intraregion and interregion calls

- Audio codec preference lists

Unified CM Support for Locations and Regions

Cisco Unified Communications Manager supports 2,000 locations and 2,000 regions per cluster. To deploy up to 2,000 locations and regions, you must configure the following service parameters in the **Clusterwide Parameters** > **(System - Location and Region) and Clusterwide Parameters** > **(System - RSVP)** configuration menus:

- Default Intraregion Max Audio Bit Rate

- Default Interregion Max Audio Bit Rate

- Default Intraregion Max Video Call Bit Rate (Includes Audio)

- Default Interregion Max Video Call Bit Rate (Includes Audio)

- Default Intraregion Max Immersive Call Bit Rate (Includes Audio)

- Default Interregion Max Immersive Video Call Bit Rate (Includes Audio)

- Default Audio Codec Preference List between Regions

- Default Audio Codec Preference List within Regions

When adding regions, you should select **Use System Default** for the Maximum Audio Bit Rate and Maximum Session Bit Rate for Video Call values.

Changing these values for individual regions and locations from the default has an impact on server initialization and publisher upgrade times. Hence, with a total of 2,000 regions and 2,000 locations, you can modify up to 200 of them to use non-default values. With a total of 1,000 or fewer regions and locations, you can modify up to 500 of them to use non-default values. Table 13-3 summarizes these limits.

Table 13-3   Number of Allowed Non-Default Regions and Locations

Number of non-default regions and locations	Maximum number of regions	Maximum number of locations
0 to 200	2,000	2,000
200 to 500	1,000	1,000

Note   The Maximum Audio Bit Rate is used by both voice calls and fax calls. If you plan to use G.729 as the interregion codec, use T.38 Fax Relay for fax calls. If you plan to use fax pass-through over the WAN, use audio preference lists to prefer G.729 for audio-only calls and G.711 for fax calls.

Location Bandwidth Manager

The Location Bandwidth Manager (LBM) is a Unified CM Feature Service managed from the serviceability web pages and is responsible for all of the Enhanced Location CAC bandwidth functions. The LBM can run on any Unified CM subscriber node or as a standalone service on a dedicated Unified CM node in the cluster. A minimum of one instance of LBM must run in each cluster to enable Enhanced Location CAC in the cluster. However, Cisco recommends running LBM on each subscriber node in the cluster that is also running the Cisco CallManager service.

The LBM performs the following functions:

- Assembles topology of locations and links

- Calculates the effective paths across the topology

- Services bandwidth requests from the Cisco CallManager service (Unified CM call control)
- Replicates the bandwidth information to other LBMs
- Provides configured and dynamic information to serviceability
- Updates Location Real-Time Monitoring Tool (RTMT) counters

The LBM Service is enabled by default when upgrading Cisco Unified CM from earlier releases that support only traditional Location CAC. For new installations, the LBM service must be activated manually.

During initialization, the LBM reads local locations information from the database, such as: locations audio, video, and immersive bandwidth values; intra-location bandwidth data; and location-to-location link audio, video, and immersive bandwidth values and weight. Using the link data, each LBM in a cluster creates a local assembly of the paths from one location to every other location. This is referred to as the *assembled topology*. In a cluster, each LBM accesses the same data and thus creates the same local copy of the assembled topology during initialization.

At runtime, the LBM applies reservations along the computed paths in the local assembled topology of locations and links, and it replicates the reservations to other LBMs in the cluster. If intercluster Enhanced Location CAC is configured and activated, the LBM can be configured to replicate the assembled topology to other clusters (see Intercluster Enhanced Location CAC, page 13-14, for more details).

By default the Cisco CallManager service communicates with the local LBM service; however, LBM groups can be used to manage this communication. LBM groups provide an active and standby LBM in order to create redundancy for Unified CM call control. Figure 13-4 illustrates LBM redundancy.

Figure 13-4   Location Bandwidth Manager Redundancy
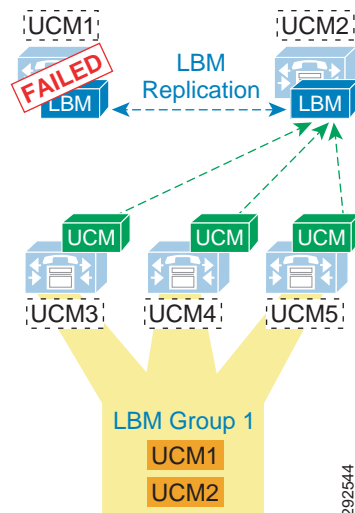
Figure 13-4 shows five Unified CM servers: UCM1 and UCM2 are dedicated LBM servers (only LBM service enabled); UCM3, UCM4, and UCM5 are Unified CM subscribers (Cisco CallManager service enabled). An LBM Group has been configured with UCM1 as active and UCM2 as standby, and it is applied to subscribers UCM3, UCM4, and UCM5. This configuration allows for UCM3, UCM4, and UCM5 to query UCM1 for all bandwidth requests. If UCM1 fails for any reason, the subscribers will fail-over to the standby UCM2. This example is used to illustrate how the LBM Group configuration functions and is not a recommended configuration (see recommendations below).

Because LBM is directly involved in processing requests for every call that is processed by the CallManager service that it is serving, it is important to follow these simple design recommendations in order to ensure optimal functionality of the LBM.

The recommended configuration is to run LBM co-resident with the Cisco CallManager service (call processing). If redundancy of the LBM service is required, it is important not to over-subscribe any given LBM. It is also important make sure LBM is no more than a primary and secondary in any given deployment. This means that LBM should not have the load of more than 2 CallManager services during failure scenarios, and the load of only one CallManager service during normal operation. The LBM group can be used to configure a co-resident LBM as the primary, another local (on the same LAN) LBM as secondary, and lastly the service parameter as a failsafe mechanism to ensure that all calls processed by that CallManager service do not fail. There are many reasons for these recommendations. It is difficult, at best, to determine the load of any LBM because it is directly related to the call-processing load of the CallManager service that it is serving. There are also considerations for delay. As soon as an LBM is off-box from a CallManager service, there is an added delay caused by packetization and processing for every call serviced by the CallManager service. Compounding call-processing delay can bring the overall delay budget to an unacceptable level for any given call flow to a ringing state, and result in a poor user experience. Following these design recommendations will reduce the overall call-processing delay.

The order in which the Unified CM Cisco CallManager service uses the LBM is as follows:

- LBM Group designation
- Local LBM (co-resident)
- Service parameter **Call Treatment when no LBM available** (Default = **allow calls**)

Enhanced Location CAC Design and Deployment Recommendations and Considerations

- The Location Bandwidth Manager (LBM) is a Unified CM Feature Service. It is responsible for modeling the topology and servicing Unified CM bandwidth requests.
- LBMs within the cluster create a fully meshed communications network via XML over TCP for the replication of bandwidth change notifications between LBMs.
- Cisco recommends deploying the LBM service co-resident with a Unified CM subscriber running the Cisco CallManager call processing service.
- If redundancy is required for the LBM service, create an LBM Group for each Unified CM subscriber running the Cisco CallManager call processing service. Add the co-resident LBM service as the primary LBM, and the LBM from another Unified CM subscriber on the same LAN as a secondary LBM. This will ensure that the Cisco CallManager call processing service uses the co-resident LBM as primary, the LBM on another local (same LAN) Unified CM subscriber as secondary, and the service parameter **Call Treatment when no LBM available** as tertiary source for LBM requests.

Note   Cisco recommends having LBM back up more than one Cisco CallManager service. Assuming that the LBM is handling the load of the co-resident CallManager service, and during failure of another CallManager service, this would equate to the load of 2 call processing servers on a single LBM.

- For deployments with cluster over the WAN and local failover, intracluster LBM traffic is already calculated into the WAN bandwidth calculations. See the section on clustering over the WAN Local Failover Deployment Model, page 10-47, for details on bandwidth calculations.

Intercluster Enhanced Location CAC

Intercluster Enhanced Location CAC extends the concept of network modeling across multiple clusters. In intercluster Enhanced Location CAC, each cluster manages its locally configured topology of locations and links and then propagates this local topology to other remote clusters that are part of the LBM intercluster replication network. Upon receiving a remote cluster's topology, the LBM assembles this into its own local topology and creates a global topology. Through this process the global topology is then identical across all clusters, providing each cluster a global view of enterprise network topology for end-to-end CAC. Figure 13-5 illustrates the concept of a global topology with a simplistic hub-and-spoke network topology as an example.

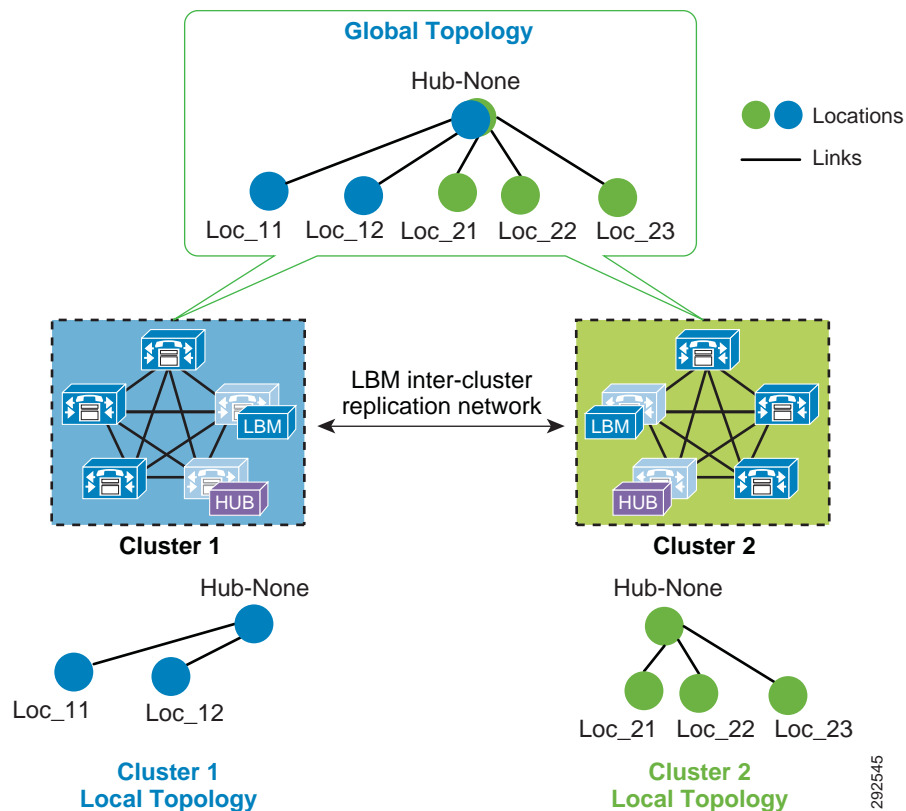*Figure 13-5*   *Example of a Global Topology for a Simple Hub-and-Spoke Network*

Figure 13-5 shows two clusters, Cluster 1 and Cluster 2, each with a locally configured hub-and-spoke network topology. Cluster 1 has configured Hub_None with links to Loc_11 and Loc_12, while Cluster 2 has configured Hub_None with links to Loc_21, Loc_22, and Loc_23. Upon enabling intercluster Enhanced Location CAC, Cluster 1 sends its local topology to Cluster 2, as does Cluster 2 to Cluster 1. After each cluster obtains a copy of the remote cluster's topology, each cluster overlays the remote cluster's topology over their own. The overlay is accomplished through common locations, which are locations that are configured with the same name. Because both Cluster 1 and Cluster 2 have the common location Hub_None with the same name, each cluster will overlay the other's network topology with Hub_None as a common location, thus creating a global topology where Hub_None is the hub and Loc_11, Loc_12, Loc_21, Loc_22 and Loc_23 are all spoke locations. This is an example of a simple network topology, but more complex topologies would be processed in the same way.
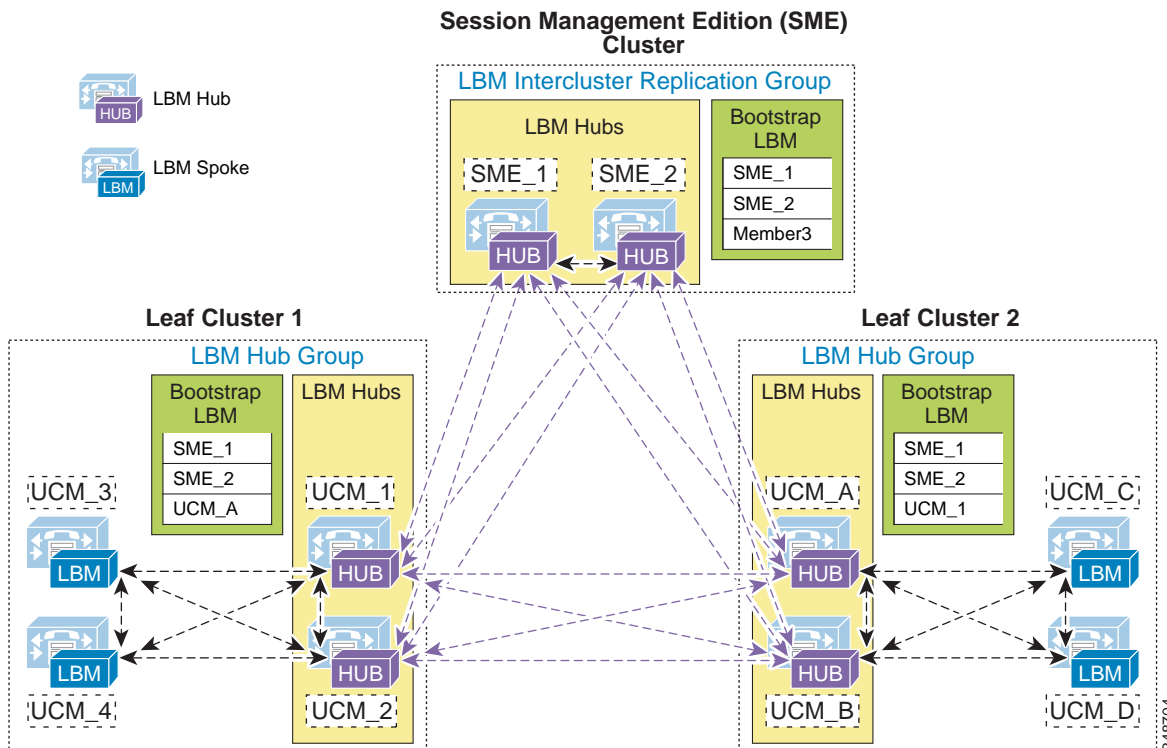
LBM Hub Replication Network

The intercluster LBM replication network is a separate replication network of designated LBMs called LBM hubs. LBM hubs create a separate full mesh with one another and replicate their local cluster's topology to other remote clusters. Each cluster effectively receives the topologies from every other remote cluster in order to create a global topology. The designated LBMs for the intercluster replication network are called *LBM hubs*. The LBMs that replicate only within a cluster are called *LBM spokes*. The LBM hubs are designated in configuration through the LBM **intercluster replication group**. The LBM role assignment for any LBM in a cluster can also be changed to a hub or spoke role in the intercluster replication group configuration (For further information on the LBM hub group configuration, refer to the Cisco Unified Communications Manager product documentation available at http://www.cisco.com/en/US/products/sw/voicesw/ps556/tsd_products_support_series_home.html.)

In the LBM intercluster replication group, there is also a concept of bootstrap LBM. Bootstrap LBMs are LBM hubs that provide all other LBM hubs with the connectivity details required to create the full-mesh hub replication network. Bootstrap LBM is a role that any LBM hub can have. If all LBM hubs point to a single LBM hub, that single LBM hub will tell all other LBM hubs how to connect to one another. Each replication group can reference up to three bootstrap LBMs.

Once the LBM hub group is configured on each cluster, the designated LBM hubs will create the full-mesh intercluster replication network. Figure 13-6 illustrates an intercluster replication network configuration with LBM hub groups set up between three clusters (Leaf Cluster 1, Leaf Cluster 2, and a Session Management Edition (SME) cluster) to form the intercluster replication network. The SME cluster is used here only as an example and is not required for this specific setup. The SME cluster could simply be another regular cluster handling endpoint registrations.

Figure 13-6   Example Intercluster Replication Network for Three Clusters

In Figure 13-6, two LBMs from each cluster have been designated as the LBM hubs for their cluster. These LBM hubs form the intercluster LBM replication network. The bootstrap LBMs configured in each LBM intercluster replication group are designated as SME_1 and SME_2. These two LBM hubs from the SME cluster serve as points of contact or bootstrap LBMs for the entire intercluster LBM replication network. This means that each LBM in each cluster connects to SME_1, replicates its local topology to SME_1, and gets the remote topology from SME_1. They also get the connectivity information for the other leaf clusters from SME_1, connect to the other remote clusters, and replicate their topologies. This creates the full-mesh replication network. If SME_1 is unavailable, the LBM hubs will connect to SME_2. If SME_2 is unavailable, Leaf Cluster 1 LBMs will connect to UCM_A and Leaf Cluster 2 LBMs will connect to UCM_1 as a backup measure in case the SME cluster is unavailable. This is just an example configuration to illustrate the components of the intercluster LBM replication network.

The LBM has the following roles with respect to the LBM intercluster replication network:

- Bootstrap LBMs
  - Remote LBM hubs responsible for interconnecting all LBM hubs in the replication network
  - Can be any hub in the network
  - Can indicate up to three bootstrap LBM hubs per LBM intercluster replication group
- LBM hubs (local LBMs)
  - Communicate directly to other remote hubs as part of the intercluster LBM replication network
- LBM spokes (local LBMs)
  - Communicate directly to local LBM hubs in the cluster and indirectly to the remote LBM hubs through the local LBM hubs
- LBM hub replication network — Bandwidth deduction and adjustment messages
  - LBM optimizes the LBM messages by choosing a sender and receiver from each cluster.
  - The LBM sender and receiver of the cluster is determined by lowest IP address.
  - The LBM hubs that receive messages from remote clusters, in turn forward the received messages to the LBM spokes in their local cluster.

LBM hubs can also be configured to encrypt their communications. This allows intercluster ELCAC to be deployed in environments where it is critical to encrypt traffic between clusters because the links between clusters might reside over unprotected networks. For further information on configuring encrypted signaling between LBM hubs, refer to the Cisco Unified Communications Manager product documentation available at

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

Common Locations (Shared Locations) and Links

As mentioned previously, common locations are locations that are named the same across clusters. Common locations play a key role in how the LBM creates the global topology and how it associates a single location across multiple clusters. A location with the same name between two or more clusters is considered the same location and is thus a shared location across those clusters. So if a location is meant to be shared between multiple clusters, it is required to have exactly the same name. After replication, the LBM will check for configuration discrepancies across locations and links. Any discrepancy in bandwidth value or weight between common locations and links can be seen in serviceability, and the LBM calculates the locations and link paths with the most restrictive values for bandwidth and the lowest value (least cost) for weight.

Common locations and links can be configured across clusters for a number of different reasons. You might have a number of clusters that manage devices in the same physical site and use the same WAN uplinks, and therefore the same location needs to be configured on each cluster in order to associate that location to the local devices on each cluster. You might also have clusters that manage their own topology, yet these topologies interconnect at specific locations and you will have to configure these locations as common locations across each cluster so that, when the global topology is being created, the clusters have the common interconnecting locations and links on each cluster to link each remote topology together effectively. Figure 13-7 illustrates linking topologies together and shows the common topology that each cluster shares.

Figure 13-7   Using Common Locations and Links to Create a Global Topology
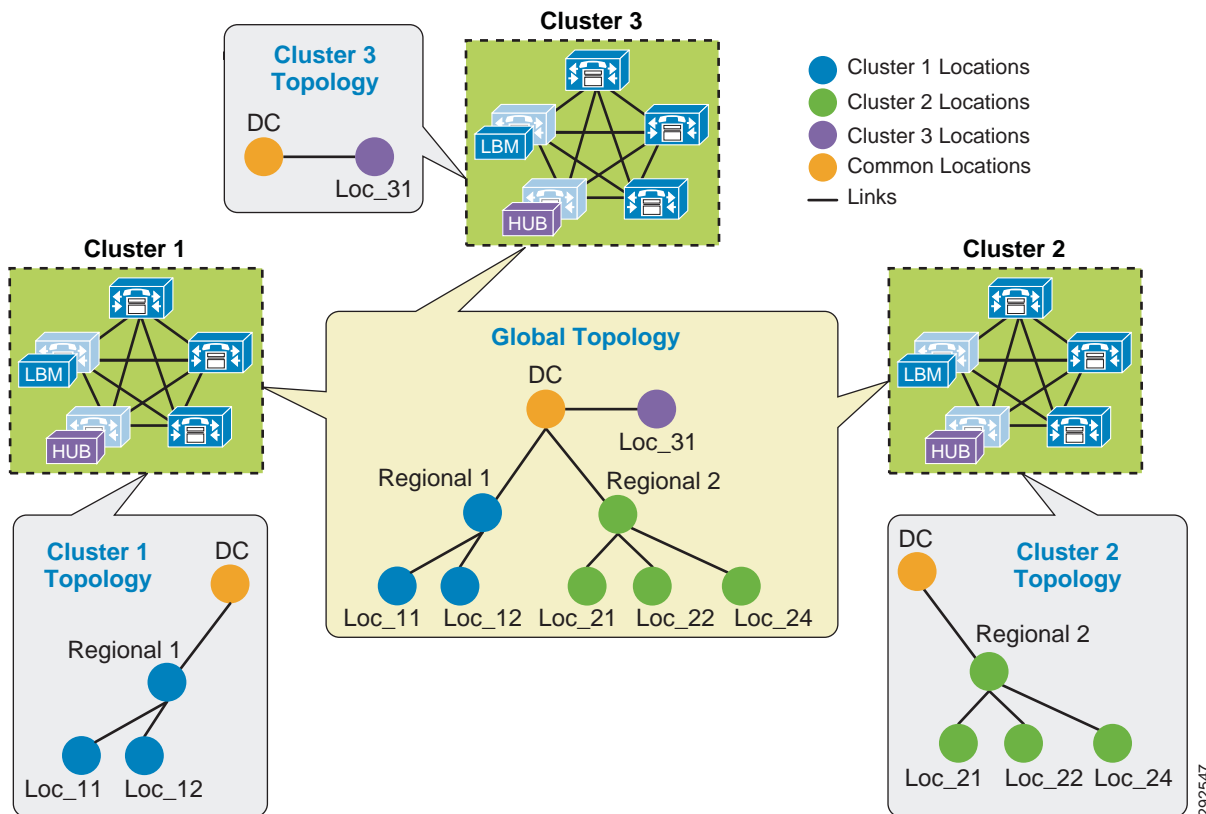
In Figure 13-7, Cluster 1 has devices in locations Regional 1, Loc_11, and Loc_12, but it requires configuring DC and a link from Regional 1 to DC in order to link to the rest of the global topology. Cluster 2 is similar, with devices in Regional 2 and Loc_21, Loc_22, and Loc_23, and it requires configuring DC and a link from DC to Regional 2 to map into the global topology. Cluster 3 has devices in Loc_31 only, and it requires configuring DC and a link to DC from Loc_31 to map into Cluster 1 and Cluster 2 topologies. Alternatively, Regional 1 and Regional 2 could be the common locations configured on all clusters instead of DC, as is illustrated in Figure 13-8.

*Figure 13-8   Alternative Topology Using Different Common Locations*
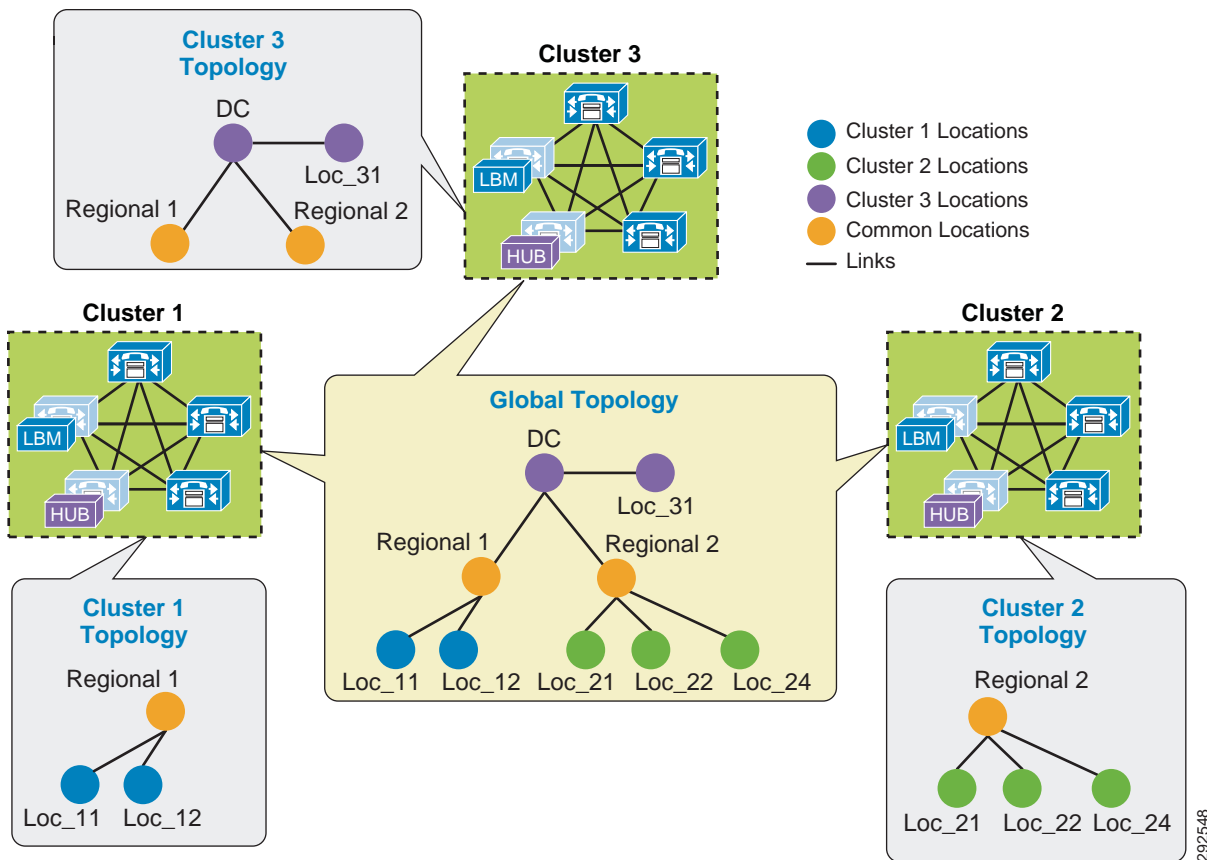
The key to topology mapping from cluster to cluster is to ensure that at least one cluster has a common location with another cluster so that the topologies interconnect accordingly.

Shadow Location

The *shadow location* is used to enable a SIP trunk to pass Enhanced Location CAC information such as location name and Video-Traffic-Class (discussed below), among other things, required for Enhanced Location CAC to function between clusters. In order to pass this location information across clusters, the SIP intercluster trunk (ICT) must be assigned to the "shadow" location. The shadow location cannot have a link to other locations, and therefore no bandwidth can be reserved between the shadow location and other locations. Any device other than a SIP ICT that is assigned to the shadow location will be treated as if it was associated to Hub_None. That is important to know because if a device other than a SIP ICT ends up in the shadow location, bandwidth deductions will be made from that device as if it were in Hub_None, and that could have varying effects depending on the location and links configuration.

When the SIP ICT is enabled for Enhanced Location CAC, it passes information in the SIP Call-Info header that allows the originating and terminating clusters to process the location bandwidth deductions end-to-end. Figure 13-9 illustrates an example of a call between two clusters and some details about the information passed. This is only to illustrate how location information is passed from cluster to cluster and how bandwidth deductions are made.

Figure 13-9   *Location Information Passed Between Clusters over SIP*
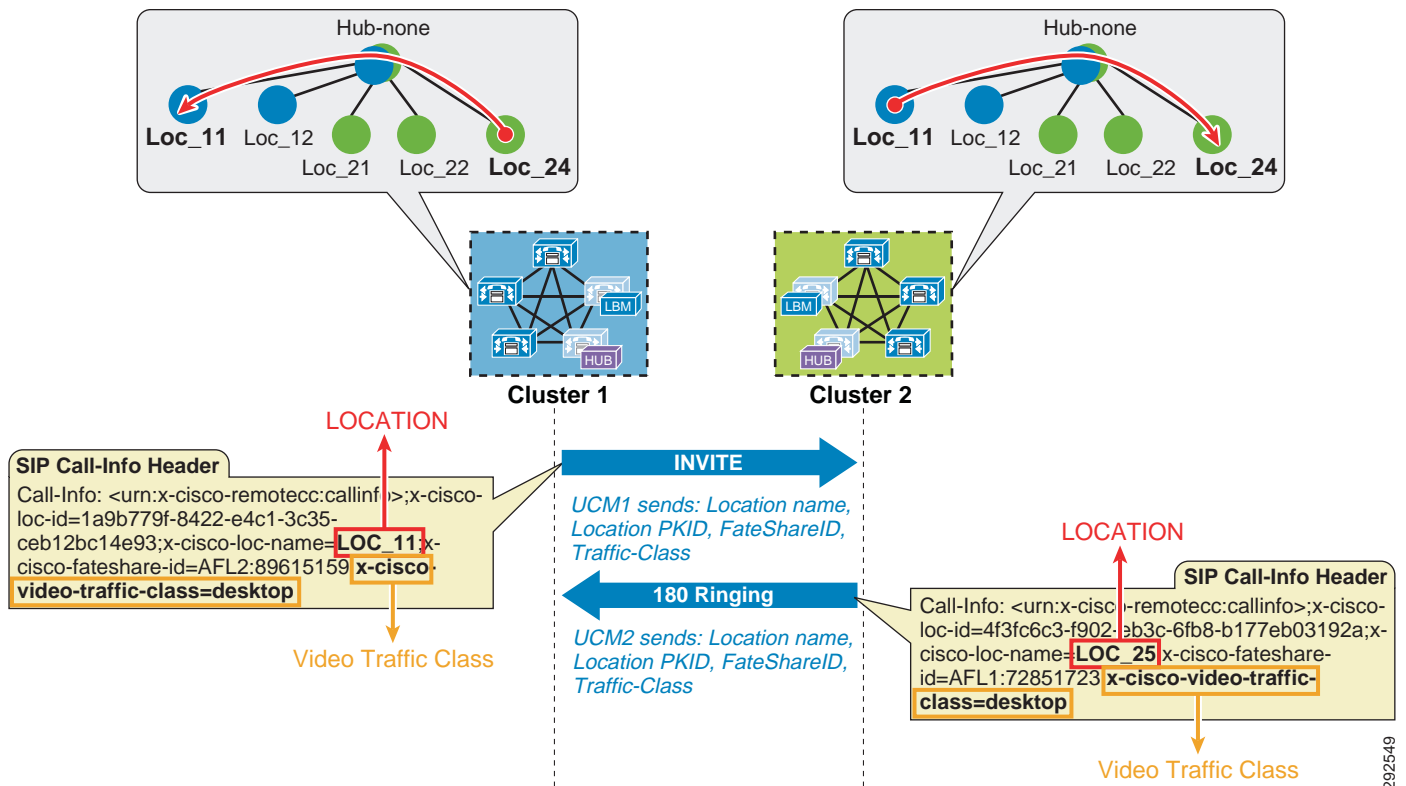
In Figure 13-9, Cluster 1 sends an invite to Cluster 2 and populates the call-info header with the calling parties location name and Video-Traffic-Class, among other pertinent information such as unique call-ID. When Cluster 2 receives the invite with the information, it looks up the terminating party and performs a CAC request on the path between the calling party's and called party's locations from the global topology that it has in memory from LBM replication. If it is successful, Cluster 2 will replicate the reservation and extend the call to the terminating device and return a 180 ringing with the location information of the called party back to Cluster 1. When Cluster 1 receives the 180 ringing, it gets the terminating device's location name and goes through the same bandwidth lookup process using the same unique call-ID that it calculates from the information passed in the call-info header. If it is successful, it too continues with the call flow. Because both clusters use the same information in the call-info header, they will deduct bandwidth for the same call using the same call-ID, thus avoiding any double bandwidth deductions.

Location and Link Management Cluster

In order to avoid configuration overhead, a Location and Link Management Cluster can be configured to manage all locations and links in the global topology. All other clusters uniquely configure the locations that they require for location-to-device association and do not configure links or any bandwidth values other than unlimited. It should be noted that the Location and Link Management Cluster is a design concept and is simply any cluster that is configured with the entire global topology of locations and links, while all other clusters in the LBM replication network are configured only with locations set to unlimited bandwidth values and without configured links. When intercluster Enhanced Location CAC is enabled and the LBM replication network is configured, all clusters replicate their view of the network. The designated Location and Link Management Cluster has the entire global topology with

locations, links, and bandwidth values; and once those values are replicated, all clusters use those values because they are the most restrictive. This design alleviates configuration overhead in deployments where a large number of common locations are required across multiple clusters.

Recommendations

- Locations and link management cluster:
  - One cluster should be chosen as the management cluster (the cluster chosen to manage locations and links).
  - The management cluster should be configured with the following:

    All locations within the enterprise will be configured in this cluster.

    All bandwidth values and weights for all locations and links will be managed in this cluster.
- All other clusters in the enterprise:
  - All other clusters in the enterprise should configure *only* the locations required for association to devices but should *not* configure the links between locations. This link information will come from the management cluster when intercluster Enhanced Location CAC is enabled.
  - When intercluster Enhanced Location CAC is enabled, all of the locations and links will be replicated from the management cluster and LBM will use the lowest, most restrictive bandwidth and weight value.
- LBM will always use the lowest most restrictive bandwidth and lowest weight value after replication.

Benefits

- Manage enterprise CAC topology from a single cluster.
- Alleviates location and link configuration overhead when clusters share a large number of common locations.
- Alleviates configuration mistakes in locations and links across clusters.
- Other clusters in the enterprise require the configuration only of locations needed for location-to-device and endpoint association.
- Provides a single cluster for monitoring of the global locations topology.

Figure 13-10 illustrates Cisco Unified Communications Manager Session Management Edition (SME) as a Location and Link Management Cluster for three leaf clusters.

Figure 13-10   Example of SME as a Location and Link Management Cluster
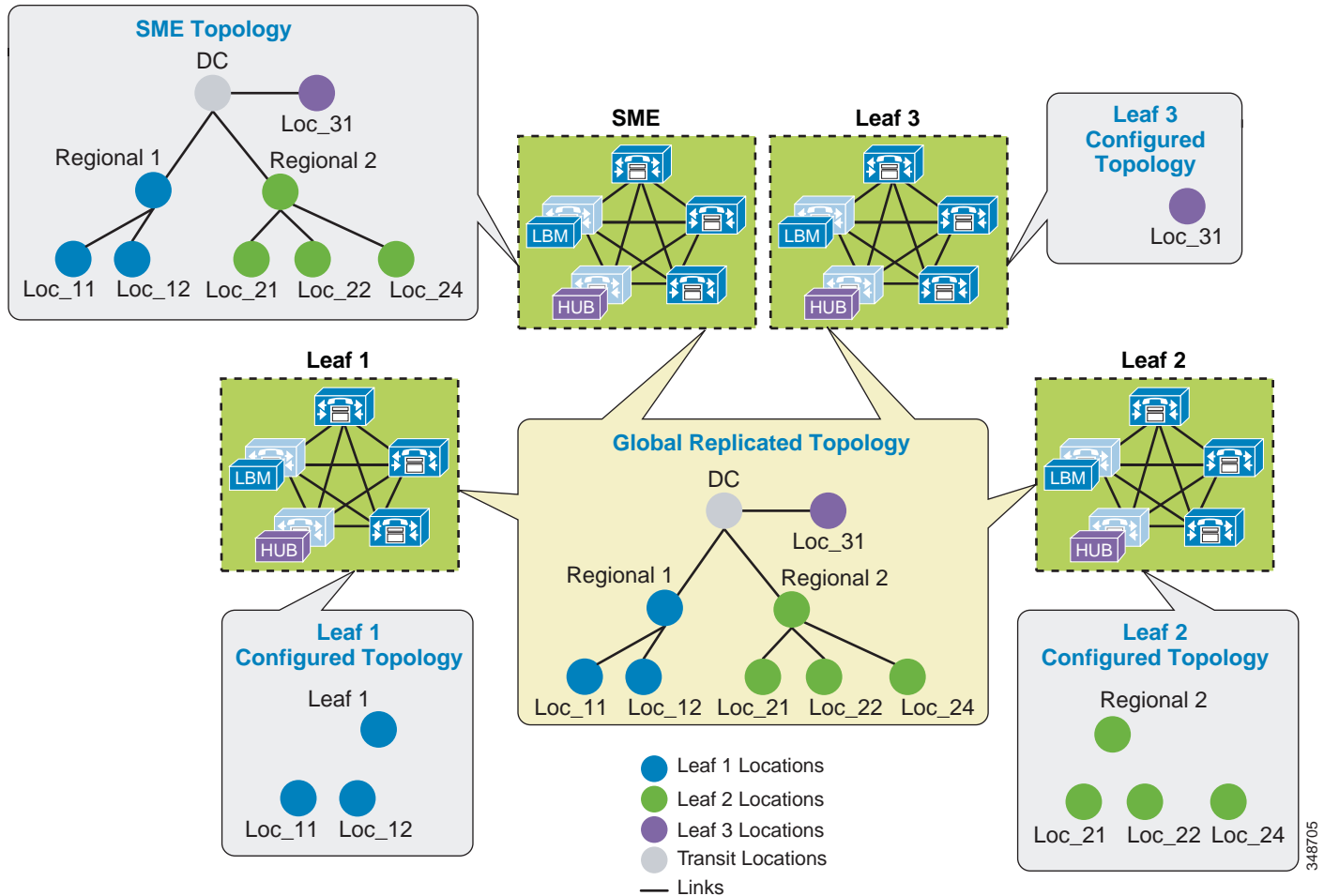
In Figure 13-10 there are three leaf clusters, each with devices in only a regional and remote locations. SME has the entire global topology configured with locations and links, and intercluster LBM replication is enabled between all four clusters. None of the clusters in this example share locations, although all of the locations are common locations because SME has configured the entire location and link topology. Note that Leaf 1, Leaf 2, and Leaf 3 configure only locations that they require to associate to devices and endpoints, while SME has the entire global topology configured. After intercluster replication, all clusters will have the global topology.

Intercluster Enhanced Location CAC Design and Deployment Recommendations and Considerations

- A cluster requires the location to be configured locally for location-to-device association.
- Each cluster should be configured with the immediately neighboring locations so that each cluster's topology can inter-connect. This does not apply to Location and Link Management Cluster deployments.
- Links need to be configured to establish points of interconnect between remote topologies. This does not apply to Location and Link Management Cluster deployments.

- Discrepancies of bandwidth limits and weights on common locations and links are resolved by using the lowest bandwidth and weight values.

- Naming locations consistently across clusters is critical. Follow the practice, "Same location, same name; different location, different name."

- The Hub_None location should be renamed to be unique in each cluster or else it will be a common (shared) location by other clusters.

- Cluster-ID should be unique on each cluster for serviceability reports to be usable.

- All LBM hubs are fully meshed between clusters.

- An LBM hub is responsible for communicating to hubs in remote clusters.

- An LBM spoke does not directly communicate with other remote clusters. LBM spokes receive and send messages to remote clusters through the Local LBM Hub.

- LBM Hub Groups

    - Used to assign LBMs to the Hub role

    - Used to define three remote hub members that replicate hub contact information for all of the hubs in the LBM hub replication network

    - An LBM is a hub when it is assigned to an LBM hub group.

    - An LBM is a spoke when it is not assigned to an LBM hub group.

- If a cluster has no LBM hub, or if the LBM hub is not running, the cluster will be isolated and will not participate in the intercluster LBM replication network.

Performance Guidelines

- Maximum of 2,000 locally configured locations. This limit of 2,000 locations also applies to the Location and Link Management Cluster.

- Maximum of 8,000 total replicated locations with intercluster CAC

Enhanced Location CAC for TelePresence Immersive Video

Since TelePresence endpoints now provide a diverse range of collaborative experiences from the desktop to the conference room, Enhanced Location CAC includes support to provide CAC for TelePresence immersive video calls. This section discusses the features in Enhanced Location CAC that support TelePresence immersive video CAC.

Video Call Traffic Class

Video Call Traffic Class is an attribute that is assigned to all endpoints, and that can also be enabled on SIP trunks, to determine the video classification type of the endpoint or trunk. This enables Unified CM to classify various call flows as either immersive, desktop video, or both, and to deduct accordingly from the appropriate location and/or link bandwidth allocations of video bandwidth, immersive bandwidth, or both. For TelePresence endpoints there is a non-configurable Video Call Traffic Class of **immersive** assigned to the endpoint. A SIP trunk can be classified as desktop, immersive, or mixed video in order to deduct bandwidth reservations of a SIP trunk call. All other endpoints and trunks have a non-configurable Video Call Traffic Class of **desktop video**. More detail on endpoint and trunk classification is provided in the subsections below.

TelePresence immersive endpoints mark their media with a DSCP value of CS4 by default, and desktop video endpoints mark their media with AF41 by default, as per recommended QoS settings. For Cisco endpoints this is accomplished through the configurable Unified CM QoS service parameters **DSCP for Video calls** and **DSCP for TelePresence calls**. When a Cisco TelePresence endpoint registers with Unified CM, it downloads a configuration file and applies the QoS setting of **DSCP for TelePresence calls**. When a Unified Communications video-capable endpoint registers with Unified CM, it downloads a configuration file and applies the QoS setting of **DSCP for Video calls**. All third-party video endpoints require manual configuration of the endpoints themselves and are statically configured, meaning they do not change QoS marking depending on the call type; therefore, it is important to match the Enhanced Location CAC bandwidth allocation to the correct DSCP. Unified CM achieves this by deducting desktop video calls from the Video Bandwidth location and link allocation for devices that have a Video Call Traffic Class of **desktop**. End-to-end TelePresence immersive video calls are deducted from the Immersive Video Bandwidth location and link allocation for devices or trunks with the Video Call Traffic Class of **immersive**. This ensures that end-to-end desktop video and immersive video calls are marked correctly and counted correctly for call admission control. For calls between desktop devices and TelePresence immersive devices, bandwidth is deducted from both the Video Bandwidth and the Immersive Video Bandwidth location and link allocations.

Endpoint Classification

Cisco TelePresence endpoints have a fixed non-configurable Video Call Traffic Class of **immersive** and are identified by Unified CM as immersive. Telepresence endpoints are defined in Unified CM by the device type. When a device is added in Unified CM, any device with TelePresence in the name of the device type is classified as **immersive**, as are the generic single-screen and multi-screen room systems. Another way to check the capabilities of the endpoints in the Unified CM is to go to the **Cisco Unified Reporting Tool > System Reports > Unified CM Phone Feature List**. In the feature drop down list, select **Immersive Video Support for TelePresence Devices**; in the product drop down list, select **All**. This will display all of the device types that are classified as **immersive**. All other endpoints have a fixed Video Call Traffic Class of **desktop** due to their lack of the non-configurable immersive attribute.

Bandwidth reservations are determined by the classification of endpoints in a video call, and they deduct bandwidth from the locations and links bandwidth pools as listed in Table 13-4.

*Table 13-4   Bandwidth Pool Usage per Endpoint Type*

Endpoint A	Endpoint B	Locations and Links Pool Used
Immersive video	Immersive video	Immersive bandwidth
Immersive video	Desktop video	Immersive and video bandwidth
Desktop video	Desktop video	Video bandwidth
Audio-only call	Any	Audio bandwidth

SIP Trunk Classification

A SIP trunk can also be classified as desktop, immersive, or mixed video in order to deduct bandwidth reservations of a SIP trunk call, and the classification is determined by the calling device type and Video Call Traffic Class of the SIP trunk. The SIP trunk can be configured through the SIP Profile trunk-specific information as:

- Immersive — High-definition immersive video
- Desktop — Standard desktop video
- Mixed — A mix of immersive and desktop video

A SIP trunk can be classified with any of these three classifications and is used primarily to classify Video or TelePresence Multipoint Control Units (MCUs), a video device at a fixed location, a Unified CM cluster supporting traditional Location CAC, or a Cisco TelePresence System Video Communications Server (VCS).

Bandwidth reservations are determined by the classification of an endpoint and a SIP trunk in a video call, and they deduct bandwidth from the locations and links bandwidth pools as listed in Table 13-5.

Table 13-5   Bandwidth Pool Usage per SIP Trunk and Endpoint Type

Endpoint	SIP Trunk	Locations and Links Pool Used
TelePresence endpoint	Immersive	Immersive bandwidth
TelePresence endpoint	Desktop	Immersive and video bandwidth
TelePresence endpoint	Mixed	Immersive and video bandwidth
Desktop endpoint	Immersive	Immersive and video bandwidth
Desktop endpoint	Desktop	Video bandwidth
Desktop endpoint	Mixed	Immersive and video bandwidth
Non-video endpoint	Any	Audio bandwidth

By default, all video calls from either immersive or desktop endpoints are deducted from the locations and links video bandwidth pool. To change this behavior, set Unified CM's CallManager service parameter **Use Video BandwidthPool for Immersive Video Calls** to **False**, and this will enable the immersive video bandwidth deductions. After this is enabled, immersive and desktop video calls will be deducted out of their respective pools.

As described earlier, a video call between a Unified Communications video endpoint (desktop Video Call Traffic Class) and a TelePresence endpoint (immersive Video Call Traffic Class) will mark their media asymmetrically and, when immersive video CAC is enabled, will deduct bandwidth from both video and immersive locations and links bandwidth pools. Figure 13-11 illustrates this.

Figure 13-11   Enhanced Location CAC Bandwidth Deductions and Media Marking for a Multi-Site Deployment
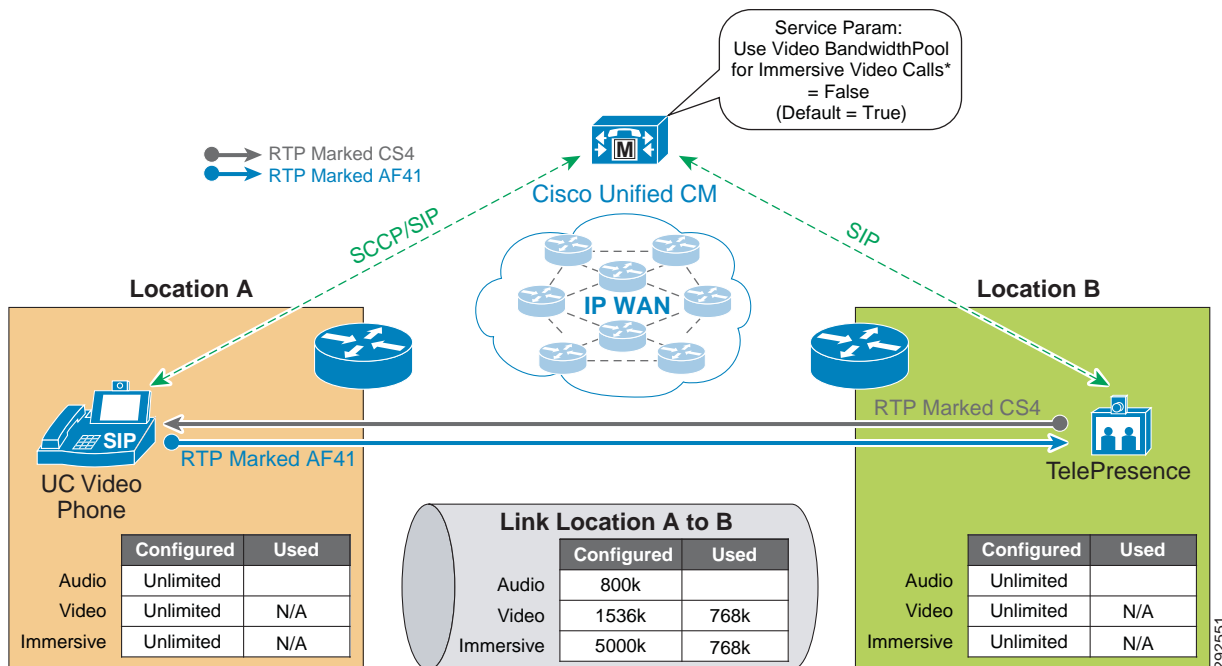
Examples of Various Call Flows and Location and Link Bandwidth Pool Deductions

The following call flows depict the expected behavior of locations and links bandwidth deductions when the Unified CM service parameter **Use Video BandwidthPool for Immersive Video Calls** is set to **False**.

Figure 13-12 illustrates an end-to-end TelePresence immersive video call between TP-A in Location L1 and TP-B in Location L2. End-to-end immersive video endpoint calls deduct bandwidth from the immersive bandwidth pool of the locations and the links along the effective path.

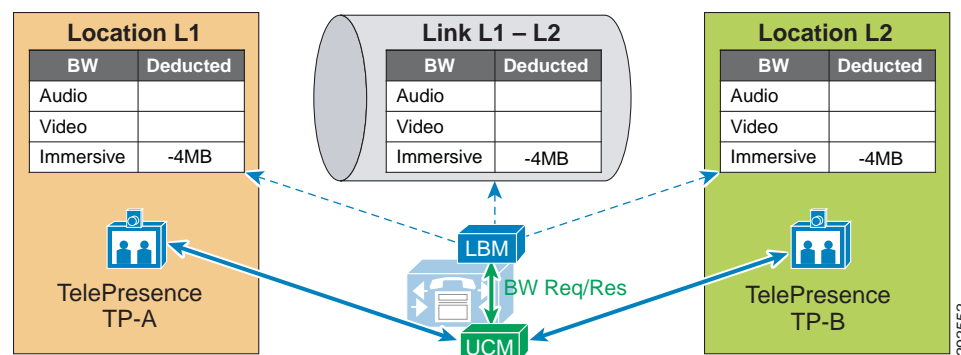*Figure 13-12   Call Flow for End-to-End TelePresence Immersive Video*

Figure 13-13 illustrates an end-to-end desktop video call between DP-A in Location L1 and DP-B in Location L2. End-to-end desktop video endpoint calls deduct bandwidth from the video bandwidth pool of the locations and the links along the effective path.

Figure 13-13   Call Flow for End-to-End Desktop Video

Figure 13-14 illustrates a video call between desktop video endpoint DP-A in Location L1 and TelePresence video endpoint TP-B in Location L2. Interoperating calls between desktop video endpoints and TelePresence video endpoints deduct bandwidth from both video and immersive locations and the links bandwidth pools along the effective path.
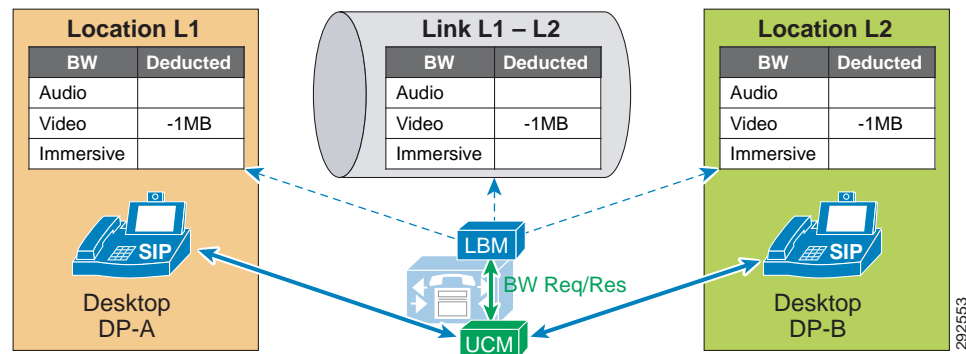
Figure 13-14   Call Flow for Desktop-to-TelePresence Video
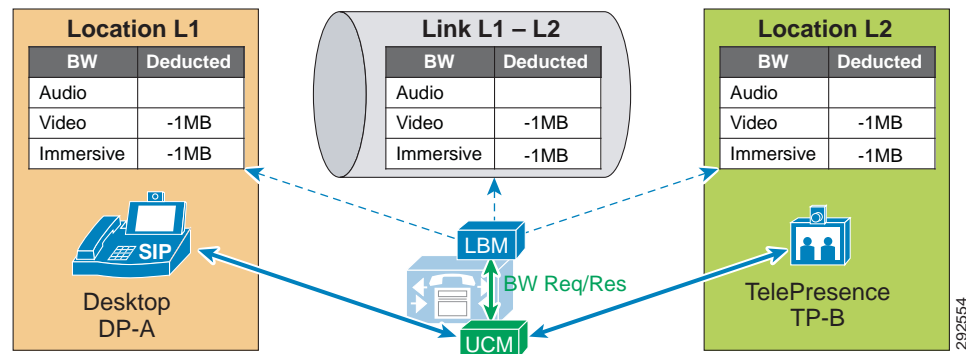
In Figure 13-15, a desktop video endpoint and two TelePresence endpoints call a SIP trunk configured with a Video Traffic Class of **immersive** that points to a TelePresence MCU. Bandwidth is deducted along the effective path from the immersive locations and the links bandwidth pools for the calls that are end-to-end immersive and from both video and immersive locations and the links bandwidth pools for the call that is desktop-to-immersive.

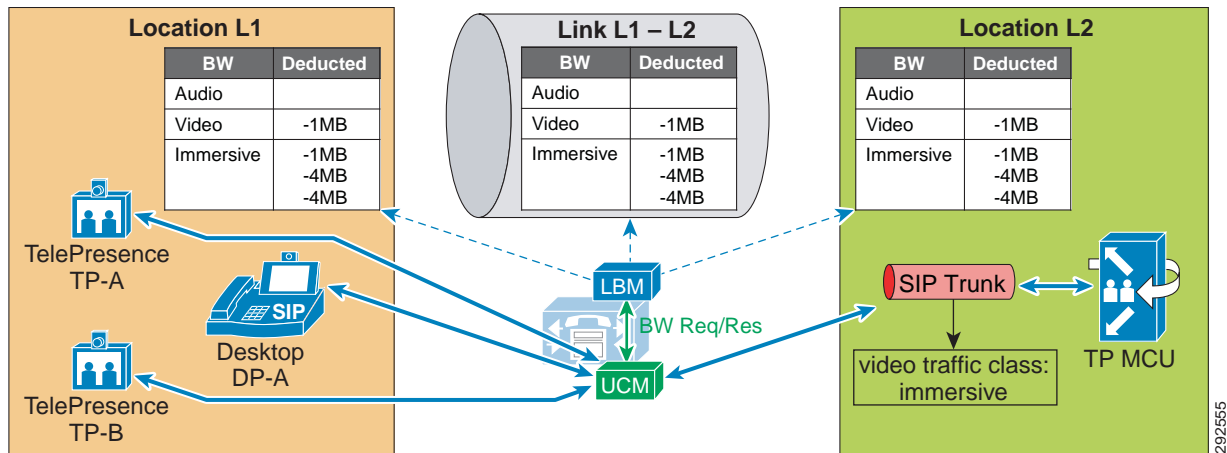*Figure 13-15   Call Flow for a Video Conference with an MCU*

Figure 13-16 illustrates an end-to-end immersive video call across clusters, which deducts bandwidth from the immersive bandwidth pool of the locations and links along the effective path.

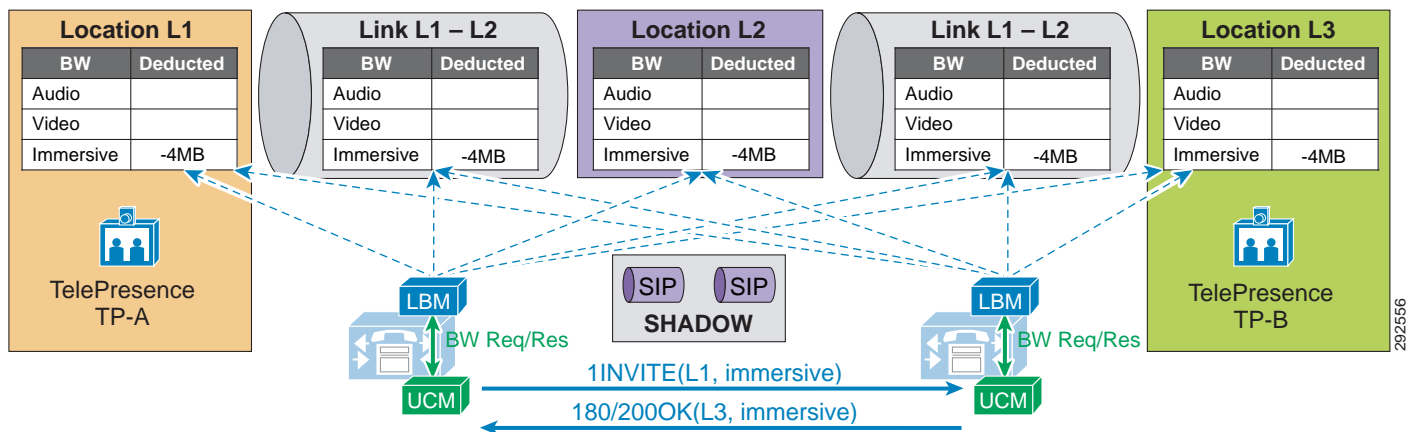*Figure 13-16   Call Flow for End-to-End TelePresence Immersive Video Across Clusters*

Figure 13-17 illustrates an end-to-end desktop video call across clusters, which deducts bandwidth from the video bandwidth pool of the locations and links along the effective path.

Figure 13-17 **Call Flow for End-to-End Desktop Video Call Across Clusters**

Figure 13-18 illustrates a desktop video endpoint calling a TelePresence endpoint across clusters. the call deducts bandwidth from both video and immersive bandwidth pools of the locations and links along the effective path.

Figure 13-18 **Call Flow for Desktop-to-TelePresence Video Across Clusters**
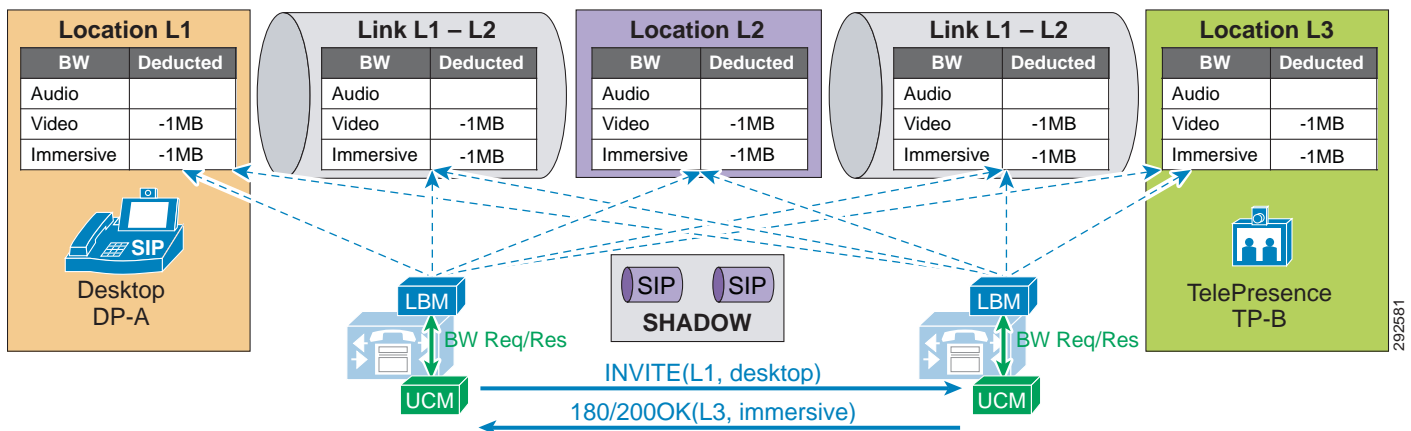
Video Bandwidth Utilization and Admission Control

When Unified CM negotiates an audio or video call, a number of separate streams are established between the endpoints involved in the call. For video calls with content sharing, this can result in as many as 8 (or possibly more) unidirectional streams. For an audio-only call typically the bare minimum is 2 streams, one in each direction. This section discusses bandwidth utilization on the network and how Unified CM accounts for this in admission control bandwidth accounting.

For the purpose of the discussion in this section, please note the following:

- The figures in this section use a bidirectional arrow (<-->) to represent two unidirectional streams.

- The following points summarize how Unified CM Enhanced Location CAC deducts bandwidth from the configured audio, video, and immersive allocations. For more information, see the section on :

    - Audio (audio-only calls): RTP bit rate + IP and UDP header overhead

    - Video (video calls): RTP bit rate only

    - Immersive (video calls by Cisco TelePresence endpoints): RTP bit rate only

- Bandwidth deductions in Enhanced Location CAC:

    - Bandwidth deductions are made for bidirectional RTP streams and are assumed to be symmetrically routed (both streams routed over the same path). For example, a G.711 audio call of 80 kbps is 80 kbps in each direction over a full-duplex network; that is 80 kbps on the transmit pair of wires and 80 kbps on the receive pair of wires, equating to 80 kbps full-duplex. (See Figure 13-19.) Note that traffic is not always routed symmetrically in the WAN. Check with your network administrator when necessary to ensure that admission control is correctly accounting for the media as it is routed in the network over the WAN.

    - Real-Time Transport Control Protocol (RTCP) bandwidth overhead is not part of Unified CM bandwidth allocation and should be part of network provisioning. RTCP is quite common in most call flows and is commonly used for statistical information about the streams. It is also used to synchronize audio in video calls to ensure proper lip-sync. In some cases it can be enabled or disabled on the endpoint. RFC 3550 recommends that the fraction of the session bandwidth added for RTCP should be fixed at 5%. What this means is that it is common practice for the RTCP session to be up to 5% of the associated RTP session. So when calculating bandwidth consumption on the network, you should add the RTCP overhead for each RTP session. For example, if you have a G.711 audio call of 80 kbps with RTCP enabled, you will be using up to 84 kbps per session (4 kbps RTCP overhead) for both RTP and RTCP. This calculation is not part of Enhanced Location CAC deductions but should be part of network provisioning.

✎ **Note**   There are, however, methods to re-mark this traffic to another Differentiated Services Code Point (DSCP). For example, RTCP uses odd-numbered UDP ports while RTP uses even-numbered UDP ports. Therefore, classification based on UDP port ranges is possible. Network-Based Application Recognition (NBAR) is another option as it allows for classification and re-marking based on the RTP header **Payload Type** field. For more information on NBAR, see http://www.cisco.com. However, if the endpoint marking is trusted in the network, then RTCP overhead should be provisioned in the network within the same QoS class as audio RTP (default marking is EF). It should also be noted that RTCP is marked by the endpoint with the same marking as RTP; by default this is EF (verify that RTCP is also marked as EF).

Figure 13-19   A Basic Audio-Only Call with RTCP Enabled
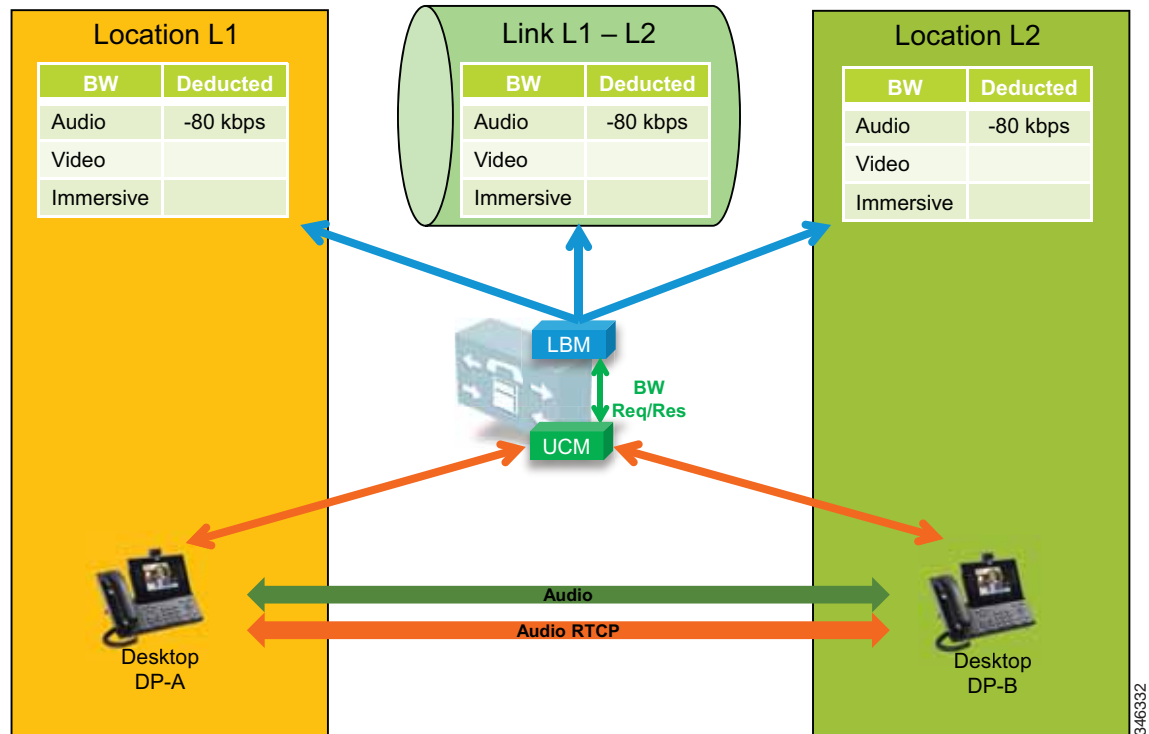
In Figure 13-19 two desktop video phones have established an audio-only call. In this call flow four streams are negotiated: two audio streams illustrated by a single bidirectional arrow and two RTCP streams also illustrated by a bidirectional arrow. For this call, the Location Bandwidth Manager (LBM) deducts 80 kbps (bit rate + IP/UDP overhead) between location L1 and location L2 for a call established between desktop phones DP-A and DP-B. The actual bandwidth consumed at Layer 3 in the network with RTCP enabled would be between 80 kbps and 84 kbps, as discussed previously in this section.

In Figure 13-20 two desktop video phones have established a video call. In this call flow eight streams are negotiated: two audio streams, two audio-associated RTCP streams, two video streams, and two video-associated RTCP streams. Again for this illustration one bidirectional arrow is used to depict two unidirectional streams. This particular call is 1024 kbps, with 64 kbps of G.711 audio and 960 kbps of video (bit rate only for audio and video allocations of video calls). So in this case the LBM deducts 1024 kbps between locations L1 and L2 for a video call established between desktop phones DP-A and DP-B. RTCP is overhead that should be accounted for in provisioning, depending on how it is marked or re-marked by the network.

Figure 13-20   A Basic Video Call with RTCP Enabled
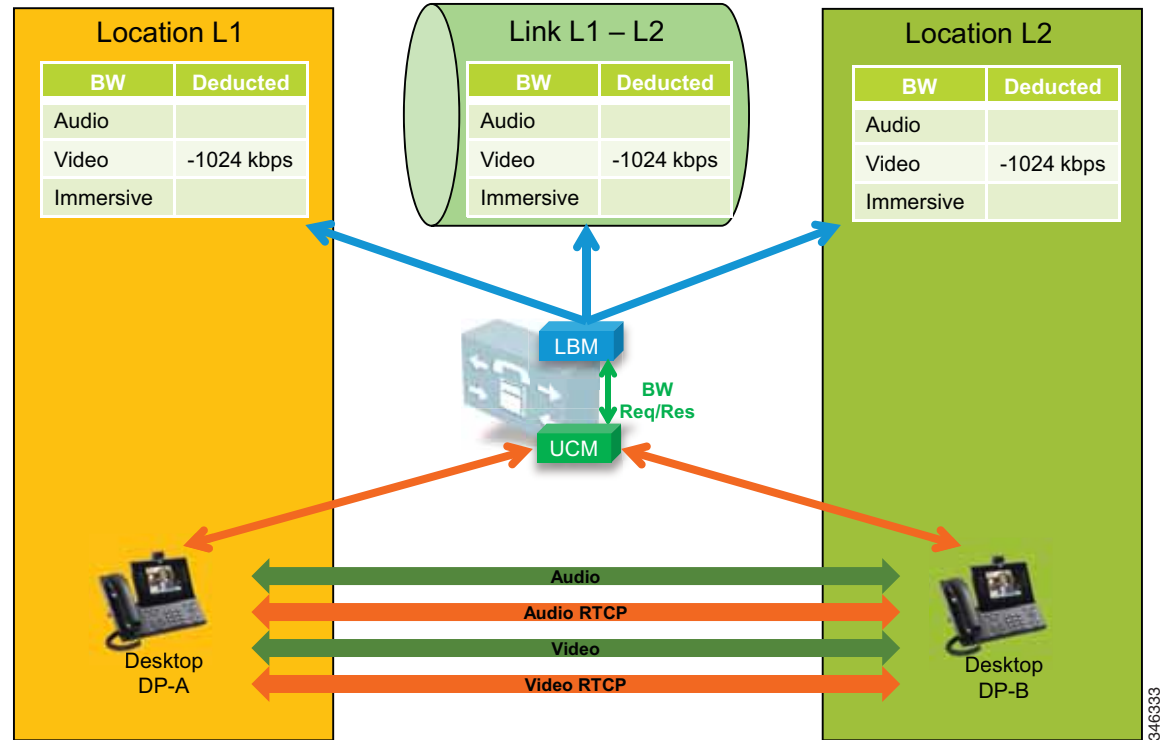
The example in Figure 13-21 is of a video call with presentation sharing. This is a more complex call with regard to the number of associated streams and bandwidth allocation when compared to bandwidth used on the network, and therefore it must be provisioned in the network. Figure 13-21 illustrates a video call with RTCP enabled and Binary Floor Control Protocol (BFCP) enabled for presentation sharing. All SIP-enabled telepresence multipurpose or personal endpoints such as a the Cisco TelePresence System EX, MX, SX, C, and Profile Series function in the same manner.

Figure 13-21   Video Call with RTCP and BFCP Enabled and Presentation Sharing
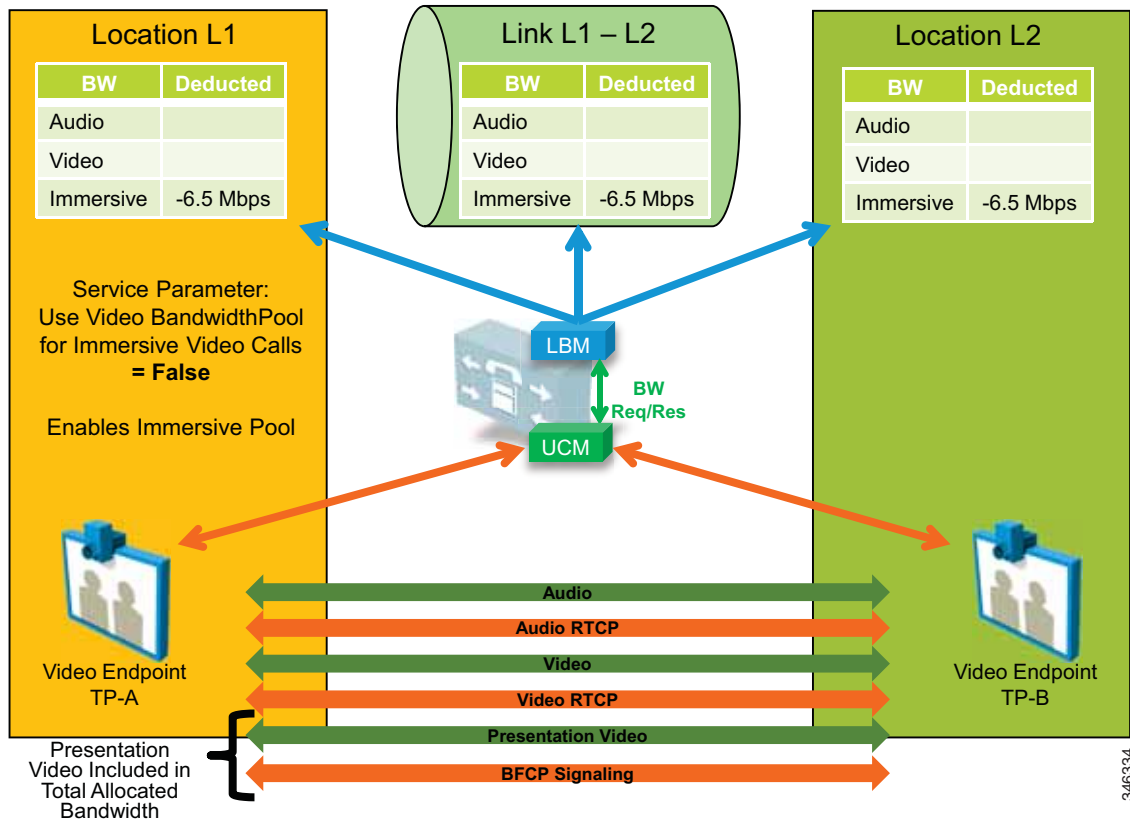

When a video call is established between two video endpoints, audio and video streams are established and bandwidth is deducted for the negotiated rate. Unified CM uses regions to determine the maximum bit rate for the call. For example, with a Cisco TelePresence System EX90 at the highest detail of 1080p at 30 frames per second (fps), the negotiated rate between regions would have to be set at 6.5 Mbps. EX90s used in this scenario would average around 6.1 Mbps for this session. When the endpoints start presentation sharing during the session, BFCP is negotiated between the endpoints and a new video stream is enabled at either 5 fps or 30 fps, depending on endpoint configuration. When this occurs, the endpoints will throttle down their main video stream to include the presentation video so that the entire session does not use more than the allocated bandwidth of 6.5 Mbps. Thus, the average bandwidth consumption remains the same with or without presentation sharing.

Note   The presentation video stream is typically unidirectional in the direction of the person or persons viewing the presentation.

Telepresence immersive and office endpoints such as the Cisco TelePresence System 500, 1000, 3000, and TX9000 Series that negotiate a call between one another function a little differently in the sense that the video for presentation sharing is an additional bandwidth above and beyond what is allocated for the main video session, and thus it is not deducted from Enhanced Location CAC. Figure 13-22 illustrates this.

Figure 13-22   Video Call with RTCP and BFCP Enabled and Presentation
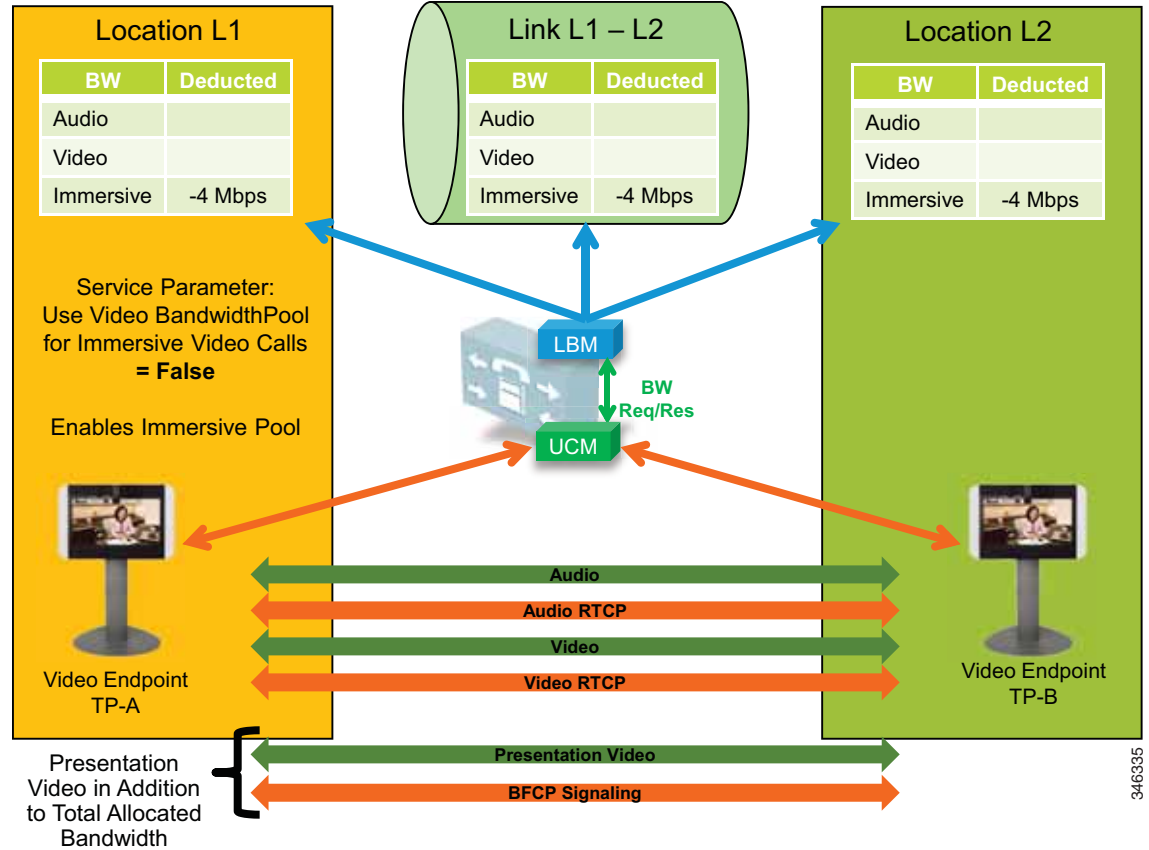
In Figure 13-22 the telepresence immersive video endpoints establish a video call and enable presentation sharing. The LBM deducts 4 Mbps for the main audio and video session from the immersive pool for the call, and video is established between the endpoints. When presentation sharing is activated, the two endpoints exchange BFCP and negotiate a presentation video stream at 5 fps or 30 fps in one direction, depending on the endpoint configuration. At 5 fps the average bandwidth used is approximately 500 kbps of additional bandwidth overhead. This bandwidth is above and beyond the 4 Mbps that was allocated for the video call and should be provisioned in the network. At 30 fps the average bit rate of the presentation video is approximately 1.5 Mbps.

Note   The Cisco TelePresence System endpoints use Telepresence Interoperability Protocol (TIP) to multiplex multiple screens and audio into two audio and video RTP streams in each direction. Therefore the actual streams on the wire may be different than what is expressed in the illustration, but the concept of additional bandwidth overhead for the presentation video is the same.

Upgrade and Migration from Location CAC to Enhanced Location CAC

Upgrading to Cisco Unified CM from a previous release that supports only traditional Location CAC, will result in the migration of Location CAC to Enhanced Location CAC. The migration consists of taking all previously defined locations bandwidth limits of audio and video bandwidth and migrating them to a link between the user-defined location and Hub_None. This effectively recreates the hub-and-spoke model that previous versions of Unified CM Location CAC supported. Figure 13-23 illustrates the migration of bandwidth information.

*Figure 13-23   Migration from Location CAC to Enhanced Location CAC After Unified CM Upgrade*
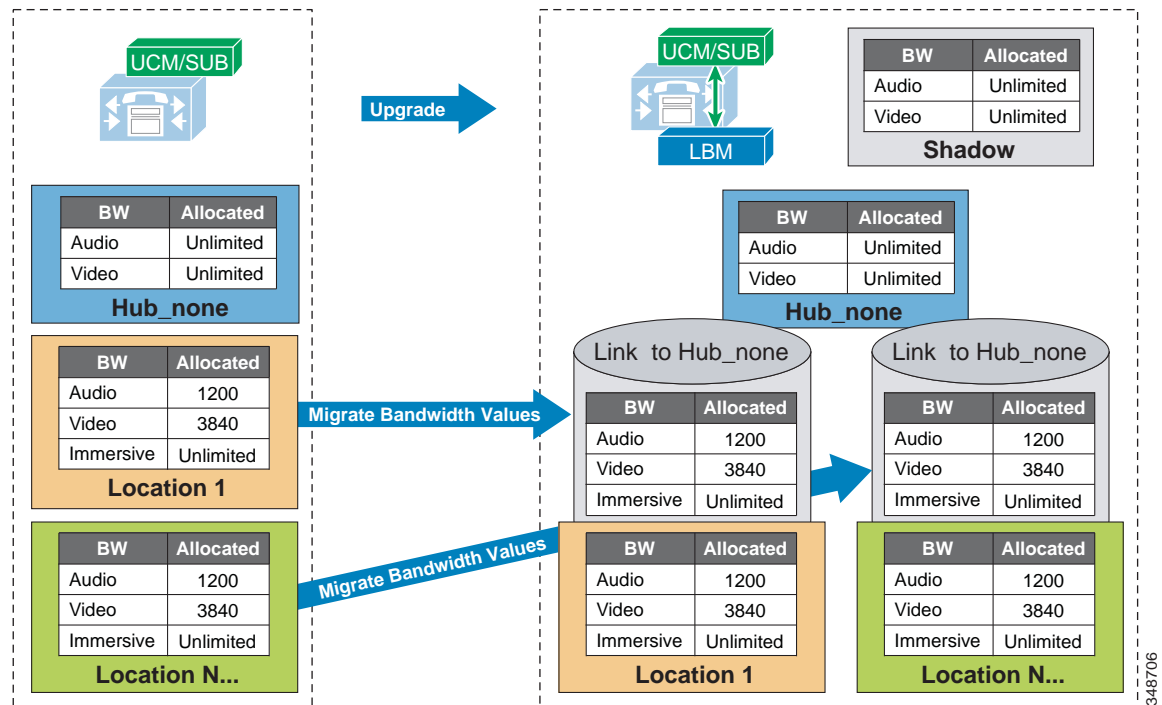
Settings after an upgrade to a Cisco Unified CM release that supports Enhanced Location CAC:

- The LBM is activated on each Unified CM subscriber running the Cisco CallManager service.
- The Cisco CallManager service communicates directly with the local LBM.
- No LBM group or LBM hub group is created.
- All LBM services are fully meshed.
- Intercluster Enhanced Location CAC is not enabled.
- All intra-location bandwidth values are set to unlimited.
- Bandwidth values assigned to locations are migrated to a link connecting the user-defined location and Hub_None.
- Immersive bandwidth is set to unlimited.
- A shadow location is created.

- Phantom and shadow locations have no links.

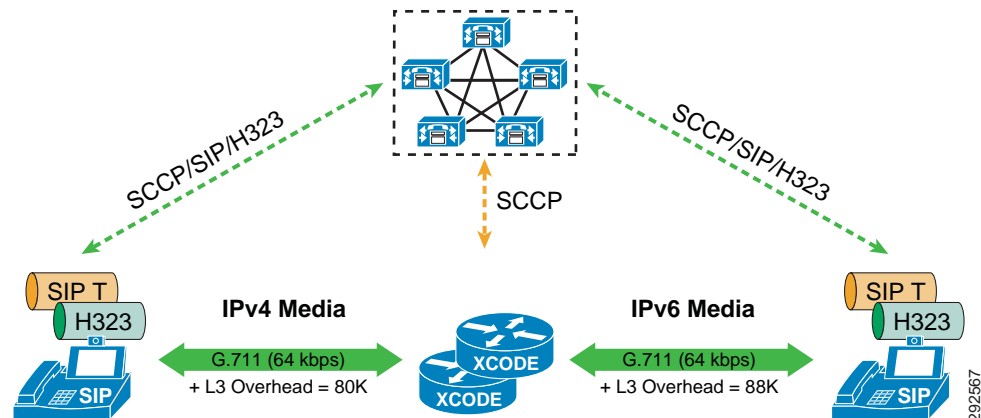- Enhanced Location CAC bandwidth adjustment for MTPs and transcoders:

    For transcoding insertion, the bit rate is different on each leg of the connection. Figure 13-24 illustrates this.

Figure 13-24   Example of Different Bit Rate for Transcoding
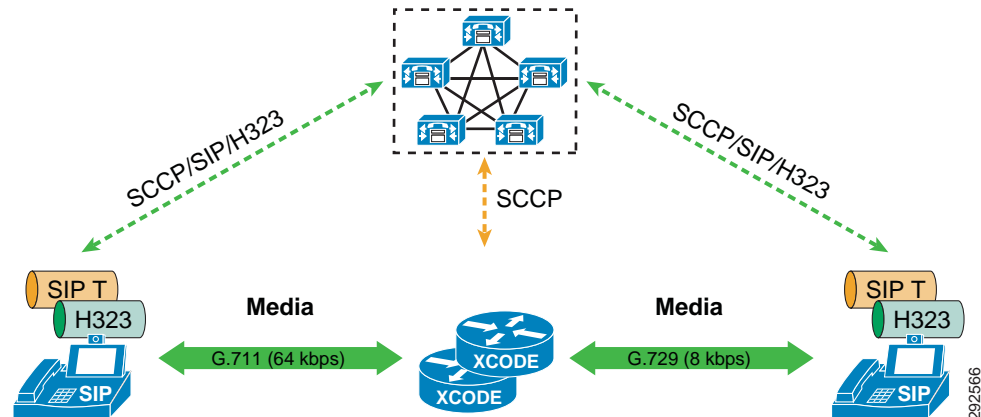
For dual stack MTP insertion, the bit rate is different on each connection but the bandwidth is different due to IP header overhead. Figure 13-25 illustrates the difference in bandwidth used for IPv4 and IPv6 networks with dual stack MTP insertion.

Figure 13-25   Bandwidth Differences for Dual Stack MTP Insertion

Enhanced Location CAC does not account for these differences in bandwidth between MTPs and transcoders. The service parameter **Locations Media Resource Audio Bit Rate Policy** determines whether the largest or smallest bandwidths should be used along the locations and links path. Lowest Bit Rate (default) or Highest Bit Rate can be used to manage these differences in bandwidth consumption.

Extension Mobility Cross Cluster with Enhanced Location CAC

Enhanced Location CAC supports designs using Extension Mobility Cross Cluster (EMCC). Unified CM provides the ability to perform Extension Mobility logins between clusters within an enterprise with a feature called Extension Mobility Cross Cluster (EMCC). For further information, see the section on .

With Enhanced Location CAC in EMCC designs, the visiting cluster passes the location of the visiting phone to the home cluster. This allows the home cluster to associate the correct location to the visiting phone during registration. The following requirements must be met for Enhanced Location CAC to function in EMCC designs:

- Cisco Unified CM 10.0 or a later release required on both home and visiting clusters
- The visiting and home clusters must be in the same intercluster LBM replication network

Both Enhanced Location CAC and EMCC can be designed and deployed according to the guidelines in the product documentation and this SRND. There are no other requirements or any specific configuration aspects to employ.

Design Considerations for Call Admission Control

This section describes how to apply the call admission control mechanisms to various IP WAN topologies. With Unified CM Enhanced Location CAC network modeling support, Unified CM is no longer limited to supporting simple hub-and-spoke or MPLS topologies but, together with intercluster enhanced locations, can now support most any network topology in any Unified CM deployment model. Enhanced Location CAC is still a statically defined mechanism that does not query the network, and therefore the administrator still needs to provision Unified CM accordingly whenever network changes affect admission control. This is where a network-aware mechanism such as RSVP can fill that gap and provide support for dynamic changes in the network, such as when network failures occur and media streams take different paths in the network. This is often the case in designs with load-balanced dual or multi-homed WAN uplinks or unequally sized primary and backup WAN uplinks.

To learn how Enhanced Location CAC functions and how to design and deploy Enhanced location CAC, see the section on .

In this section explores a few typical topologies and explains how Enhanced Location CAC can be designed to manage them.
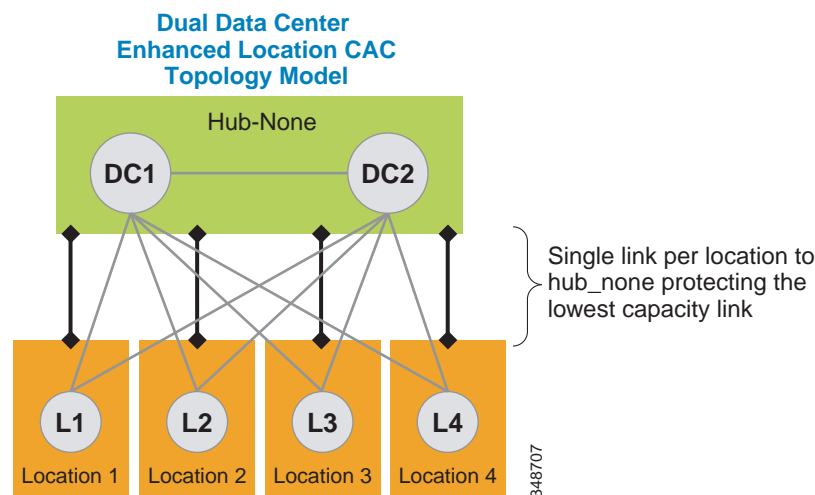
Dual Data Center Design

Figure 13-26 illustrates a simple dual data center WAN network design where each remote site has a single WAN uplink to each data center. The data centers are interconnected by a high-speed WAN connection that is over-provisioned for data traffic.

Figure 13-26   Dual Data Center WAN Network
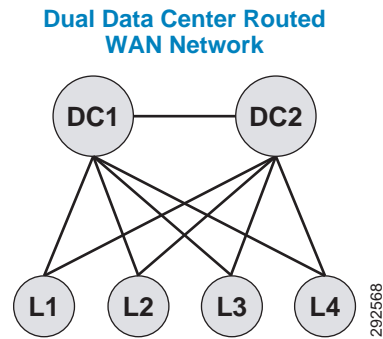
Typically these WAN uplinks from the remote sites to the data centers are load-balanced or in a primary/backup configuration, and there are limited ways for a static CAC mechanism to handle these scenarios. Although you could configure this multi-path topology in Enhanced Location CAC, only one path would be calculated as the effective path and would remain statically so until the weight metric was changed. A better way to support this type of network topology is to configure the two data centers as one data center or hub location in Enhanced Location CAC and configure a single link to each remote site location. Figure 13-27 illustrates an Enhanced Location (E-L) CAC locations and links overlay.

Figure 13-27   Enhanced Location CAC Topology Model for Dual Data Centers

Design Recommendations

The following design recommendations for dual data centers with remote dual or more links to remote locations apply to both load-balanced and primary/backup WAN designs:

- A single location (Hub_None) represents both data centers.

- A single link between the remote locations and Hub_None protects the remote site uplinks from over-subscription during normal conditions or failure of the highest bandwidth capacity links.

- The capacity of link bandwidth allocation between the remote site and Hub_None should be equal to the lowest bandwidth capacity for the applicable Unified Communications media for a single link. For example, if each WAN uplink can support 2 Mbps of audio traffic marked EF, then the link audio bandwidth value should be no more than 2 Mbps to support a failure condition or equal-cost path routing.

MPLS Clouds

When designing for Multiprotocol Label Switching (MPLS) any-to-any connectivity type clouds in the Enhanced Location CAC network model, a single location can serve as the MPLS cloud. This location will not have any devices associated to it, but all of the sites that have uplinks to this cloud will have links configured to the location. In this way the MPLS cloud serves as a transit location for interconnecting multiple variable-sized bandwidth WAN uplinks to other remote locations. The illustrations in this section depict a number of different MPLS networks and their equivalent locations and links model.

In Figure 13-28, Hub_None represents the MPLS cloud serving as a transit location interconnecting the campus location where servers, endpoints, and devices are located, with remote locations where only endpoints and devices are located. Each link to Hub_None from the remote location may be sized according to the WAN uplink bandwidth allocated for audio, video, and immersive media.
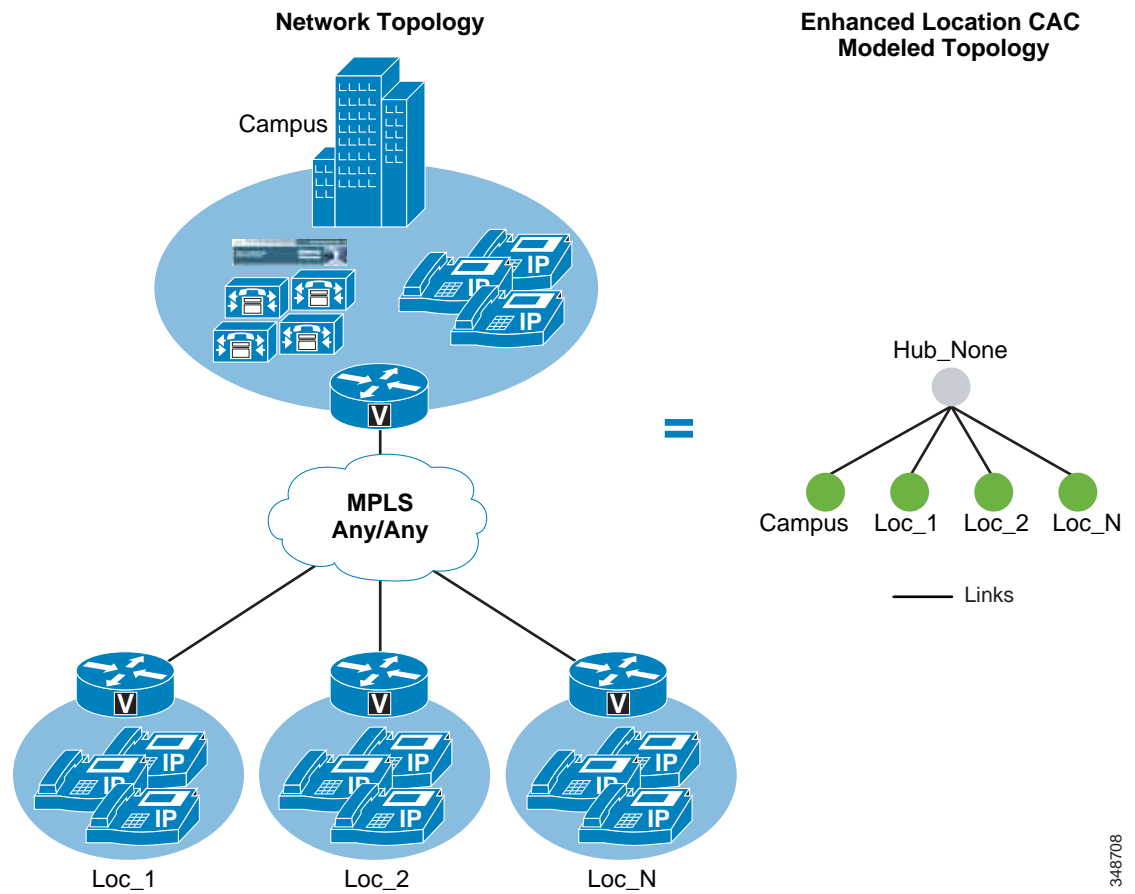
Figure 13-28   *Single MPLS Cloud*

Figure 13-29 shows two MPLS clouds that serve as transit locations interconnecting the campus location where servers, endpoints, and devices are located, with remote locations where only endpoints and devices are located. The campus also connects to both clouds. Each link to the MPLS cloud from the remote location may be sized according to the WAN uplink bandwidth allocated for audio, vide, and immersive media. This design is typical in enterprises that span continents, with a separate MPLS cloud from different providers in each geographical location.
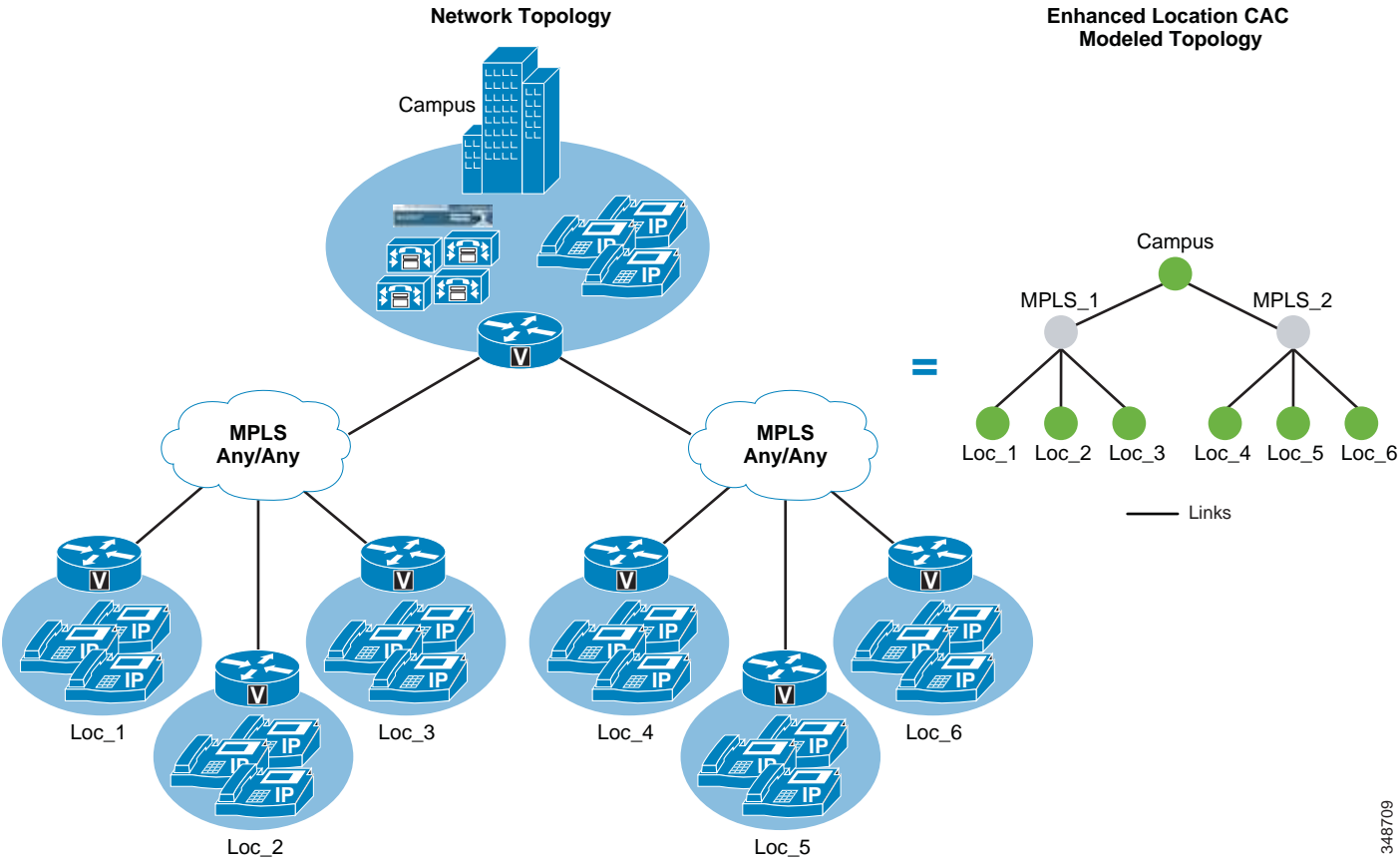
Figure 13-29   *Separate MPLS Clouds*

Figure 13-30 shows multiple MPLS clouds from different providers, where each site has one connection to each cloud and uses the MPLS clouds in either an equal-cost load-balanced manner or in a primary/backup scenario. In any case, this design is equivalent to the dual data center design where a single location represents both clouds and a single link represents the lowest capacity link of the two.

Figure 13-30   Remote Sites Connected to Dual MPLS Clouds
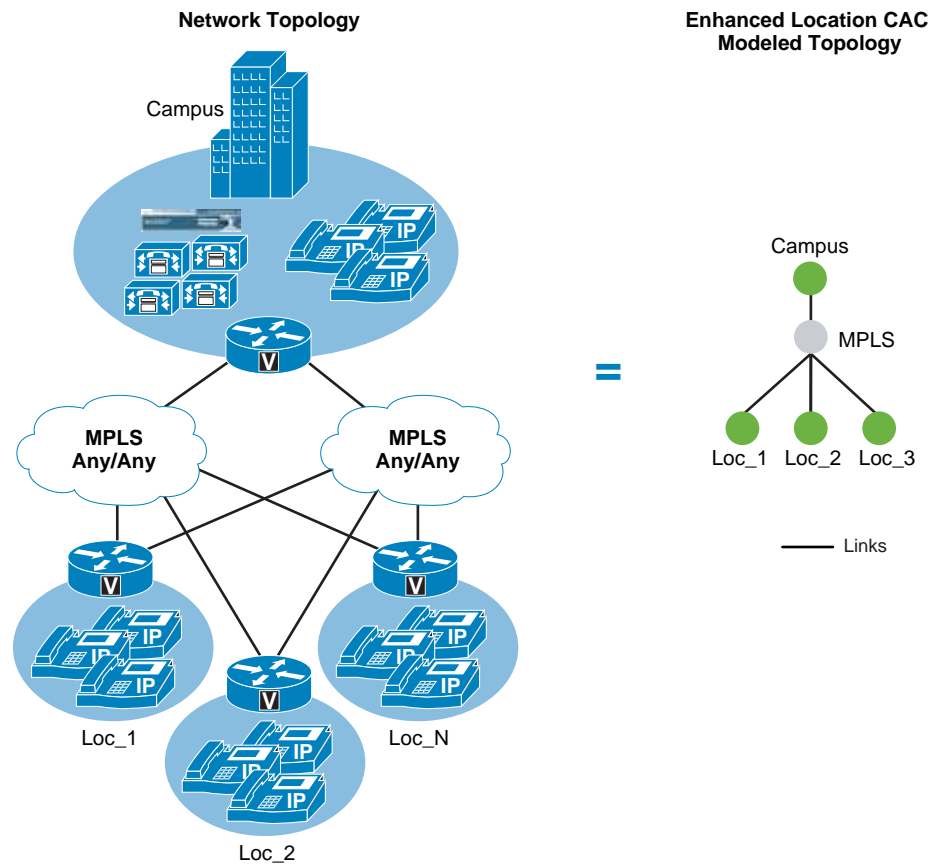
Design Recommendations

- The MPLS cloud should be configured as a location that does not contain any endpoints but is used as a hub to interconnect locations.

- The MPLS cloud serves as a transit location for interconnecting multiple variable-sized bandwidth WAN uplinks to other remote locations.

- Remote sites with connectivity to dual MPLS clouds should treat those connections as a single link and size to the lowest capacity of the links in order to avoid oversubscription during network failure conditions.

Call Admission Control Design Recommendations for Video Deployments

This section discusses Enhanced Location CAC and the design considerations and recommendations applicable to Quality of Service (QoS) when designing video deployments.

Admission control and QoS are complementary and in most cases co-dependent. Current Cisco product offerings such as audio and video endpoints, voice and video gateways, voice messaging, and conferencing all support native QoS packet marking based on IP Differentiated Services Code Point (IP DSCP). Note, however, that Jabber for Windows clients specifically do not follow the same native marking ability that other clients do, due to how the Windows operating systems requires the use of Group Policy Objects (GPO) using application, IP addresses, and UDP/TCP port ranges to mark traffic with DSCP. Group Policy Objects are very similar in function to network access lists in their ability to mark traffic.

QoS is critical to admission control because without it the network has no way of prioritizing the media to ensure that admitted traffic gets the network resources that it requires above that of non-admitted or other traffic classifications. In Unified CM's CallManager service parameters for QoS, there are five main QoS settings that are applicable to endpoint media classification and that also allow immersive and desktop classified endpoints (see the section on Enhanced Location CAC for TelePresence Immersive Video, page 13-22) to have different QoS markings for their media based on their video classification of immersive or desktop. Table 13-6 shows the five main DSCP settings along with their default settings and Per Hop Behavior (PHB) equivalents.

*Table 13-6   QoS settings for Endpoint Media Classification*

Cisco CallManager Service parameters > Clusterwide Parameters (System - QOS)	Default Value	PHB Equivalent
DSCP for Audio Calls	46	EF
DSCP for Video Calls	34	AF41
DSCP for Audio Portion of Video Calls	34	AF41
DSCP for TelePresence Calls	32	CS4
DSCP for Audio Portion of TelePresence Calls	32	CS4

The **DSCP for Audio Calls** setting is used for any device that makes an audio-only call. The **DSCP for Video Calls** setting is used for the audio and video traffic of any device that is classified as "desktop." **DSCP for TelePresence Calls** is used for the audio and video traffic of any device that is classified as "immersive." The **DSCP for Audio Portion of Video Calls** and **DSCP for Audio Portion of TelePresence Calls** are currently applicable only to the Cisco DX Series endpoints and only differentiate the audio portion of video calls dependent on the video call type based on classification.

As mentioned in the section on Enhanced Location CAC for TelePresence Immersive Video, page 13-22, Cisco Unified CM E-LCAC has the ability to perform admission control for TelePresence calls separately from other video calls. E-LCAC does this through a classification of endpoints and SIP trunks as "immersive" or "desktop." This classification gives Unified CM the ability to deduct bandwidth from a separate immersive bandwidth pool for those devices and trunks classified as immersive. By default LBM deducts ALL video, no matter the classification, from the video bandwidth pool (Unified CM's CallManager service parameter **Use Video BandwidthPool for Immersive Video Calls** set to **True**).

Also by default, all immersive classified endpoints have a DSCP set to CS4 (DSCP 32; **DSCP for TelePresence Calls**), while desktop endpoints have a DSCP set to AF41 (DSCP 34; **DSCP for Video Calls**). The default settings for QoS and E-LCAC differentiate DSCP but deduct all video from the same E-LCAC bandwidth pool. Figure 13-31 illustrates the QoS and E-LCAC bandwidth pool associations and defaults for immersive and desktop classified devices.

Figure 13-31   Default QoS Settings for CAC Bandwidth Pools
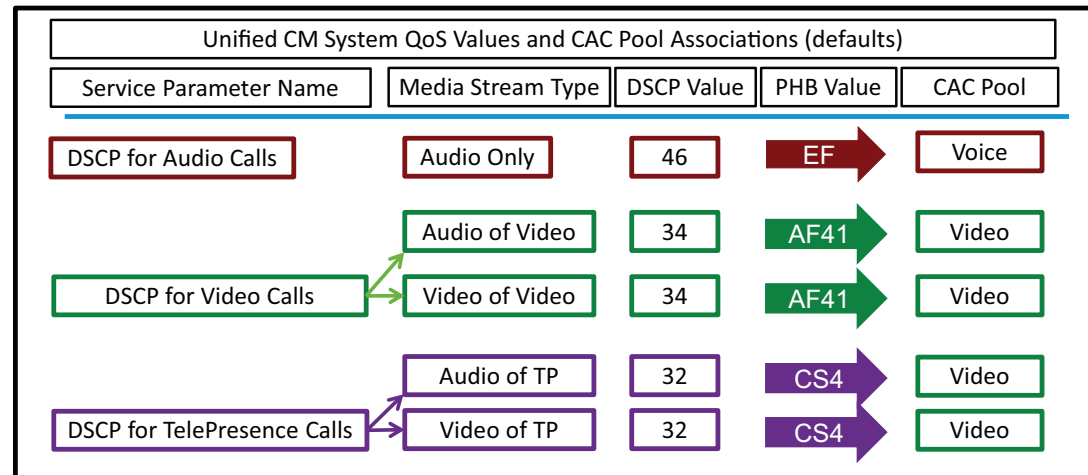
DSCP for TelePresence Calls is the immersive classification, and **DSCP for Video Calls** is the desktop classification.

Enhanced Location CAC Design Considerations and Recommendations

When designing Enhanced Location CAC for video, follow the design recommendations and considerations listed in this section.

Design Recommendations

The following design recommendations apply to video solutions that employ Enhanced Location CAC:

- If you are deploying Unified Communications video (desktop classification) and TelePresence video (immersive classification) where differentiation between desktop video and TelePresence video is a requirement, then ensure that the Unified CM service parameter **Use Video Bandwidth Pool for Immersive Video Calls** is set to **false**. This enables the immersive bandwidth pool for TelePresence calls.

- In Enhanced Location CAC, TelePresence endpoints can be managed in the same location as Unified Communications video endpoints. If TelePresence calls are not to be tracked through Enhanced Location CAC, then set the immersive location and links bandwidth pool to **unlimited**. This will ensure that CAC will not be performed on TelePresence or SIP trunks classified as immersive. If TelePresence calls are to be tracked through Enhanced Location CAC, then set immersive location and links bandwidth pool to a value according to the bit rate used and the number of calls to be allowed over the locations and link paths.

- Intercluster SIP trunks should be associated with the shadow location.

- Cisco Unified CM uses two different cluster-wide QoS service parameter to differentiate between the Differentiated Services Code Point (DSCP) settings of UC video endpoints and TelePresence endpoints. TelePresence endpoints use the **DSCP for Telepresence calls** QoS parameter while the Cisco UC video endpoints use the **DSCP for video calls** QoS service parameter.

- When marking video with the default QoS markings, the following recommendations apply:

  – For sites that deploy only UC endpoints and no TelePresence endpoints, ensure that the CS4 DSCP class is added to the AF41 QoS traffic class on inbound WAN QoS configurations to account for the inbound CS4 marked traffic, thus ensuring QoS treatment of CS4 marked media.

  – For sites that deploy only UC TelePresence endpoints and no UC endpoints, ensure that the AF41 DSCP class is added to the CS4 QoS traffic class on inbound WAN QoS configurations to account for the inbound AF41 marked traffic, thus ensuring QoS treatment of AF41 marked media.

Design Considerations

When deploying Enhanced Location CAC for immersive video calls, consider the affects of DSCP marking for both QoS classes, as the interoperable calls where an immersive classified endpoint is connected with a desktop classified endpoint are by default asymmetrically marked.

DSCP QoS Marking

The Differentiated Services Code Point (DSCP) QoS markings for TelePresence video interoperable calls are asymmetric, with AF41 used for the UC endpoints and CS4 for the TelePresence endpoints. AF41 and CS4 are default configurations in Unified CM, and changes to these defaults should align with the QoS configuration in the network infrastructure, as applicable. TelePresence endpoints mark video calls with a DSCP value of CS4, which is consistent with the default **DSCP for Telepresence calls** setting. UC endpoints mark calls with a DSCP value of AF41, which is consistent with the default **DSCP for Video calls** setting. Figure 13-32 illustrates the media marking and bandwidth accounting.

Figure 13-32   *Bandwidth Deductions and Media Marking in a Multi-Site Deployment with Enhanced Location CAC*
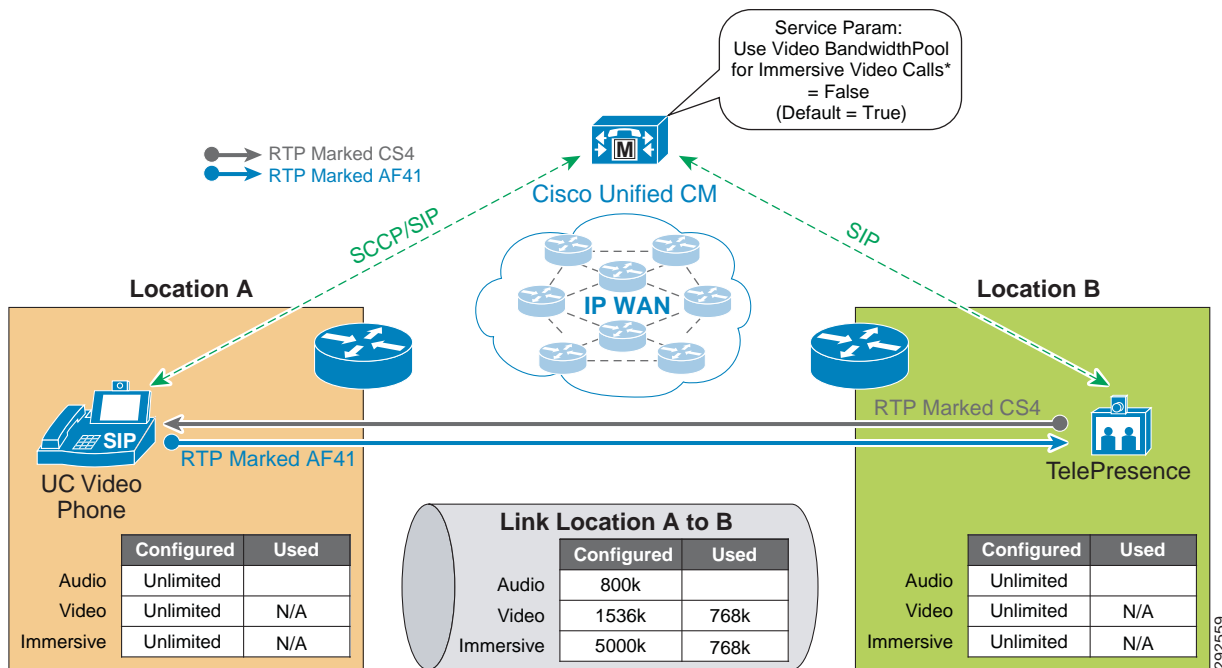
Bandwidth Accounting for TelePresence Video Interoperability Calls

Enhanced Location CAC for TelePresence-to-UC video interoperable calls deducts bandwidth from both the video and immersive locations and links bandwidth pools, as illustrated in Figure 13-32. This is by design to ensure that both types of QoS classified streams have the bandwidth required for media in both directions of the path between endpoints.

Enhanced Location CAC accounts for the bidirectional media of both AF41 and CS4 class traffic. In asymmetrically marked flows, however, the full allocated bit rate of the AF41 class is used in one direction but not the other. In the other direction, the full allocated bit rate is marked CS4. This does not represent additional bandwidth consumption but simply a difference in marking and queuing in the network for each QoS class. This manner of bandwidth accounting is required to protect each flow in each direction.

If TelePresence video (CS4) has been provisioned in the network paths separately from Unified Communications video (AF41) and TelePresence is largely scheduled and in environments where the scheduling of calls is controlled and the utilization of TelePresence is deterministic, then immersive video bandwidth for locations and links can be set to **unlimited** to avoid the double bandwidth CAC calculations. This ensures that TelePresence-to-TelePresence calls always go through unimpeded and will not be subject to admission control, while desktop video and TelePresence-to-desktop video calls will be subject to admission control and accounted for in the video bandwidth allocation.

For more information on the call flows for Enhanced Location CAC and TelePresence interoperable calls, see the section on Enhanced Location CAC for TelePresence Immersive Video, page 13-22.

Design Recommendations for Unified CM Session Management Edition Deployments with Enhanced Location CAC

Unified CM Session Management Edition (SME) is typically used for interconnecting multiple Unified CM clusters, third-party UC systems (IP- and TDM-based PBXs), PSTN connections, and centralized UC applications as well as for dial-plan and trunk aggregation. The following is a list of recommendations and design considerations to follow when deploying Unified CM SME with Enhanced Location CAC. For more information on Unified CM SME, see the chapter on Collaboration Deployment Models, page 10-1.

Recommendations and Design Considerations

- All leaf clusters that support Enhanced Location CAC should be enabled for intercluster Enhanced Location CAC with SME.

- SME can be used as a centralized bootstrap hub for the Enhanced Location CAC intercluster hub replication network. See LBM Hub Replication Network, page 13-15, for more information.

- All trunks to leaf clusters supporting Enhanced Location CAC should be SIP trunks placed in the shadow location to enable Enhanced Location CAC on the trunk between SME and the leaf clusters supporting Enhanced Location CAC.

- For TelePresence video interoperability, see the section on Call Admission Control Design Recommendations for Video Deployments, page 13-42.

- Connectivity from SME to any trunk or device other than a Unified CM that supports Enhanced Location CAC (some examples are third-party PBXs, gateways, Unified CM clusters that support only traditional Location CAC, voice messaging ports or trunks to conference bridges, Cisco Video Communications Server, and so forth) should be configured in a location other than a phantom or shadow location. The reason for this is that both phantom and shadow locations are non-terminating locations; that is, they relay information about locations and are effectively placeholders for user-defined locations on other clusters. Phantom locations are legacy locations that allow for the transmission of location information in versions of Unified CM that support only traditional Location CAC, but they are not supported with Unified CM Enhanced Location CAC. Shadow locations are special locations that enable trunks between Unified CM clusters that support Enhanced Location CAC to accomplish it end-to-end.

- SME can be used as a locations and link management cluster. See Figure 13-33 as an example of this.

- SME can support a maximum of 2,000 locations configured locally.

Figure 13-33   Unified CM SME as a Location and Link Management Cluster
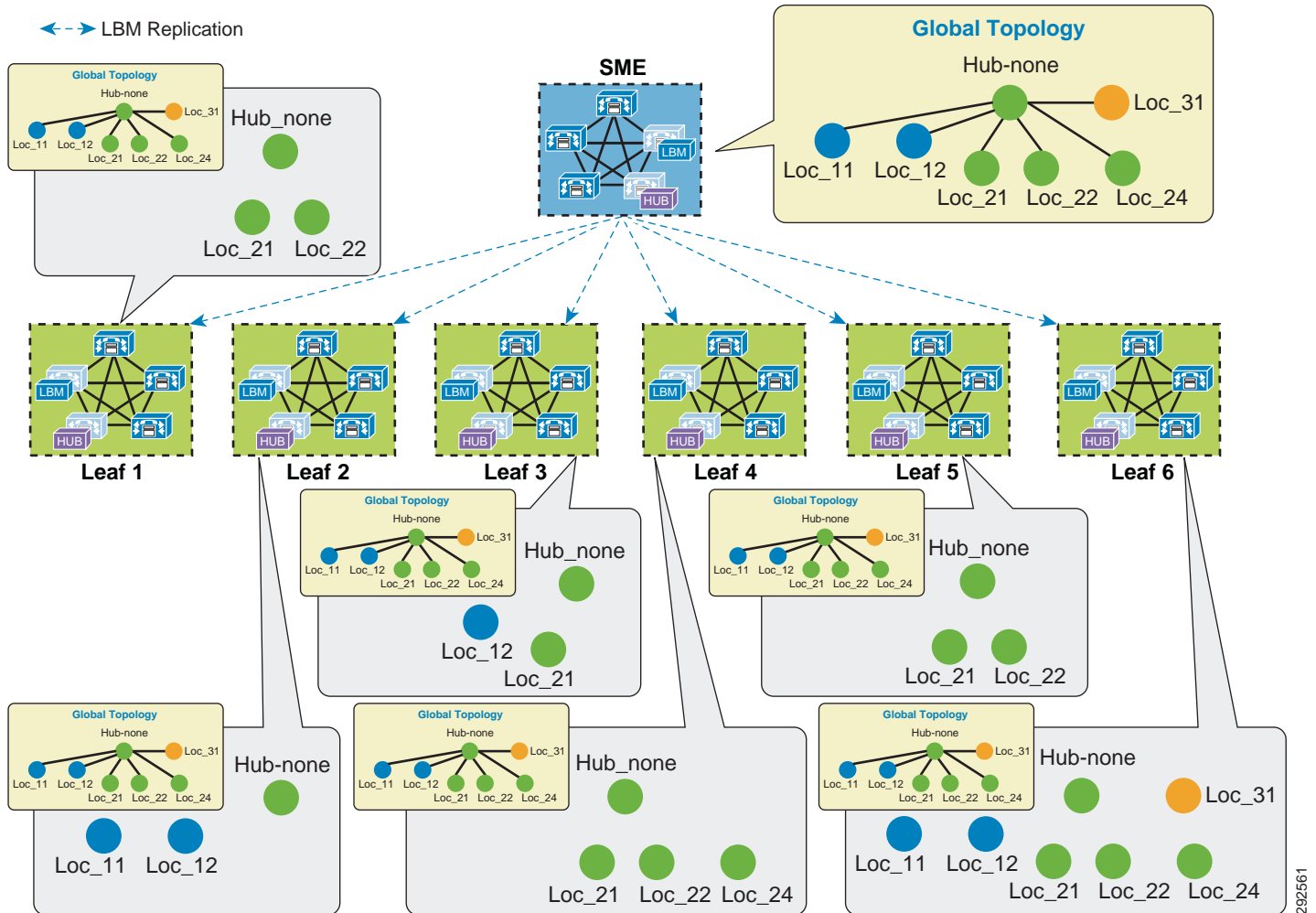

Figure 13-33 illustrates SME as a location and link management cluster. The entire location and link global topology is configured and managed in SME, and the leaf clusters configure locally only the locations that they require to associate with the end devices. When intercluster Enhanced Location CAC is enabled and locations and links are replicated, each leaf cluster will receive the global topology from SME and overlay this on their configured topology and use the global topology for call admission control. This simplifies configuration and location and link management across multiple clusters, and it diminishes the potential for misconfiguration across clusters. For more information and details on the design and deployment see the section on Location and Link Management Cluster, page 13-19.

Figure 13-34 illustrates an SME design where intercluster Enhanced Location CAC has been enabled on one or more leaf clusters (right) and where one or more leaf clusters are running a version of Unified CM that supports only traditional Location CAC (left). In this type of a deployment the locations managed by traditional Location CAC cannot be common or shared locations between clusters enabled for Enhanced Location CAC. Leaf 1 has been configured in a traditional hub and spoke, where devices are managed at various remote sites. SME and the other leaf clusters that are enabled for intercluster Enhanced Location CAC share a global topology, as illustrated in the E-L CAC Modeled Topology. Leaf1_Hub is a user-defined location in SME assigned to the SIP or H.323 intercluster trunk that

represents the hub of the Leaf 1 topology. This allows SME to deduct bandwidth for calls to and from Leaf 1 up to the Leaf1_Hub. In this way SME and Leaf 2 manage the Enhanced Location CAC locations and links while Leaf 1 manages its remote locations with traditional Location CAC.

Figure 13-34   SME Design with Enhanced Location CAC and Traditional Location CAC in Leaf Clusters
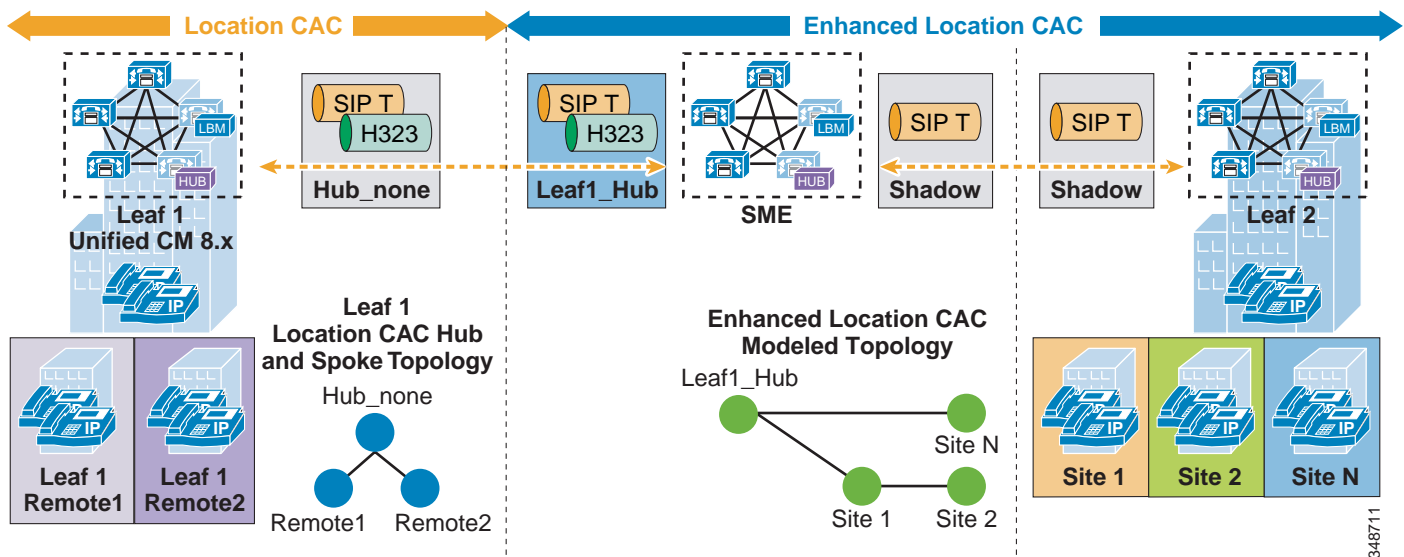
Design Recommendations for Cisco Expressway Deployments with Enhanced Location CAC
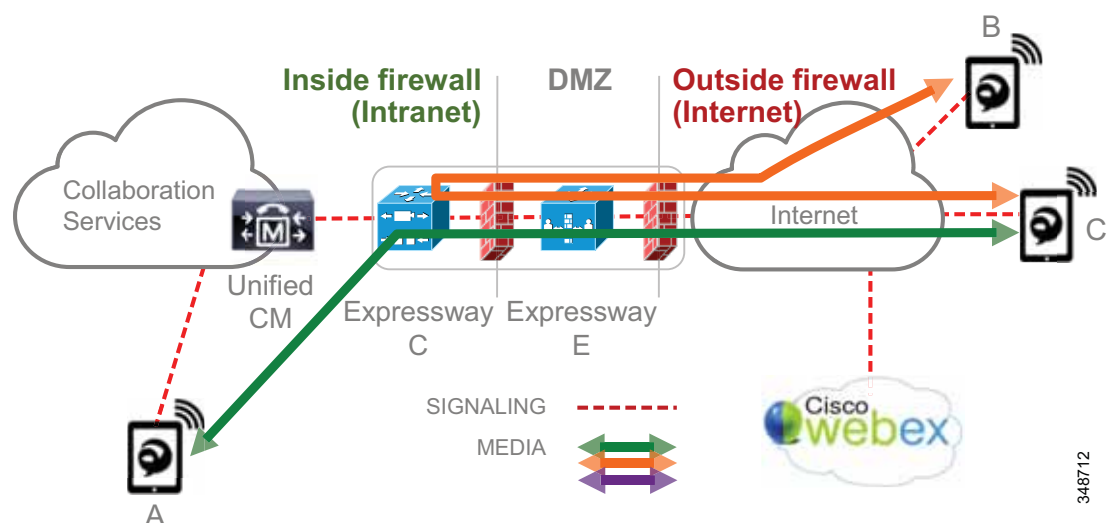
Cisco Expressway mobile and remote access capabilities provide registration of Internet-based devices to Unified CM without the use of a VPN, otherwise known as VPN-less enterprise access. This allows the endpoint or client application to register securely to Unified CM without the need for the entire operating system hosting the application to have access to the enterprise network. The following section lists the recommendations and design considerations for deploying mobile and remote access with Enhanced Location Call Admission Control (ELCAC). For more information on mobile and remote access, refer to the section on VPN-Less Access with Cisco Expressway, page 10-36.

Recommendations and Design Considerations

In the Cisco Expressway VPN-less mobile and remote access solution, endpoints supporting the feature can register to Unified CM through a Cisco Expressway deployment without the use of a VPN. Cisco Expressway C and Expressway E servers are deployed, each with redundancy for high availability. Expressway E is placed in the DMZ between the firewall to the Internet (outside) and the firewall to the enterprise (inside), while Expressway C is placed inside the enterprise. Figure 13-35 illustrates this deployment. It also illustrates the following media flows:

1. For Internet-based endpoints calling one another, the media is routed through Cisco Expressway E and Expressway C back out to the Internet, as is illustrated between endpoints B and C in Figure 13-35.

2. For Internet-based endpoints calling internal endpoints, the media flows through the Expressway E and Expressway C, as is illustrated between endpoints A and C in Figure 13-35.

Figure 13-35   Deployment of Cisco Expressway for VPN-less Access

For multiple deployments of Cisco Expressway for VPN-less access in the same enterprise, with the Internet-based endpoints registered through one Expressway pair calling Internet-based endpoints registered through another Expressway pair, the media will be routed through the enterprise. This is illustrated in Figure 13-36 with a call between endpoint D and endpoint C, both registered from the Internet but through two different Expressway pairs. The media flow will be the same whether the endpoints are registered to the same Unified CM cluster or to different Unified CM clusters.

Figure 13-36   Media Flow for a Deployment of Multiple Cisco Expressway Pairs

Figure 13-37 illustrates an example configuration for locations and links that integrate bandwidth tracking for media flows that traverse the enterprise, while still allowing media flows over the Internet without admission control.

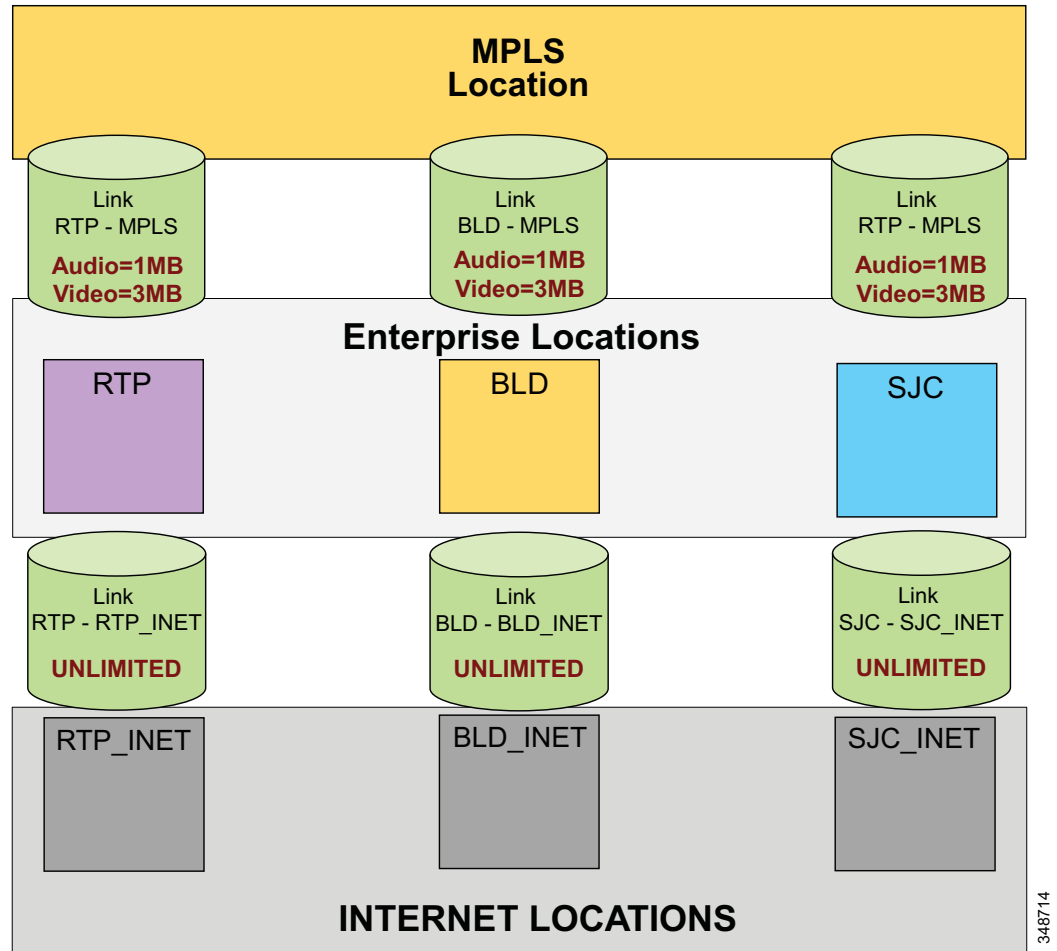*Figure 13-37*   *Locations and Links for Remote and Mobile Access*

Figure 13-37 illustrates an example deployment of ELCAC consisting of three main sites: RTP, BLD, and SJC. These sites are all connected to an MPLS provider and thus each has a separate WAN connection to the MPLS cloud. Locations and links are created accordingly so that the enterprise locations are linked directly to a location called MPLS, with bandwidth links limited for audio and video calls mapping to the network topology. Devices are located in one of the three sites when in the enterprise and thus have a location associated to them. Each of these sites has a Cisco Expressway solution for VPN-less remote and mobile access for Internet-based endpoints registering to Unified CM. Three new locations are configured for the Internet-based devices, one for each Expressway solution site, named RTP_INET, BLD_INET, and SJC_INET. These three locations represent "Internet locations" because they are locations for devices registering from the Internet to Unified CM through an Expressway pair. These locations are not interconnected with direct links. This is because calls between Expressways are routed through the enterprise and thus flow through the MPLS cloud. These Internet locations, instead, have a link to their associated enterprise location. For example, RTP_INET has a link to RTP, BLD_INET has a link to BLD, and so forth. These links between the Internet locations and the enterprise locations should be set to **unlimited** bandwidth.

As mentioned, Enhanced Location CAC for Cisco Expressway deployments requires the use of a feature in Unified CM called Device Mobility. (For details about this feature, see the section on Device Mobility, page 23-14.) Enabling device mobility on the endpoints allows Unified CM to know when the device is

registered through the Cisco Expressway or when it is registered from within the enterprise. Device mobility also enables Unified CM to provide admission control for the device as it roams between the enterprise and the Internet. Device mobility is able to do this by knowing that, when the endpoints register to Unified CM with the IP address of Expressway C, Unified CM will associate the applicable Internet location. However, when the endpoint is registered with any other IP address, Unified CM will use the enterprise location that is configured directly on the device (or from the device pool directly configured on the device). It is important to note that device mobility does not have to be deployed across the entire enterprise for this function to work. Configuration of Device Mobility in Unified CM is required only for the Expressway IP addresses, and the feature is enabled only on the devices that require the function (that is to say, those devices registering through the Internet). Figure 13-38 illustrates an overview of the device mobility configuration. Although this is a minimum configuration requirement for Device Mobility for ELCAC to function for internet-based devices, Device Mobility can be configured to support mobility for these same endpoints within the enterprise. (See the section on Device Mobility, page 23-14, for more information.)

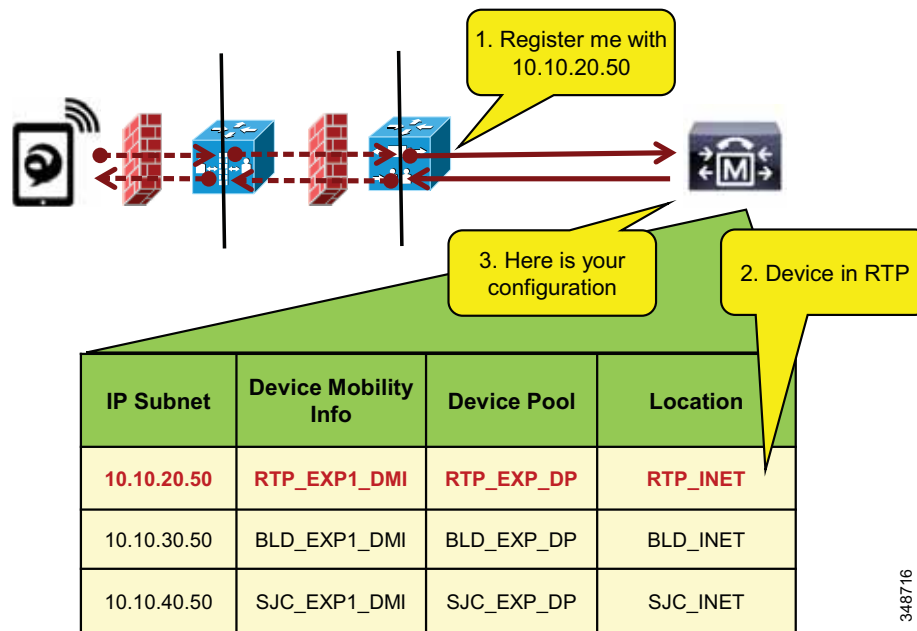*Figure 13-38   Device Mobility Configuration and Location Association*
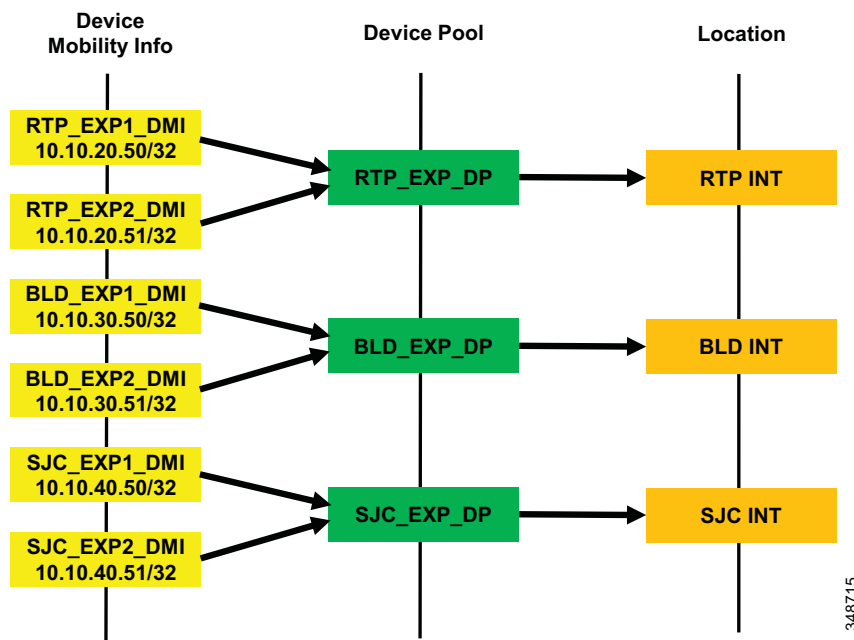
Figure 13-38 shows a simplified version of device mobility for the example deployment of ELCAC as described in Figure 13-37. The IP addresses of the Expressway C servers are configured in the device mobility information. In this example there is a redundant pair of Expressway C servers for each of the three sites, RTP, BLD, and SJC. RTP_EXP1_DMI and RTP_EXP2_DMI are configured respectively with the server IP addresses of the RTP Expressway C servers. These two are associated to a new device pool called RTP_EXP_DP, which has the location RTP_INET configured on it. Each site is configured similarly. With this configuration, when any device enabled for device mobility registers to Unified CM with the IP Address that corresponds to the device mobility information in RTP_EXP1_DMI or RTP_EXP2_DMI, it will be associated with the RTP_EXP_DP device pool and thus with the RTP_INET location.

With the above configuration, when an Internet-based device registers through the Expressway to Unified CM, it will register with the IP address of Expressway C. Unified CM then uses the IP address configured in the device mobility information and associates the device pool and thus the Internet

location associated to this device pool. This process is illustrated in Figure 13-39.

Figure 13-39   Association of Device Pool and Location Based on Expressway IP Address

IP Subnet	Device Mobility Info	Device Pool	Location
10.10.20.50	RTP_EXP1_DMI	RTP_EXP_DP	RTP_INET
10.10.30.50	BLD_EXP1_DMI	BLD_EXP_DP	BLD_INET
10.10.40.50	SJC_EXP1_DMI	SJC_EXP_DP	SJC_INET

In Figure 13-39 the client registers with Unified CM through the Expressway in RTP. Because the signaling is translated at the Expressway C in RTP, the device registers with the IP address of the Expressway C. The device pool RTP_EXP_DP is associated to the device based on this IP address. The RTP_EXP_DP pool is configured with the RTP_INET location, and therefore that location is associated to the device. Thus, when devices register to the Expressway, they get the correct location association through device mobility. When the endpoint relocates to the enterprise, it will return to its static location configuration. Also, if the endpoint relocates to another Expressway in SJC, for example, it will get the correct location association through device mobility.

Design and Deployment Best Practices for Cisco Expressway VPN-less Access with Enhanced Location CAC

- Each site with Internet access, where a Cisco Expressway solution resides, requires an Internet location and an enterprise location. Each Cisco Expressway deployment requires these location pairs. The enterprise location is associated to devices when they are in the enterprise (see locations RTP, BLD, and SJC in Figure 13-37). The Internet location is associated to the endpoints through the Device Mobility feature when the endpoints are registering from the Internet (see locations RTP_INET, BLD_INET, and SJC_INET in Figure 13-37). For example, in Figure 13-37, RTP and RTP_INET form a location pair for the physical site RTP.

- Enterprise locations are configured according to applicable enterprise ELCAC design.

- Internet locations will always have a single link to the enterprise location that they are paired with. For example, in Figure 13-37, RTP and RTP_INET form an enterprise location and internet location pair.

- Links from Internet locations to enterprise locations are set to **unlimited** bandwidth. Unlimited bandwidth between these location pairs ensures that bandwidth is not counted for calls from the Internet location to the local enterprise location, and vice versa (for example, calls from RTP to RTP_INET in Figure 13-37).

- In a Cisco Expressway solution where more than one Cisco Expressway site is deployed, and requiring multiple Internet locations, ensure that Internet locations do not have direct links between one another. Direct links between Internet locations will create multiple paths in ELCAC, and for that reason they are not recommended.