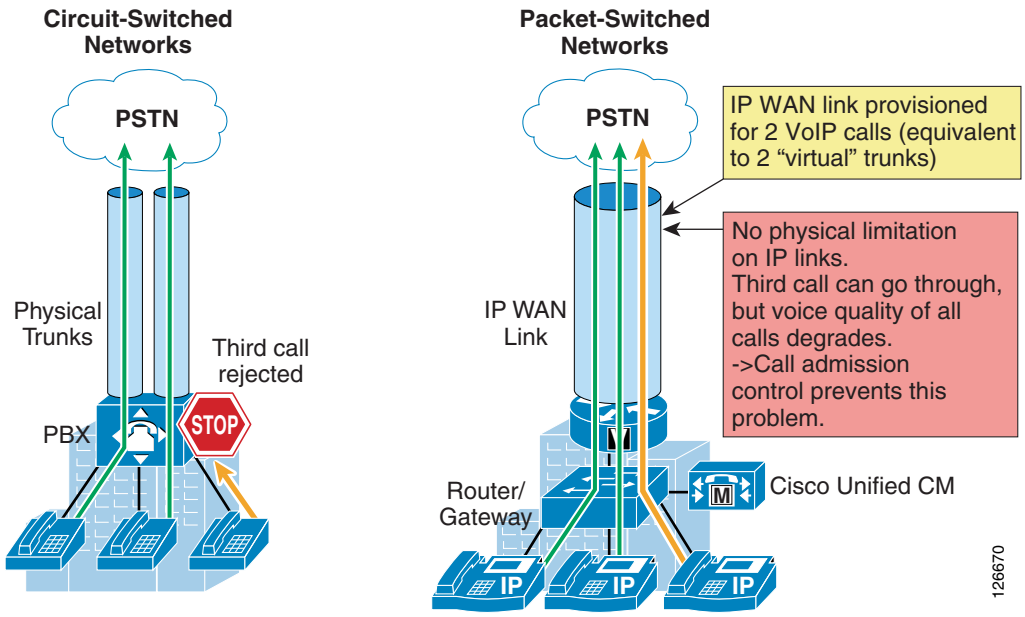
CHAPTER 11

Call Admission Control

Revised: June 28, 2012; OL-27282-05

The call admission control function is an essential 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 11-1.

Figure 11-1 Why Call Admission Control is Needed



As shown on the left side of Figure 11-1, 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 11-1. 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 Principles, page 11-3

  This section defines the two fundamental approaches to call admission control in an IP-based telephony system: topology-aware and topology-unaware call admission control.

- Call Admission Control Architecture, page 11-12

  This section describes the call admission control mechanisms available through the various components of a Cisco IP Communications system, such as Cisco Unified Communications Manager Enhanced Locations call admission control, Cisco IOS gatekeeper, RSVP, and RSVP SIP Preconditions.

- Design Considerations for Call Admission Control, page 11-93

  This section shows how to apply and combine the mechanisms described in the previous sections, based on the IP WAN topology.

What's New in This Chapter

Table 11-1 lists the topics that are new in this chapter or that have changed significantly from previous releases of this document.

Table 11-1 New or Changed Information Since the Previous Release of This Document

New or Revised Topic	Described in:	Revision Date
Enhanced Locations CAC	Unified CM Enhanced Locations Call Admission Control, page 11-12	June 28, 2012
RSVP Agent support for RTCP, BFCP and FECC	RSVP Agent Support for RTCP, BFCP and FECC Negotiation, page 11-67	June 28, 2012
Migration to RSVP Agent call admission control	Migrating to RSVP Call Admission Control, page 11-70	June 28, 2012
Example WAN topologies for call admission control	Design Considerations for Call Admission Control, page 11-93	June 28, 2012
Other updates for Cisco Unified Communications System Release 9.0	Various sections throughout this chapter	June 28, 2012

Call Admission Control Principles

As mentioned previously, call admission control is a function of the call processing agent in an IP-based telephony system, so in theory there could be as many call admission control mechanisms as there are IP-based telephony systems. However, most of the existing call admission control mechanisms fall into one of the following two main categories:

- Topology-unaware call admission control — Based on a static configuration within the call processing agent
- Topology-aware call admission control — Based on communication between the call processing agent and the network about the available resources

The remainder of this section first analyzes the principles of topology-unaware call admission control and its limitations, and then presents the principles of topology-aware call admission control.

Topology-Unaware Call Admission Control

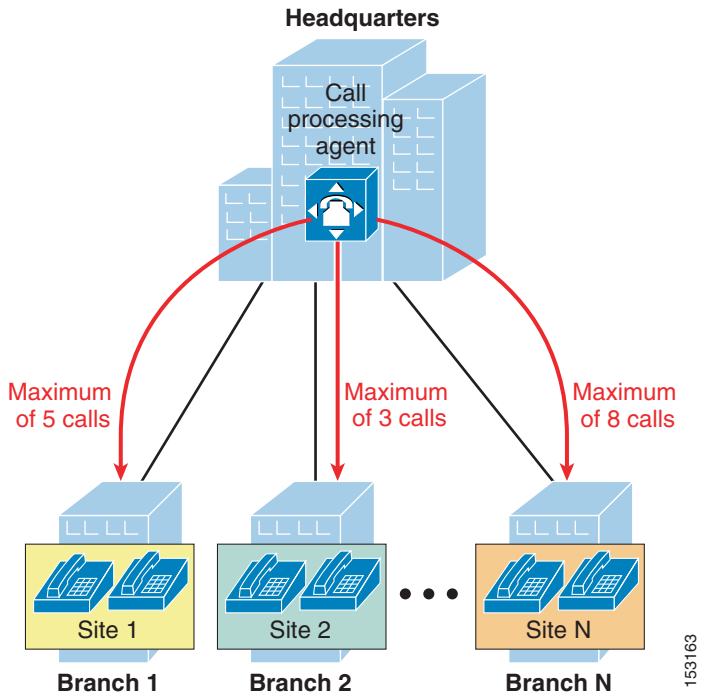
We define as topology-unaware call admission control any mechanism that is based on a static configuration within a call processing agent or IP-based PBX, aimed at limiting the number of simultaneous calls to or from a remote site connected via the IP WAN.

As shown in Figure 11-2, most of these mechanisms rely on the definition of a logical "site" entity, which generally corresponds to a geographical branch office connected to the enterprise IP WAN.

After assigning all the devices located at each branch office to the corresponding site entity, the administrator usually configures a maximum number of calls (or a maximum amount of bandwidth) to be allowed in or out of that site.

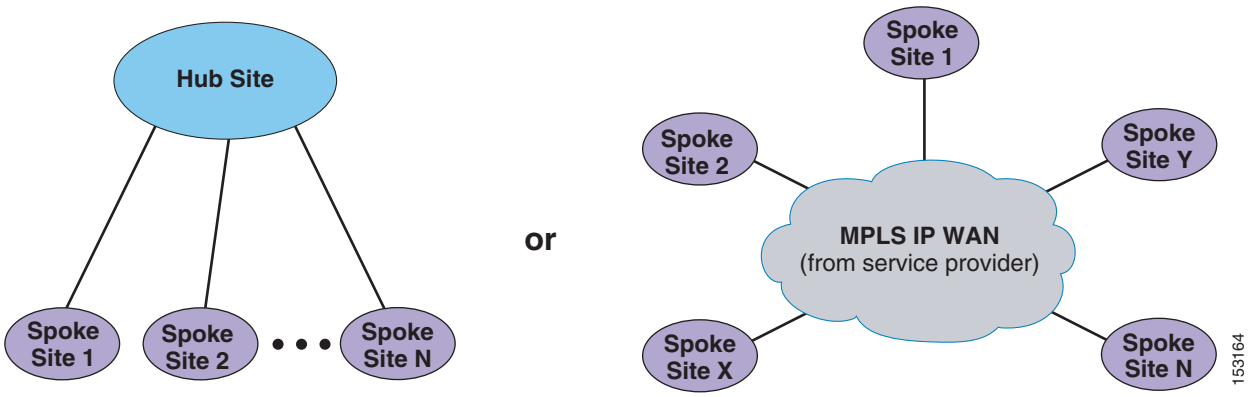
Each time a new call needs to be established, the call processing agent checks the sites to which the originating and terminating endpoints belong, and verifies whether there are available resources to place the call (in terms of number of calls or amount of bandwidth for both sites involved). If the check succeeds, the call is established and the counters for both sites are decremented. If the check fails, the call processing agent can decide how to handle the call based on a pre-configured policy. For example, it could send a network-busy signal to the caller device, or it could attempt to reroute the call over a PSTN connection.

Figure 11-2 Principles of Topology-Unaware Call Admission Control



Because of their reliance on static configurations, topology-unaware call admission control mechanisms can generally be deployed only in networks with a relatively simple IP WAN topology. In fact, most of these mechanisms mandate a simple hub-and-spoke topology or a simple MPLS-based topology (where the MPLS service is provided by a service provider), as shown in Figure 11-3.

Figure 11-3 Domain of Applicability of Topology-Unaware Call Admission Control



In a hub-and-spoke network or MPLS-based network such as those shown in Figure 11-3, each spoke site is assigned to a "site" within the call processing agent, and the number of calls or amount of bandwidth for that "site" is configured to match the bandwidth available for voice (and/or video) on the IP WAN link that connects the spoke to the IP WAN.

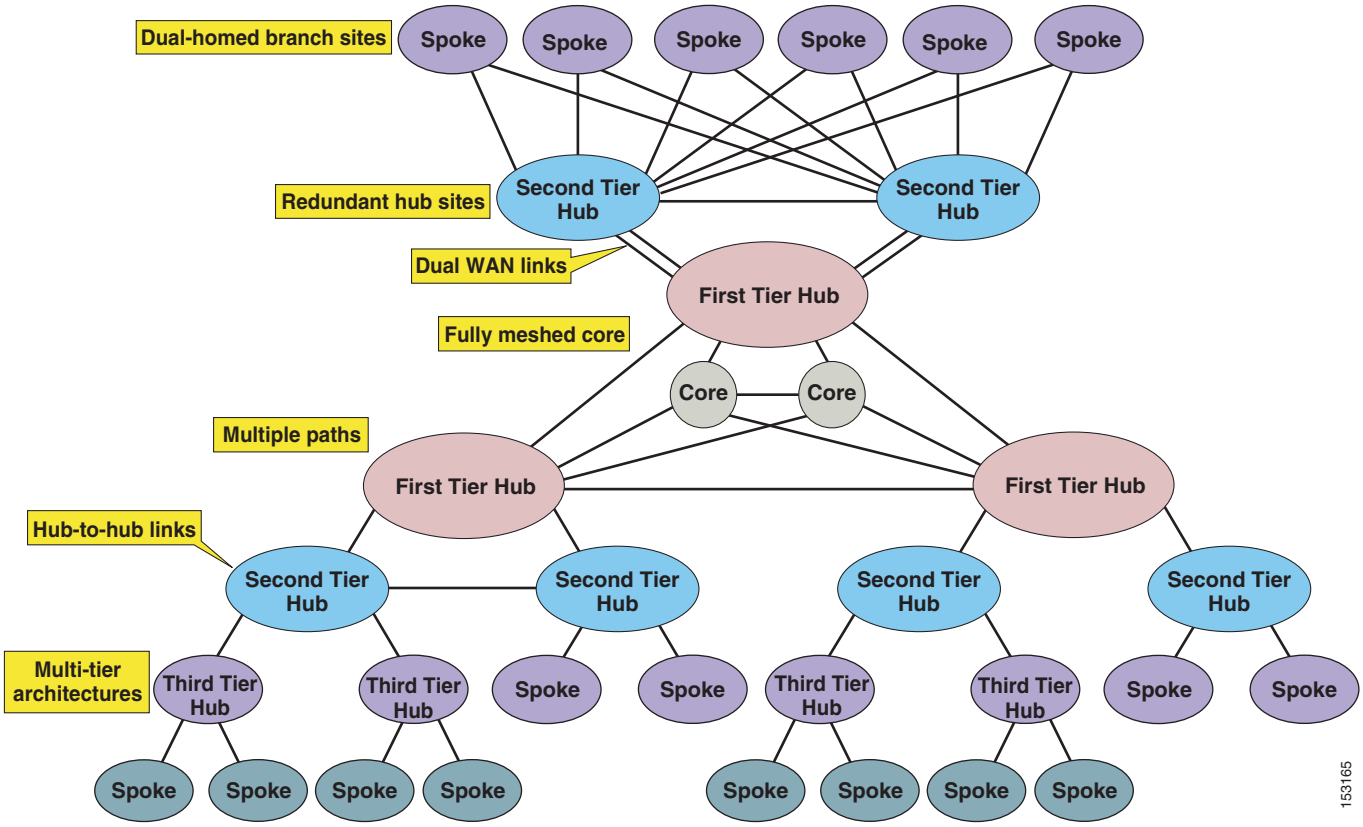
Notice the absence of redundant links from the spoke sites to the hub site and of links directly connecting two spoke sites. The next section explains why such links create problems for topology-unaware call admission control.

Limitations of Topology-Unaware Call Admission Control

In today's enterprise networks, high availability is a common requirement, and it often translates into a desire to provide redundancy for the IP WAN network connectivity.

When considering the IP WAN topology in a typical enterprise network, you are likely to encounter a number of characteristics that complicate the assumption of a pure hub-and-spoke topology. Figure 11-4 shows several of these network characteristics in a single diagram. Obviously, only the largest enterprise networks present all these characteristics at once, but it is highly likely that most IP WAN networks feature at least one of them.

Figure 11-4 Topology Characteristics of Typical Enterprise Networks

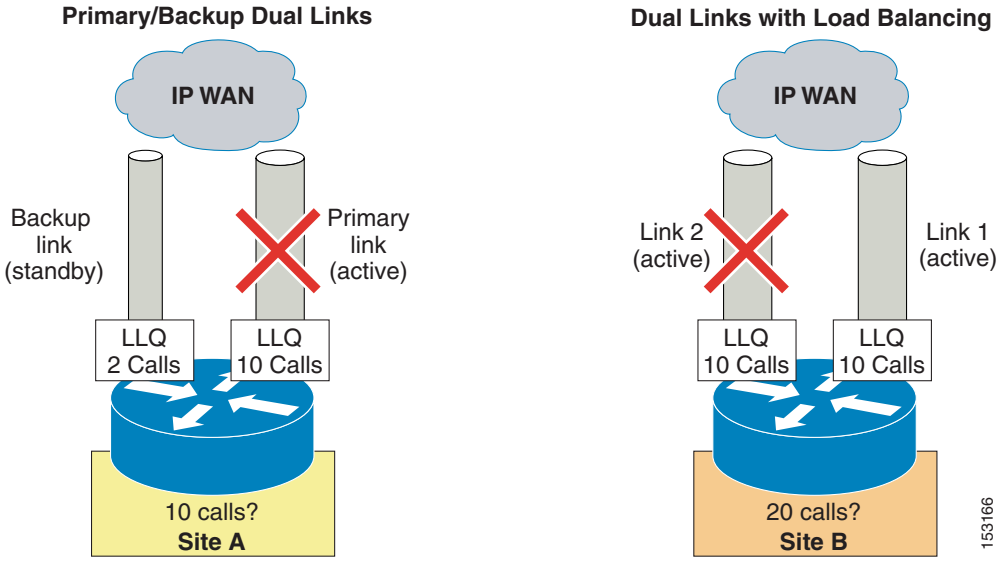


As explained in the section on Design Considerations for Call Admission Control, page 11-93, it is sometimes possible to adapt a topology-unaware call admission control mechanism to a complex network topology, but there are limitations in terms of when this approach can be used and what behavior can be achieved. For example, consider the simple case of a branch site connected to a hub site via the IP WAN, where redundancy is a network requirement. Typically, redundancy can be achieved in one of the following ways:

- A single router with a primary and a backup link to the IP WAN
- A single router with two active WAN links in a load-balancing configuration
- Two router platforms, each connected to the IP WAN, with load-balanced routing across them

The examples Figure 11-5 attempt to apply a topology-unaware call admission control mechanism to the case of a single router with a primary and backup link and the case of a single router with two active load-balanced links. (The case of two router platforms has the same call admission control implications as the latter example.)

Figure 11-5 Topology-Unaware Call Admission Control in Presence of Dual Links



For the first example in Figure 11-5, branch office A is normally connected to the IP WAN via a primary link, whose Low Latency Queuing (LLQ) bandwidth is provisioned to allow a maximum of 10 simultaneous calls. When this primary link fails, a smaller backup link becomes active and preserves the connectivity to the IP WAN. However, the LLQ bandwidth of this backup link is provisioned to allow only up to 2 simultaneous calls.

In order to deploy a topology-unaware call admission control mechanism for this branch office, we must define a "site" A in the call processing agent and configure it for a certain number of calls (or amount of bandwidth). If we choose to use 10 calls as the maximum for site A, the backup link can be overrun during failures of the primary link, thereby causing bad voice quality for all active calls. If, on the other hand, we choose 2 calls as the maximum, we will not be able to use the bandwidth provisioned for the remaining 8 calls when the primary link is active.

Now consider branch office B, which has two active links connecting it to the IP WAN. Each of these links is provisioned to allow a maximum of 10 simultaneous calls, and the routing protocol automatically performs load-balancing between them. When deploying a topology-unaware call admission control mechanism for this branch office, we must define a "site" B in the call processing agent and configure it for a certain number of calls (or amount of bandwidth). Similar to the case of branch office A, if we choose to add up the capacity of the two links and use 20 calls as the maximum for site B, there is a potential to overrun the LLQ on one of the two links during failures of the other one. For example, if link #2 fails, the system still allows 20 simultaneous calls to and from site B, which are now all routed via link #1, thus overrunning it and causing poor voice quality for all calls. On the other hand, if site B is configured for a maximum of 10 simultaneous calls, the available LLQ bandwidth is never fully utilized under normal conditions (when both links are operational).

These two simple examples show how IP WAN bandwidth provisioning in real enterprise networks is often too complex to be summarized in statically configured entries within the call processing agent. Deploying topology-unaware call admission control in such networks forces the administrator to make assumptions, develop workarounds, or accept sub-optimal use of network resources.

The optimal way to provide call admission control in the presence of a network topology that does not conform to a simple hub-and-spoke is to implement topology-aware call admission control, as described in the following section.

Note Some IP telephony systems augment classic topology-unaware call admission control with a feedback mechanism based on observed congestion in the network, which forces calls through the PSTN when voice quality deteriorates. This approach is still not equivalent to true topology-aware call admission control because it is performed after the calls have already been established and because the call processing agent still does not have knowledge of exactly where congestion is occurring. As mentioned at the beginning of the chapter, in order to be effective, call admission control must be performed before the call is set up.

Topology-Aware Call Admission Control

We define as topology-aware call admission control any mechanism aimed at limiting the number of simultaneous calls across IP WAN links that can be applied to any network topology and can dynamically adjust to topology changes.

To accomplish these goals, topology-aware call admission control must rely on real-time communications about the availability of network resources between a call processing agent (or IP-based PBX) and the network. Because the network is a distributed entity, real-time communications require a signaling protocol.

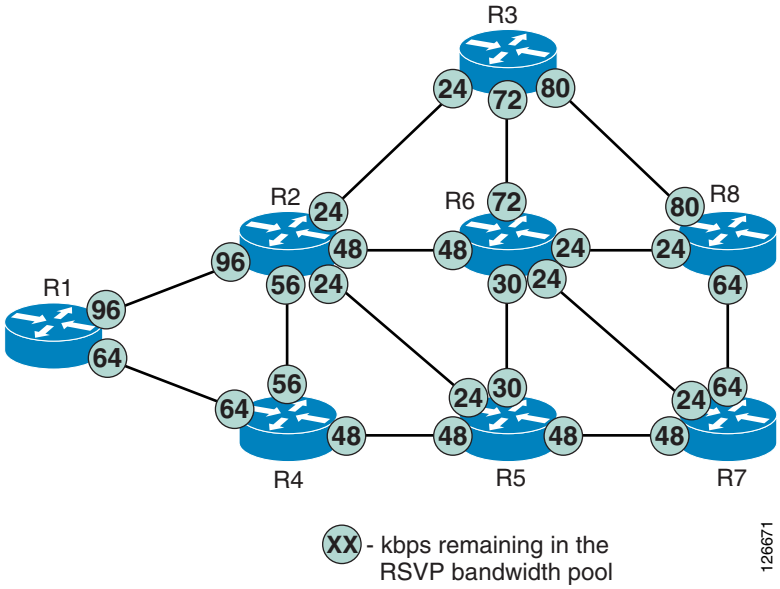
The Resource Reservation Protocol (RSVP) is the first significant industry-standard signaling protocol that enables an application to reserve bandwidth dynamically across an IP network. Using RSVP, applications can request a certain amount of bandwidth for a data flow across a network (for example, a voice call) and can receive an indication of the outcome of the reservation based on actual resource availability.

In the specific case of call admission control for voice or video calls, an IP-based PBX can synchronize the call setup process with RSVP reservations between the two remote sites and can make a routing decision based on the outcome of the reservations. Because of its distributed and dynamic nature, RSVP is capable of reserving bandwidth across any network topology, thus providing a real topology-aware call admission control mechanism.

To better understand the basic principles of how RSVP performs bandwidth reservation in a network, consider the simple example depicted in Figure 11-6. This example does not analyze the exact message exchanges and protocol behaviors, but rather focus on the end results from a functionality perspective. For more information on the RSVP message exchanges, see RSVP Principles, page 11-42.

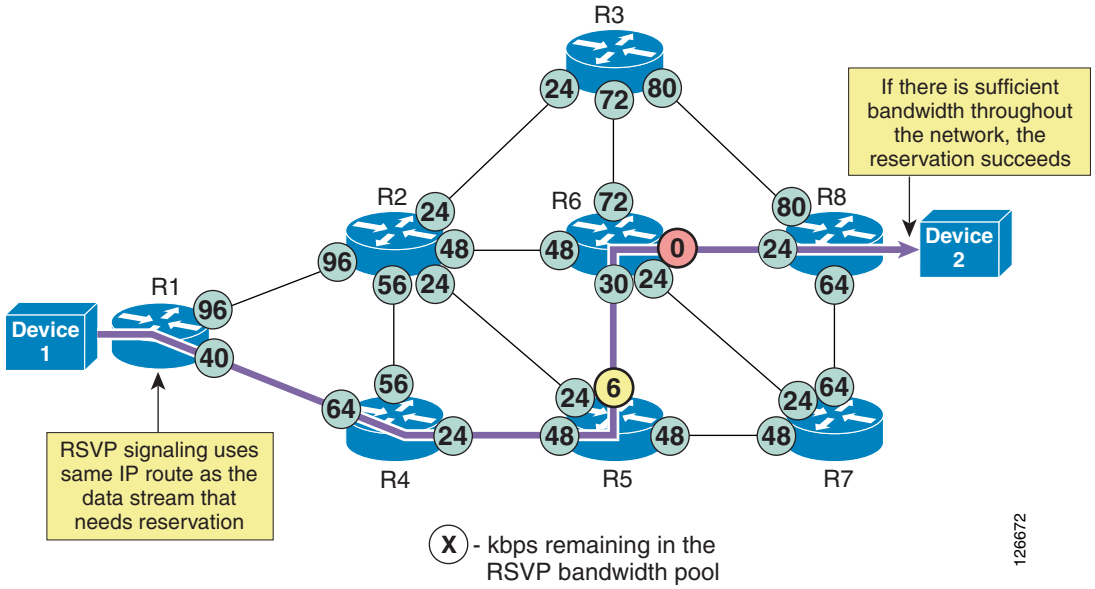
Assume that RSVP is enabled on each router interface in the network shown in Figure 11-6 and that the numbers shown in the circles represent the amount of available RSVP bandwidth remaining on each interface.

Figure 11-6 Sample Network to Show RSVP Principles



Now consider an RSVP-enabled application that wants to reserve a certain amount of bandwidth for a data stream between two devices. This scenario is depicted in Figure 11-7, which shows a particular data stream that requires 24 kbps of bandwidth from Device 1 to Device 2.

Figure 11-7 RSVP Signaling for a Successful Reservation

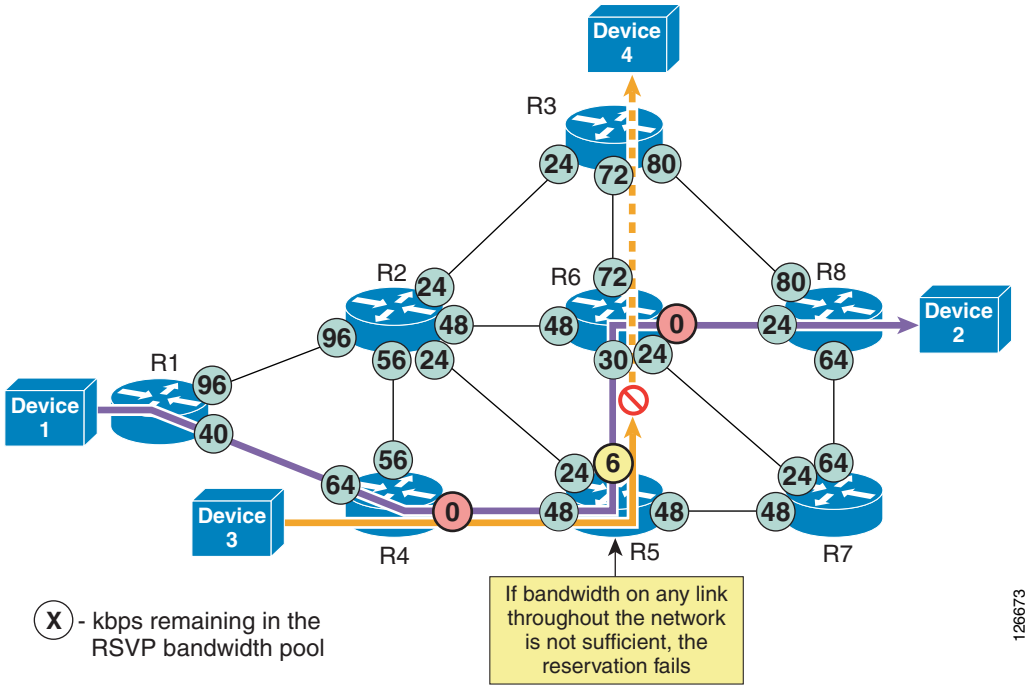


The following considerations apply to Figure 11-7:

- RSVP does not perform its own routing; instead it uses underlying routing protocols to determine where it should carry reservation requests. As routing changes paths to adapt to topology changes, RSVP adapts its reservations to the new paths wherever reservations are in place.
- The RSVP protocol attempts to establish an end-to-end reservation by checking for available bandwidth resources on all RSVP-enabled routers along the path from Device 1 to Device 2. As the RSVP messages progress through the network, the available RSVP bandwidth gets decremented by 24 kbps on the outbound router interfaces, as shown in Figure 11-7.
- The available bandwidth on all outbound interfaces is sufficient to accept the new data stream, so the reservation succeeds and the application is notified.
- RSVP reservations are unidirectional (in this case, the reservation is established from Device 1 to Device 2, and not vice versa). In the presence of bidirectional applications such as voice and videoconferencing, two reservations must be established, one in each direction.
- RSVP provides transparent operation through router nodes that do not support RSVP. If there are any routers along the path that are not RSVP-enabled, they simply ignore the RSVP messages and pass them along like any other IP packet, and a reservation can still be established. (See RSVP Principles, page 11-42, for details on protocol messages and behaviors.) However, in order to have an end-to-end QoS guarantee, you have to ensure that there is no possibility of bandwidth congestion on the links controlled by the non-RSVP routers.

After a reservation has been successfully established between Device 1 and Device 2, now assume that another application requests a 24-kbps reservation between Device 3 and Device 4, as depicted in Figure 11-8.

Figure 11-8 RSVP Signaling for an Unsuccessful Reservation



The following considerations apply to Figure 11-8:

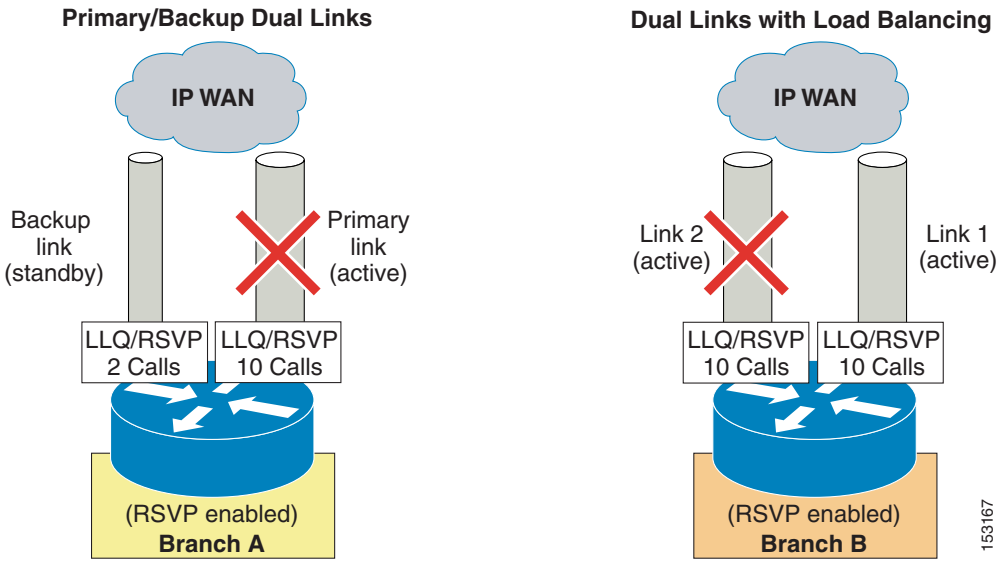
- The RSVP protocol attempts to establish an end-to-end reservation by checking for available bandwidth resources on all RSVP-enabled routers along the path from Device 3 to Device 4. As the RSVP messages progress through the network, the available RSVP bandwidth gets decremented by 24 kbps on the outbound router interfaces, as shown in Figure 11-8.
- In this example, the available bandwidth on R5's outbound interface toward R6 is not sufficient to accept the new data stream, so the reservation fails and the application is notified. The available RSVP bandwidth on each outbound interface along the path is then restored to its previous value.
- The application can then decide what to do. It could abandon the data transfer or decide to send it anyway with no QoS guarantees, as best-effort traffic.

We can now apply the topology-aware call admission control approach based on RSVP to the examples of dual-connected branch offices A and B introduced in the previous section.

As shown in Figure 11-9, branch office A has a primary link with an LLQ provisioned for 10 calls, while the backup link can accommodate only 2 calls. With this approach, RSVP is configured on both router interfaces so that the RSVP bandwidth matches the LLQ bandwidth. Branch A is also configured within the call processing agent to require RSVP reservations for all calls to or from other branches. Now calls are admitted or rejected based on the outcome of the RSVP reservations, which automatically follow the path determined by the routing protocol. Under normal conditions (when the primary link is active), up to 10 calls will be admitted; during failure of the primary link, only up to 2 calls will be admitted.

Policies can typically be set within the call processing agent to determine what to do in the case of a call admission control failure. For example, the call could be rejected, rerouted across the PSTN, or sent across the IP WAN as a best-effort call with a different DSCP marking.

Figure 11-9 Topology-Aware Call Admission Control for Dual Links



Similar considerations apply to branch B, connected to the IP WAN via two load-balanced links, as shown on the right side of Figure 11-9. RSVP is enabled on each of the two router interfaces, with a bandwidth value that matches the LLQ configuration (in this case, enough bandwidth for 10 calls). Branch B is also configured within the call processing agent to request RSVP reservations for calls to or from other branches. Again, calls are admitted or rejected based on the actual bandwidth available along

the path chosen by the routing protocol. So in a case of perfectly even load-balancing across the two links, up to 20 calls could be admitted under normal conditions (when both links are operational); if one of the two links fails, only up to 10 calls would be admitted.

In the case that one of the two links failed while more than 10 calls were active, some calls would fail to re-establish a reservation on the new path. At this point, the call processing agent would be notified and could react based on the configured policy (for example, by dropping the extra calls or by remarking them as best-effort calls).

In conclusion, topology-aware call admission control allows administrators to protect call quality with any network topology, to automatically adjust to topology changes, and to make optimal use of the network resources under all circumstances.

Special Considerations for MPLS Networks

From the call admission control perspective, a network based on MPLS differs from one based on traditional Layer 2 WAN Services with respect to support for RSVP in the "hub" of the network. Hub sites of traditional Layer 2 wide-area networks consist, in most cases, of an enterprise-controlled router that can be enabled to participate in RSVP. Because the entire network (cloud) is the "hub site" in MPLS networks, there is no enterprise-controlled hub location to enable RSVP. (For more information, see MPLS Clouds, page 11-94.) Therefore, to provide topology-aware call admission control in an MPLS environment, the Customer Edge (CE) devices of the network must be configured for RSVP support.

Because RSVP must be enabled on the CE, control of this equipment is important. If this equipment is not under the control of the enterprise, you must work with your service provider to determine if they will enable RSVP on your WAN interface and if that implementation will support advanced features such as RSVP application ID.

RSVP messages will transparently pass across the RSVP-unaware MPLS cloud, so this does not pose a problem with end-to-end RSVP capability. Configuring RSVP on the CE WAN interface will ensure that its priority queue will not be overrun. Because RSVP reservations are unidirectional, the following rules must be observed to protect the priority queue on the Provider Edge (PE) router when RSVP is not enabled in the MPLS cloud:

- The media streams must be the same size in both directions.
- The media has to be symmetrically routed.

RSVP PATH messages record the egress IP address of the RSVP-aware routers they traverse. The information in the PATH message is used to send the RSVP RESV message back via the same route. Because of this mechanism, the WAN link between CE and PE must have routable IP addresses or the RSVP Reservations will fail.

If your MPLS network does not comply with these rules, contact your local Cisco account team for further assistance before implementing RSVP.

Call Admission Control Architecture

There are several mechanisms that perform the call admission control function in a Cisco IP Communications system. This section provides design and configuration guidelines for all of these mechanisms, according to their category:

- Topology-unaware mechanisms
  - Unified CM Enhanced Locations Call Admission Control, page 11-12
  - Cisco IOS Gatekeeper Zones, page 11-40
- Topology-aware mechanisms
  - Unified CM RSVP-Enabled Locations, page 11-62
  - RSVP SIP Preconditions, page 11-73

Unified CM Enhanced Locations Call Admission Control

Cisco Unified CM 9.*x* provides Enhanced Locations call admission control (CAC) 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 Locations 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 locations-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 Locations configuration called **immersive video bandwidth**.

There are also a number of tools to administer and troubleshoot Enhanced Locations 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 Locations CAC is a model-based static CAC mechanism. Enhanced Locations 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 such as RSVP CAC. Therefore, the model needs to be updated when the WAN network topology changes. Enhanced Locations 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.

The administrator builds the network model using locations and links. Enhanced Locations CAC incorporates the following configuration components:

- Locations — A Location represents a LAN. It could contain endpoints or simply serve as a transit location between links for WAN network modeling.
- 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 shortest paths (least cost) from each location to all other locations and builds the paths. Only one "effective path" is used between a 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).
- Locations 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.
- Locations Bandwidth Manager Hub — A Locations 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 9.*x* also uses a new locations 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.

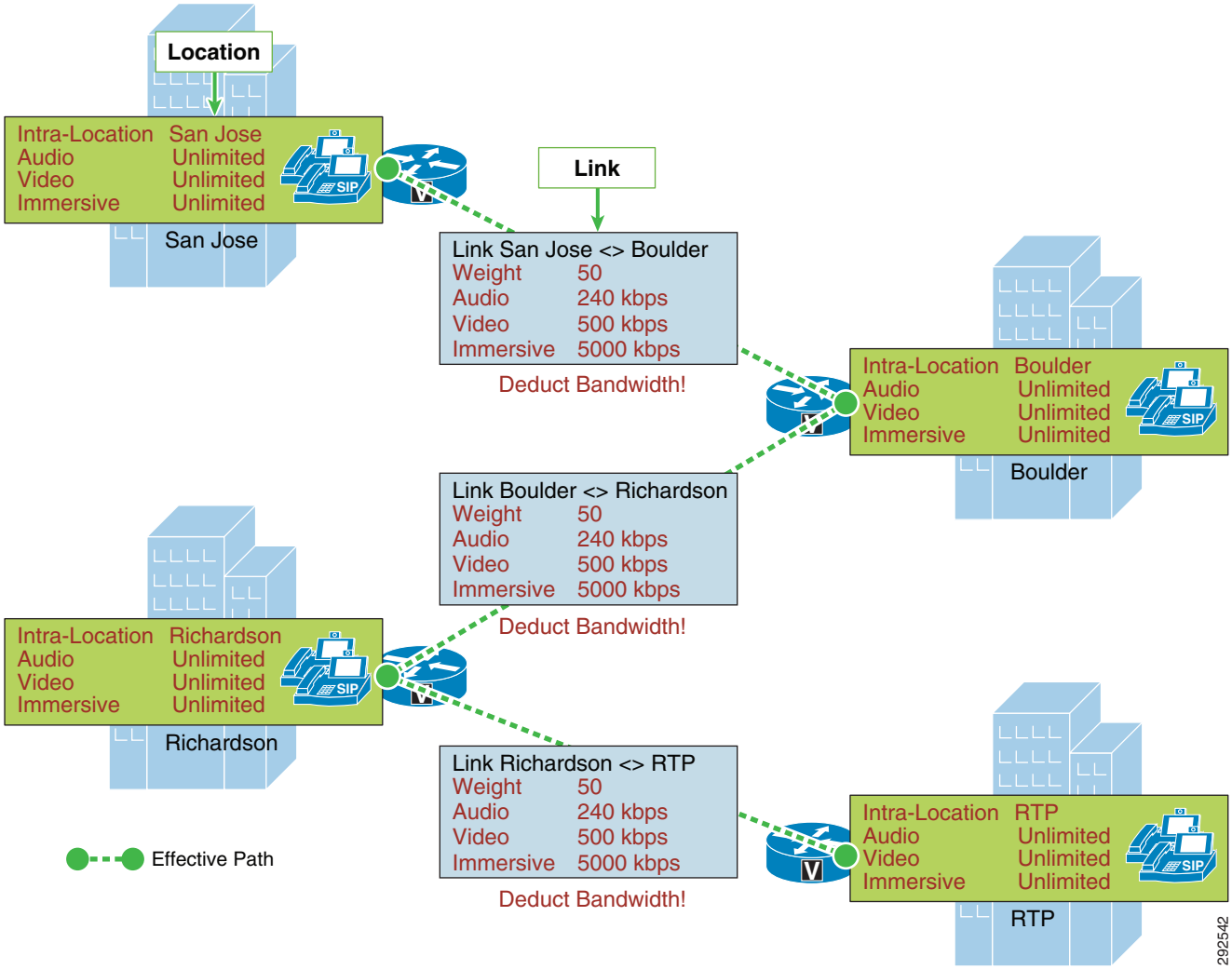
The location configuration itself consists of three main parts: links, intra-location bandwidth parameters, and RSVP locations settings. The RSVP locations settings are not considered here for Enhanced Locations CAC because they apply only to RSVP implementations. In the configuration, the link bandwidth 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 effectively limit the amount of bandwidth used on the LAN or transit location. In most networks today that consist of at least 100BASE-T or Gigabit LANs, there is little or no reason to limit bandwidth on those LANs. However, there are some deployments that can benefit from limiting high-bandwidth video calls. A

simple example might be an enterprise site with video deployed pervasively on the desktop and/or endpoints. If user calls are mostly all video-enabled, it is easy to see how a large number of 1 to 2 Mbps video calls might utilize a large percentage of available bandwidth on a LAN, and an administrator might consider limiting the number of video calls to a smaller percentage of that available LAN bandwidth. Keep in mind that this utilization might occur only during the busy hour of business or during a specific time of the year when specific traffic levels spike, and the bandwidth limit would be reached only during that time when it would be needed to ensure that the LAN is not over-subscribed with video traffic. It is also noteworthy to mention that video devices can be enabled to **Retry Video Call as Audio** if a video call to that device fails for any reason. This 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.

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 11-10. 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 11-10, 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.

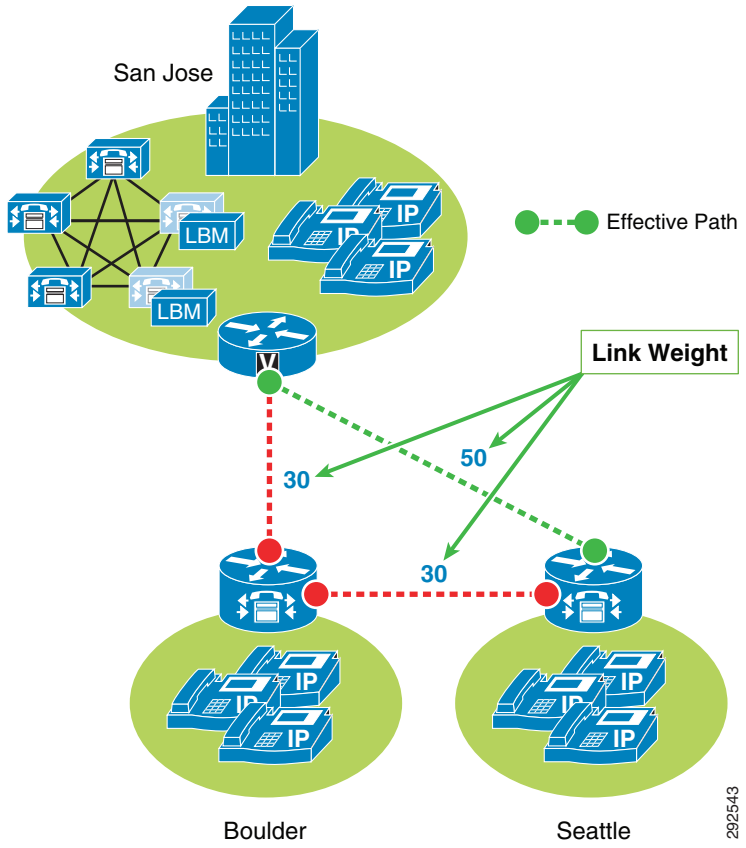
Figure 11-10 Simple Link Example with Three WAN Hops



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 11-18) 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 11-11 illustrates weight configured on links between three locations: San Jose, Boulder, and Seattle.

Figure 11-11 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

The Enhanced Locations 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 11-2 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 Layer 3 overhead. For example, a G.711 audio call consumes 80 kbps (64k bit rate + L3 overhead) 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 11-2 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 11-10 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 9-117.

Note AAR is invoked only when Enhanced Locations 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.

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 Locations 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 Locations 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 Locations call admission control.
- Intra-location bandwidth
  - 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 Locations 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 Locations 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 Locations call admission control.
- The settings for RSVP call admission control between locations — Possible settings are No Reservation, Optional, Optional (Video Desired), Mandatory, and Mandatory (Video Desired).

You can configure regions in Unified CM to define:

- The Max Audio Bit Rate used for intraregion calls
- The Max Audio Bit Rate used for interregion calls
- The Max Video Call Bit Rate (Includes Audio) used for intraregion and interregion calls. This also includes the maximum bit rate for immersive calls when applied to TelePresence endpoints.
- The link loss type for interregion calls (Possible link loss types are Low Loss and Lossy)

Unified CM Support for Locations and Regions

Cisco Unified Communications Manager supports 2,000 locations and 2,000 regions with Cisco MCS-7845 servers. 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 and Interregion Link Loss Type

When adding regions, you should select **Use System Default** for the Max Audio Bit Rate and Max Video Call Bit Rate values. If you are using RSVP call admission control, you should also select **Use System Default** for the RSVP parameter.

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 11-3 summarizes these limits.

Table 11-3 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 Max 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, change the default Interregion Max Audio Bit Rate to 64 kbps (G.722 or G.711), or else add a region for fax machines to each location with a non-default bit rate of 64 kbps (G.722 or G.711), subject to the limits in Table 11-3.

Note Irrespective of the MCS model you are using, your Cisco Partner or Cisco Systems Engineer should always use the Cisco Unified Communications Sizing Tool (http://tools.cisco.com/cucst) to validate all designs that incorporate a large number of remote sites, because there are many interdependent variables that can affect Unified CM cluster scalability (such as regions, locations, gateways, media resources, and so forth). Use the Sizing Tool to accurately determine the number of servers or clusters required to meet your design criteria.

Locations Bandwidth Manager

The Locations Bandwidth Manager (LBM) is a Unified CM Feature Service managed from the serviceability web pages and responsible for all of the Enhanced Locations CAC bandwidth functions. The LBM can run on any Unified CM subscriber or as a standalone service on a dedicated Unified CM server in the cluster. A minimum of one instance of LBM must run in each cluster to enable Enhanced Locations CAC in the cluster. For most installations, Cisco recommends the LBM are:

- Locations and links path assembly
- Bandwidth calculations over the effective paths in the assembly
- Servicing bandwidth requests from the Cisco CallManager service (Unified CM call control)
- Replication of bandwidth information to other LBMs within the cluster and between clusters when intercluster Enhanced Locations CAC is enabled
- Providing configured and dynamic information to serviceability
- Updating Location Real-Time Monitoring Tool (RTMT) counters
- Using Extensible Markup Language (XML) over TCP for communication to/from the Cisco CallManager service as well as between LBMs.

The LBM Service is enabled by default when upgrading to Cisco Unified CM 9.x from earlier releases. 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 (inter-location bandwidth data). 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 Locations CAC is configured and activated, the LBM replicates the assembled topology to other clusters (see Intercluster Enhanced Locations CAC, page 11-22, 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 11-12 illustrates LBM redundancy.

Figure 11-12 Locations Bandwidth Manager Redundancy

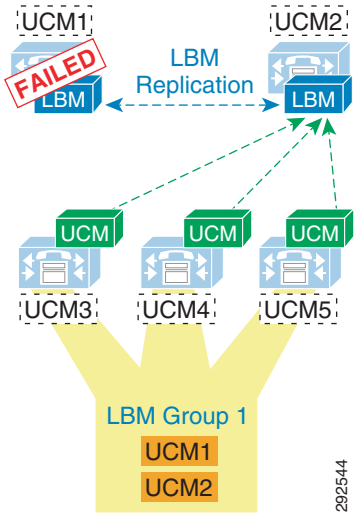


Figure 11-12 shows five Unified CM servers: UCM1 and UCM2 are dedicated LBM servers (only LBM service enabled); UCM3, UCM4, and UCM5 are Unified CM subscribers (Cisco CallManager service enabled). An LBM Group has been configured with UCM1 as active and UCM2 as standby, and it is applied to subscribers 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.

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

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

Enhanced Locations CAC Design and Deployment Recommendations and Considerations

- The Locations Bandwidth Manager (LBM) is a Unified CM Feature Service.
- LBM is responsible for modeling the topology and servicing Unified CM bandwidth requests.
- All LBMs are fully meshed within the cluster.

- The Enhanced Locations CAC LBM replication network is used to replicate the modeled topology as well as the bandwidth allocations within the cluster and across multiple clusters.
- Recommendations for LBM Group usage are as follows:
  - Manage how the Cisco CallManager service interacts with LBM (co-resident or dedicated).
  - Minimize LBM full-mesh bandwidth requirements in clustering over the WAN or dual data center deployments.
  - Deploy a minimum of two LBMs per call processing site for redundancy, either co-resident or dedicated.
  - Off-load active LBMs to inactive stand-by subscribers.
- Current recommendation is to deploy the LBM service co-resident with a Unified CM subscriber running the Cisco CallManager call processing service.

LBM Group Recommendations

- Configure each Unified CM subscriber to have a local LBM running and active.
- A minimum of two LBMs in a redundant LBM group configuration should be active at each call processing site, such as in clustering over the WAN designs.
- For load reduction of active subscribers, use dedicated LBMs or enable LBM on the inactive stand-by subscribers in 1:1 Unified CM redundancy models.

Intercluster Enhanced Locations CAC

Intercluster Enhanced Locations CAC extends the concept of network modeling across multiple clusters. In intercluster Enhanced Locations 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 11-13 illustrates the concept of a global topology with a simplistic hub-and-spoke network topology as an example.

Figure 11-13 Example of a Global Topology for a Simple Hub-and-Spoke Network

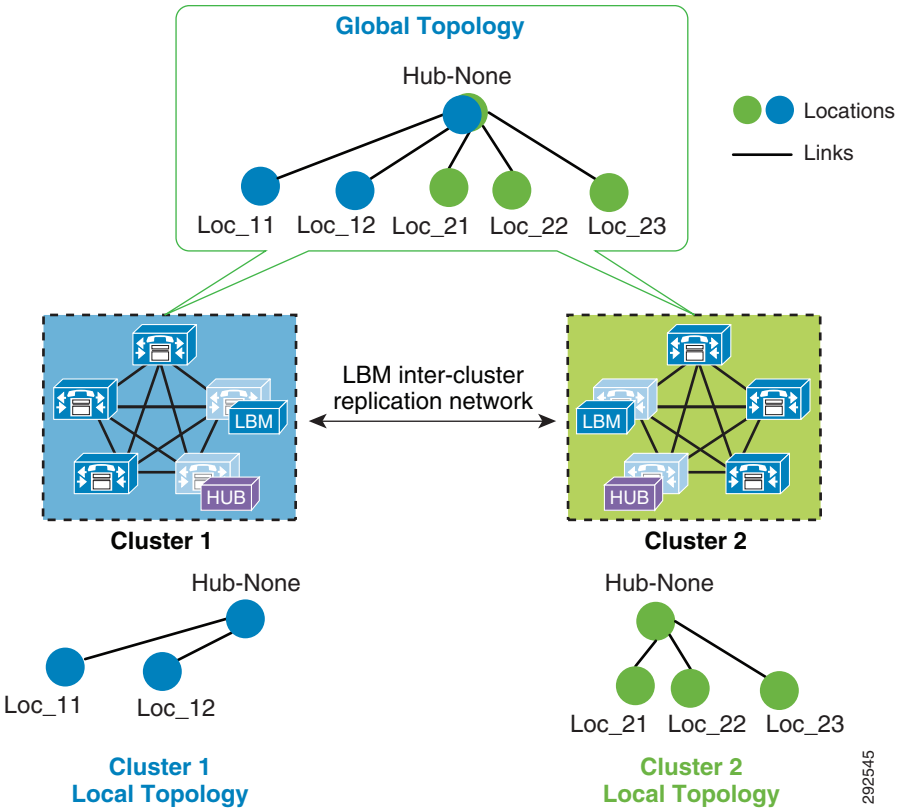


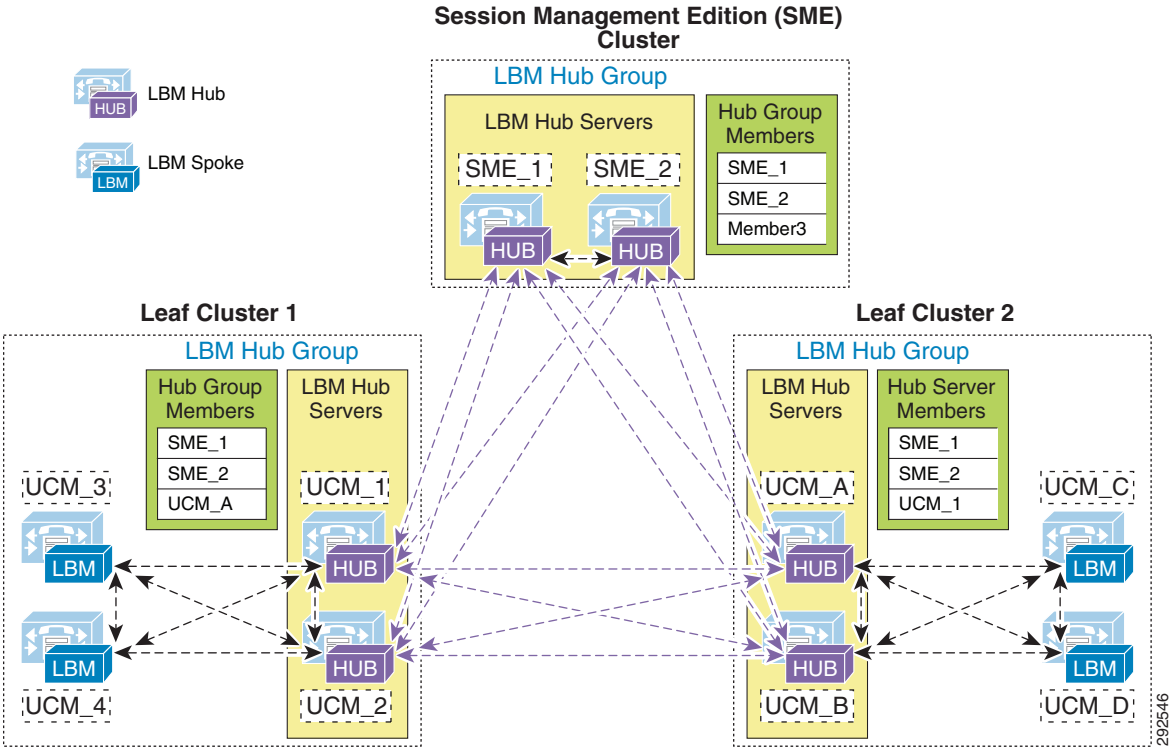
Figure 11-13 shows two clusters, Cluster 1 and Cluster 2, each with a locally configured hub-and-spoke network topology. Cluster 1 has configured Hub_None with links to Loc_11 and Loc_12, while Cluster 2 has configured Hub_None with links to Loc_21, Loc_22, and Loc_23. Upon enabling intercluster Enhanced Locations 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 network of designated LBMs that create a full-mesh with one another and replicate their local cluster's topology. In turn, each receives all remote clusters' topologies in order to create the global topology. The designated LBMs for the intercluster replication network are called LBM hubs, and the LBMs that replicate only within a cluster are called LBM spokes. The LBM hubs are designated in configuration through the LBM hub group. The LBM hub group has two main configuration areas called hub group members and hub group usage information. The hub group members are LBM hubs in remote clusters that are part of the LBM replication network. A maximum of three members can be configured. The members designated in the LBM hub group members serve as bootstrap servers for the entire intercluster replication network, providing each LBM

hub in each cluster with the connectivity details of other remote clusters with whom they are connected. The LBM Hub group usage information consists of the LBM hubs and spokes in the local cluster. Moving an LBM service into or out of the LBM Hub group determines the hub or spoke role. (See Cisco Unified Communications Manager product documentation for further information on the LBM hub group configuration.) Once the LBM hub group is configured on each cluster in the designated LBM, hubs will create the full mesh intercluster replication network. Figure 11-14 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.

Figure 11-14 Example Intercluster Replication Network for Three Cluster



In Figure 11-14, two LBM servers from each cluster have been designated as the LBM hubs for their cluster. These LBM hub servers form the intercluster LBM replication network. The LBM hub group members configured in each LBM hub group are designated as SME_1 and SME_2. These two LBM servers from the SME cluster serve as points of contact 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:

- LBM Hub group members
  - Remote hub servers responsible for interconnecting all LBM hubs in the replication network
  - Can be any hub in the network
  - Can indicate up to 3 per hub group
- LBM Hub Servers (Local LBMs)
  - Communicate directly to other remote hub servers as part of the intercluster LBM replication network
- LBM Spoke Servers (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
  - If a cluster has multiple LBM hubs, the LBM hub with the lowest IPv4 (entire) address will function as the sender of messages to other remote clusters. Only one hub per cluster will forward messages to remote clusters. This limits the amount of replication traffic in the intercluster replication network.
  - The LBM hub that functions as the sender for messages in the cluster selects one LBM hub from each cluster and forwards messages to that LBM.
  - The LBM hubs that receive messages from remote clusters, in turn forward the received messages to the LBM spokes in their local cluster.
  - Forwarded messages have a unique random string associated with them that allows receivers to determine if a messages has already been received and thus drop messages that they have received twice to prevent any replication storm or looping.
  - Other LBM hubs in the cluster that receive the forwarded message will not forward on to LBM spokes because the message is not directly from a remote cluster. This avoids hubs sending duplicate messages from remote clusters.

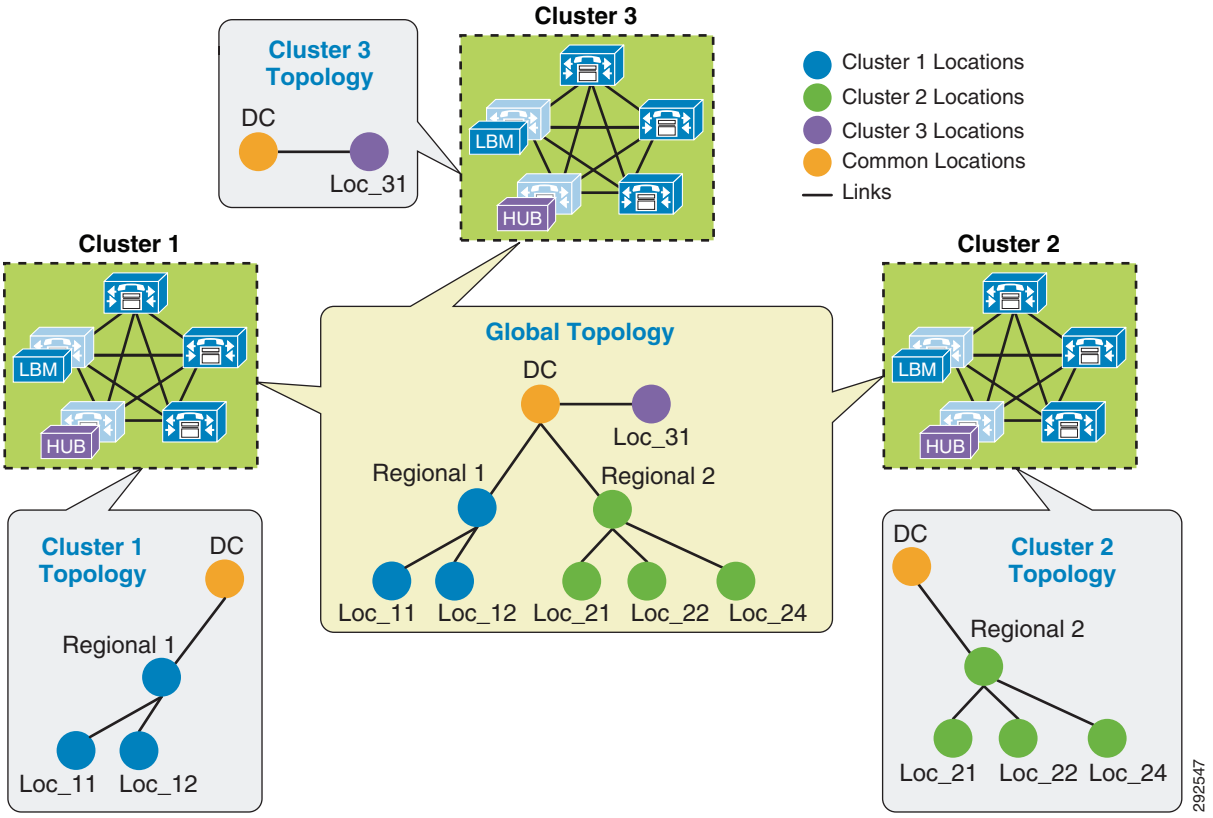
Common Locations (Shared Locations) and Links

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

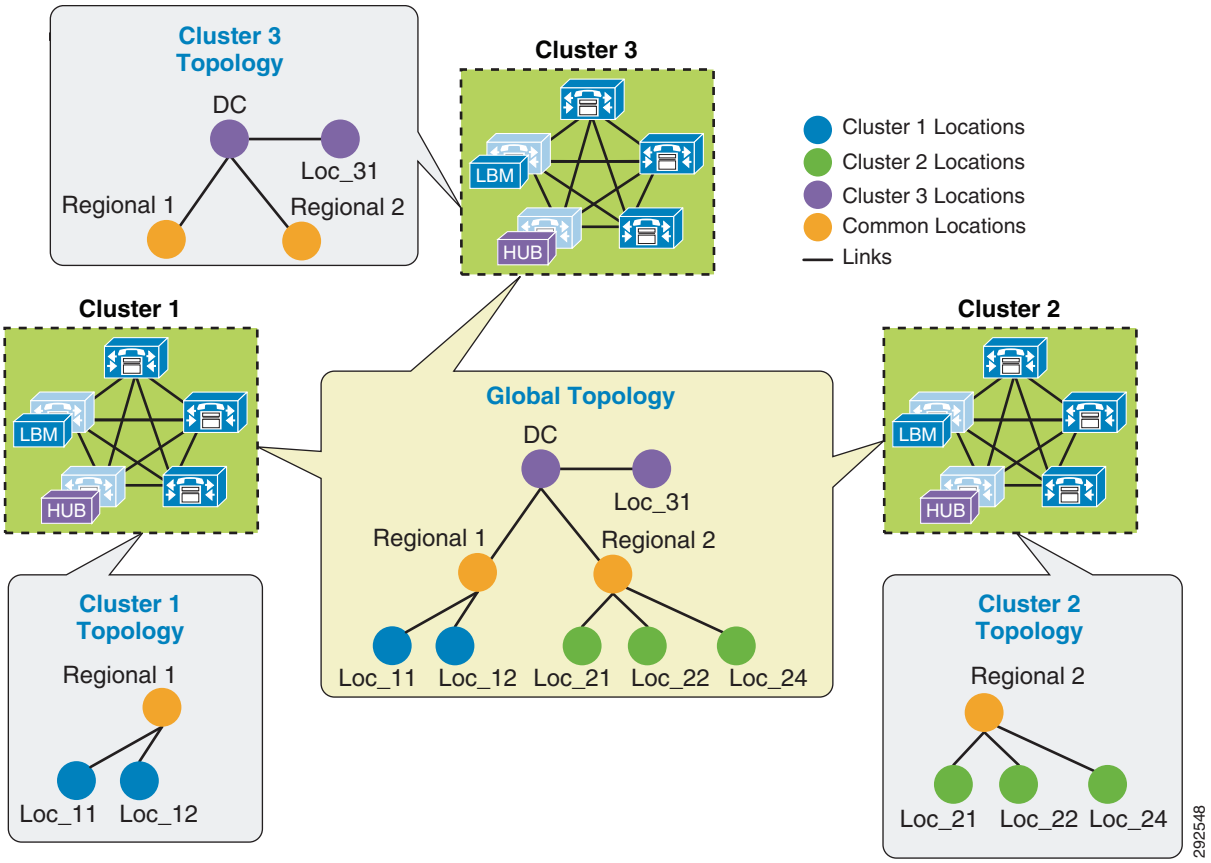
clusters have the common interconnecting locations and links on each cluster to link each remote topology together effectively. Figure 11-15 illustrates linking topologies together and shows the common topology that each cluster shares.

Figure 11-15 Using Common Locations and Links to Create a Global Topology



In Figure 11-15, 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 11-16.

Figure 11-16 Alternative Topology Using Different Common Locations



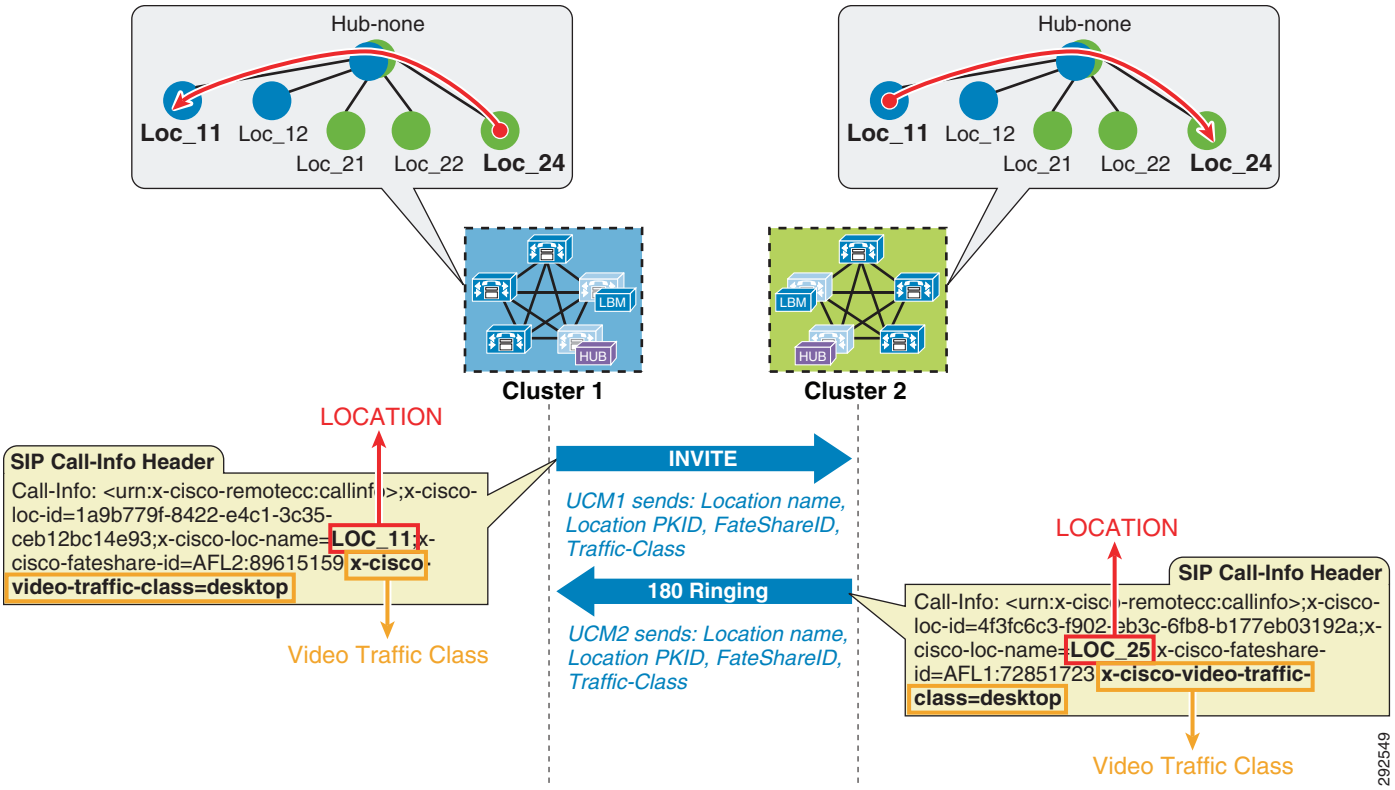
The key to topology mapping from cluster to cluster is to ensure that at least one cluster has a common location with another cluster so that the topologies interconnect accordingly.

Shadow Location

The *shadow location* is used to enable a SIP trunk to pass Enhanced Locations CAC information such as location name and Video-Traffic-Class (discussed below), among other things, required for Enhanced Locations CAC to function across clusters. In order to pass this location information across clusters, the SIP intercluster trunk (ICT) must be assigned to the "shadow" location. Similar to the "phantom" location, it 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 Locations 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 11-17 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 11-17 Shadow Location Used to Pass Information Between Clusters



In Figure 11-17, 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 locations 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 with unlimited bandwidth values and no configured links. When intercluster Enhanced Locations 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

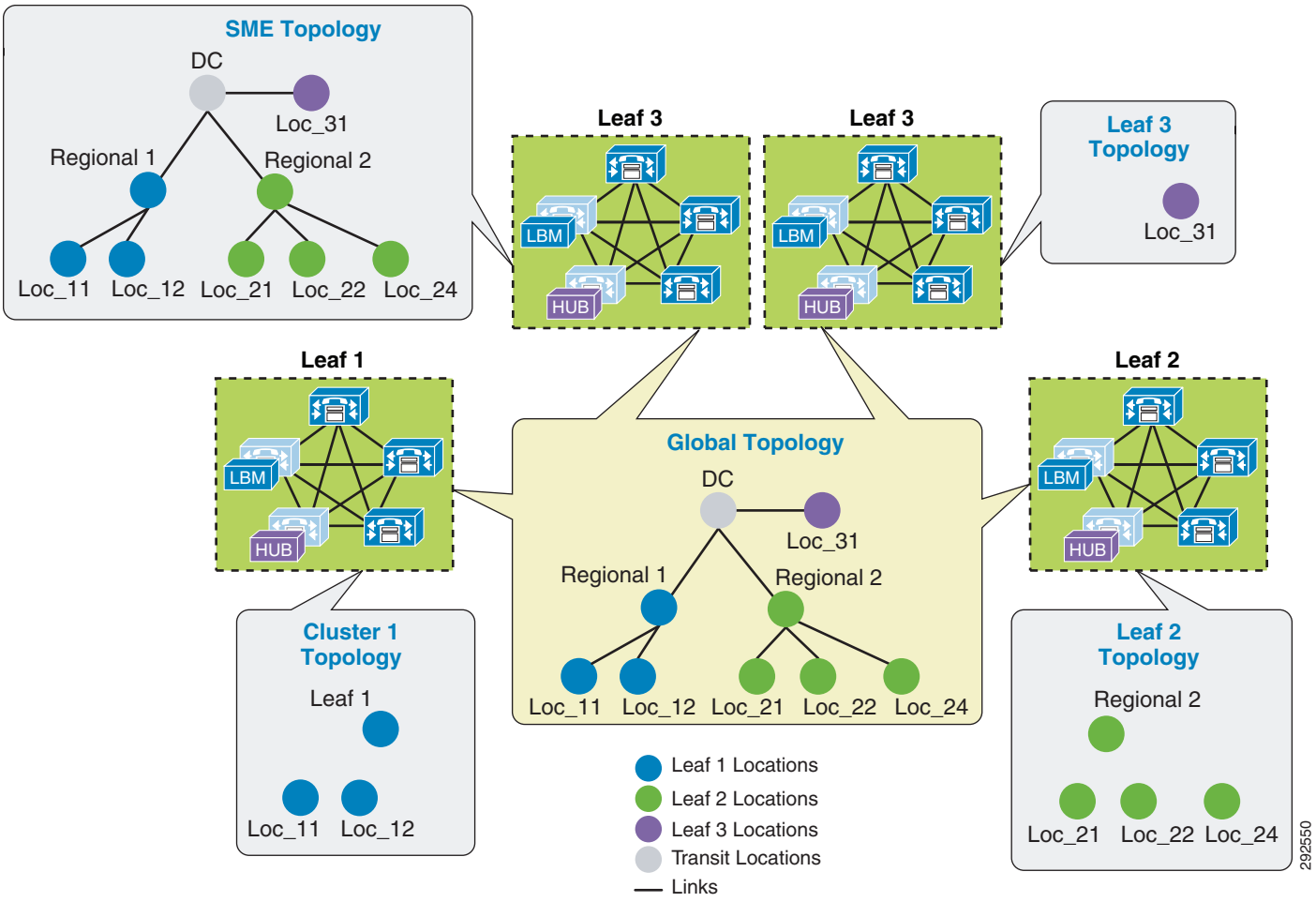
- Management cluster in the LBM replication network
  - All links and locations are managed in the management cluster.
  - Locations, bandwidth values, links, and weights
- Other clusters in the LBM replication network
  - Use unlimited bandwidth values on the intra-locations bandwidth parameters.
  - Do not configure links.
- 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 11-18 illustrates Cisco Unified Communications Manager Session Management Edition (SME) as a Location and Link Management Cluster for three leaf clusters.

Figure 11-18 Example of SME as a Location and Link Management Cluster



In Figure 11-18 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 Locations CAC Design and Deployment Recommendations and Considerations

- Oversubscription in bandwidth reservations can be incurred since reservations are made locally and replicated out to the rest of the LBM replication network. To avoid QoS impacts during oversubscription, observe the following guidelines:
  - Oversubscription is transient and will correct itself as the calls, using the oversubscribed locations and links in the path, clear and relinquish the bandwidth.
  - Bandwidth overhead should be provisioned in the QoS network policy to accommodate oversubscription. Cisco recommends over-provisioning by a minimum of one call of the highest bandwidth value in each QoS class (audio and video) for applicable locations and links. For audio-only implementations this may be a single call at 24 kbps or 80 kbps. For video implementations this may be the bit rate of a single video call.
  - Locations and links where CAC limits are often reached during the busy hour or in general, are prime candidates for over-provisioning of the QoS bandwidth capacity.
  - In cases of very high busy hour call completions (BHCC) and long delays between Unified CM clusters, Cisco recommends monitoring the locations and links to determine the amount of oversubscription during the busy hour, then ensure that the network is over-provisioned with an equal amount of bandwidth.
- 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 Locations CAC for TelePresence Immersive Video

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

Video Call Traffic Class

Video Call Traffic Class is a 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. SIP trunks can be configured through the SIP Profile as either desktop video, high definition immersive video, or a system that has both classifications of video endpoints, such as a Cisco TelePresence System Video Communications Server (VCS). All other endpoints and trunks have a non-configurable Video Call Traffic Class of **desktop video**.

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 Locations 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.

TelePresence Endpoints

TelePresence endpoints have a fixed non-configurable Video Call Traffic Class of **immersive** and are identified by Unified CM as immersive.

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 11-4.

Table 11-4 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 Trunks

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 Communications system such as Unified CM prior to version 9.0, 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 11-5.

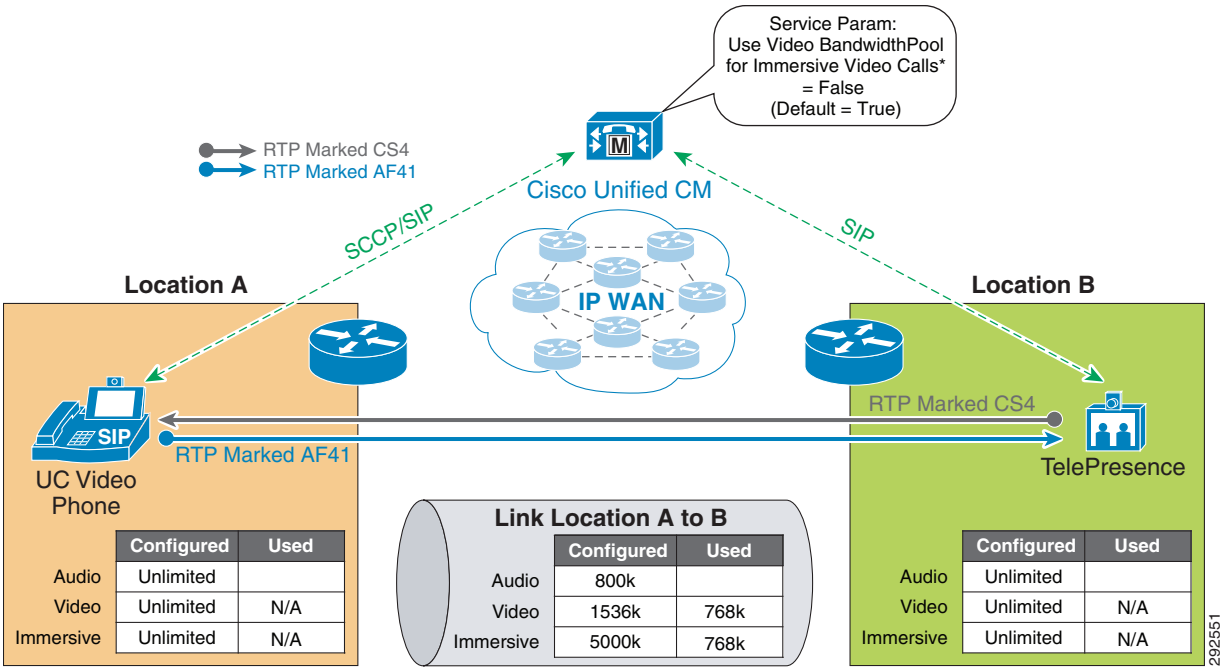
Table 11-5 Bandwidth Pool Usage per SIP Trunk and Endpoint Type

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 is deducted from the locations and links video bandwidth pool. To change this behavior, set the Unified CM service parameter **Use Video BandwidthPool for Immersive Video Calls** to **False**, and this will enable the immersive video bandwidth deductions.

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 11-19 illustrates this.

Figure 11-19 Enhanced Locations 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 11-20 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 11-20 Call Flow for End-to-End TelePresence Immersive Video

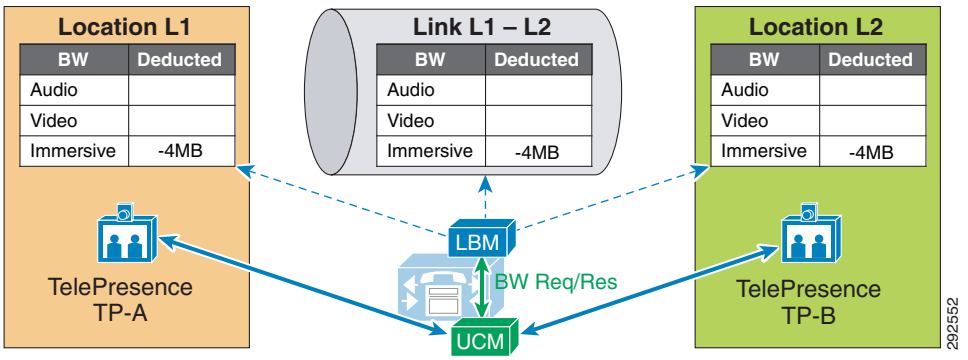


Figure 11-21 illustrates an end-to-end desktop video call between DP-A in Location L1 and DP-B in Location L2. End-to-end desktop video endpoint calls deduct bandwidth from the video bandwidth pool of the locations and the links along the effective path.

Figure 11-21 Call Flow for End-to-End Desktop Video

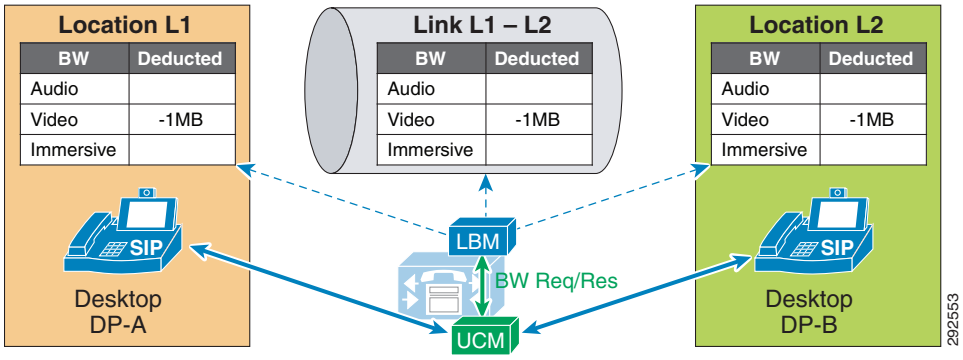
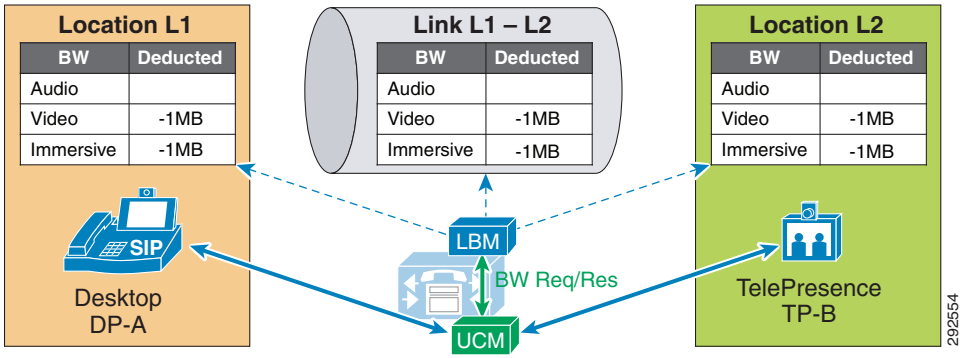


Figure 11-22 illustrates a video call between desktop video endpoint DP-A in Location L1 and TelePresence video endpoint TP-B in Location L2. Interoperating calls between desktop video endpoints and TelePresence video endpoints deduct bandwidth from both video and immersive locations and the links bandwidth pools along the effective path.

Figure 11-22 Call Flow for Desktop-to-TelePresence Video



In Figure 11-23, 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 11-23 Call Flow for a Video Conference with an MCU

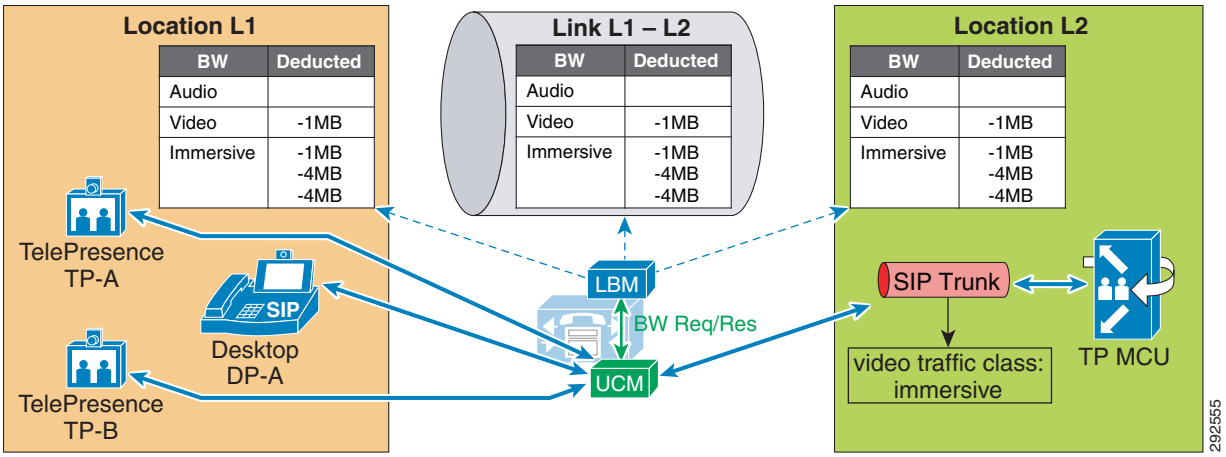


Figure 11-24 illustrates an end-to-end immersive video call across clusters, which deducts bandwidth from the immersive bandwidth pool of the locations and links along the effective path.

Figure 11-24 Call Flow for End-to-End TelePresence Immersive Video Across Clusters

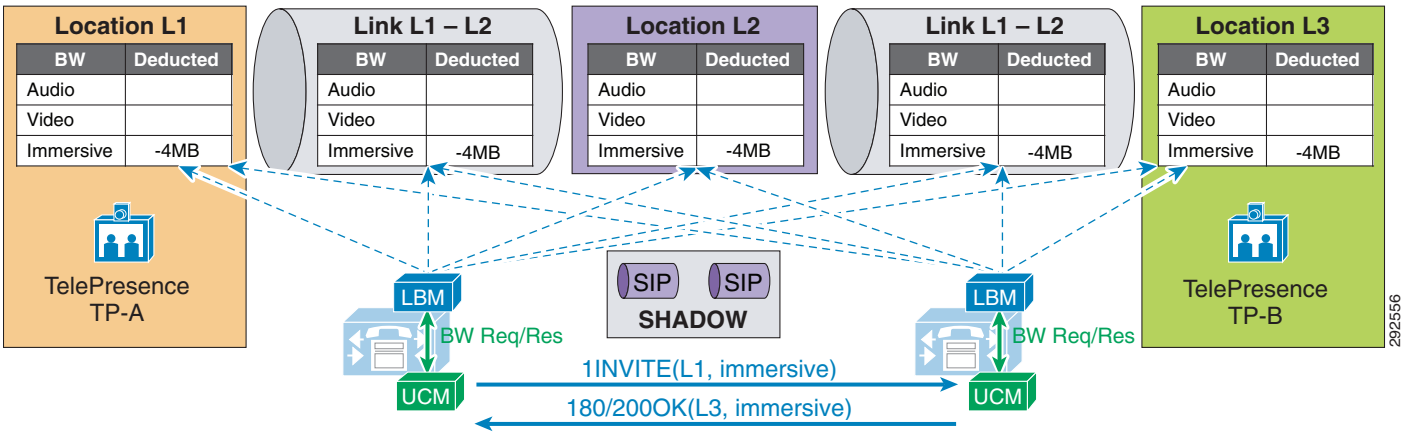


Figure 11-25 illustrates an end-to-end desktop video call across clusters, which deducts bandwidth from the video bandwidth pool of the locations and links along the effective path.

Figure 11-25 Call Flow for End-to-End Desktop Video Call Across Clusters

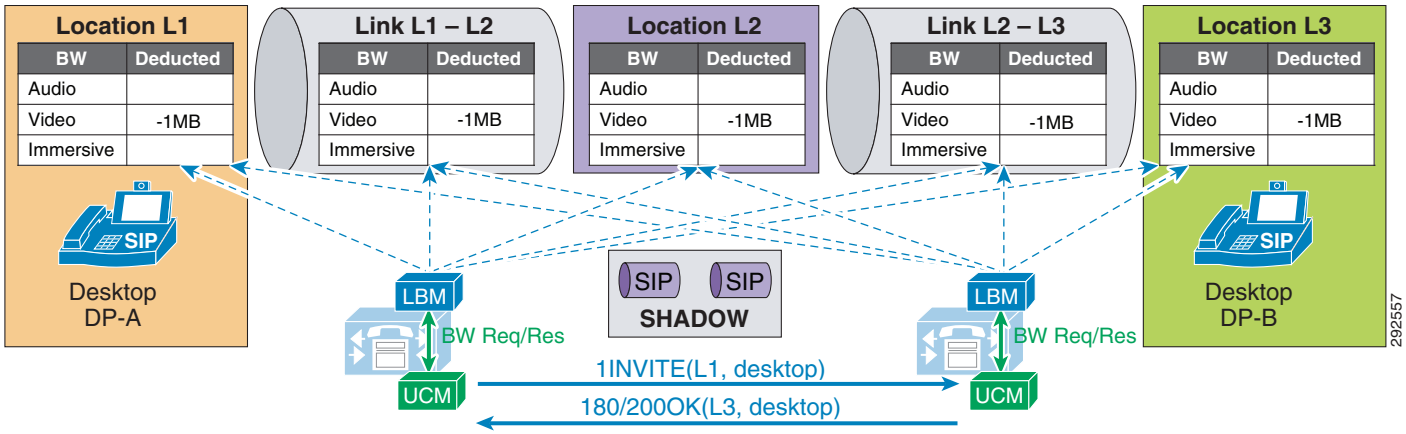
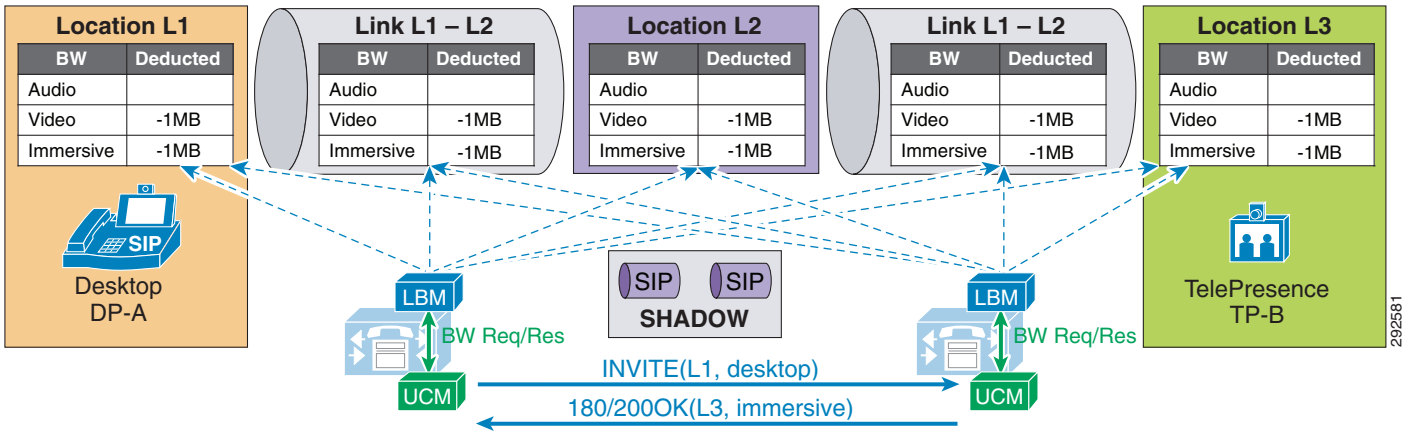


Figure 11-26 illustrates a desktop video endpoint calling a TelePresence endpoint across clusters. the call deducts bandwidth from both video and immersive bandwidth pools of the locations and links along the effective path.

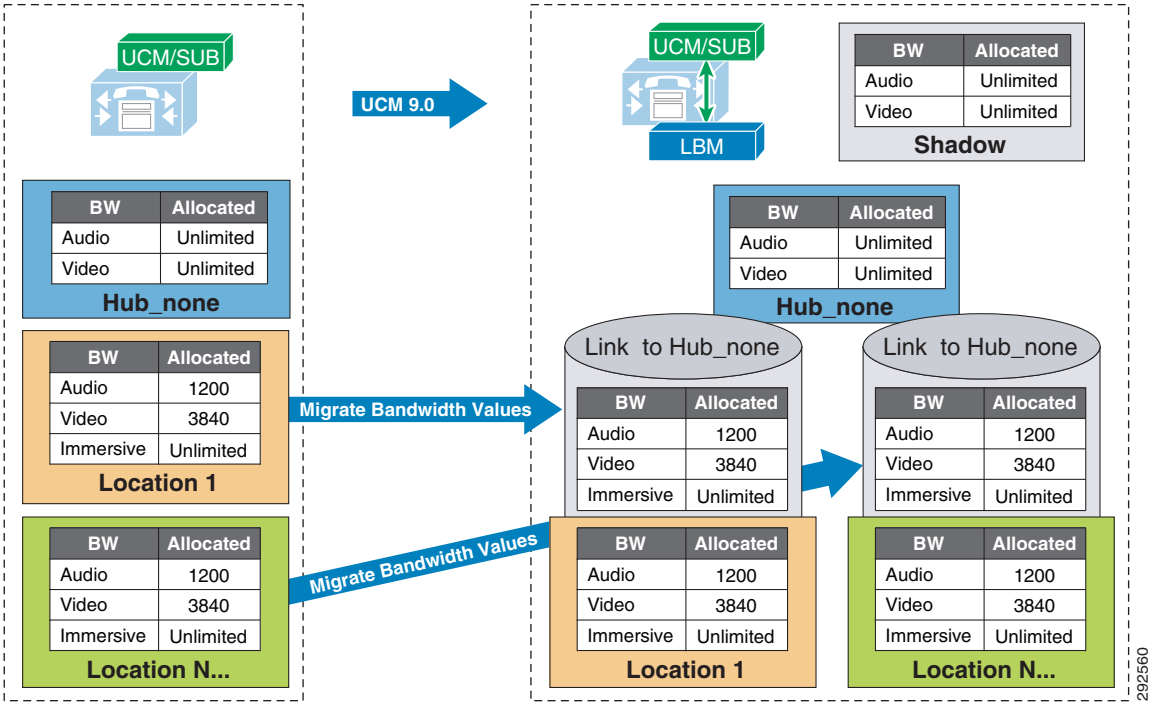
Figure 11-26 Call Flow for Desktop-to-TelePresence Video Across Clusters



Upgrade and Migration from Locations CAC to Enhanced Locations CAC

Upgrading to Cisco Unified CM 9.*x* from a previous release will result in the migration of Locations CAC to Enhanced Locations 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 Locations CAC supported. Figure 11-27 illustrates the migration of bandwidth information.

Figure 11-27 Migration from Locations CAC to Enhanced Locations CAC After Unified CM Upgrade



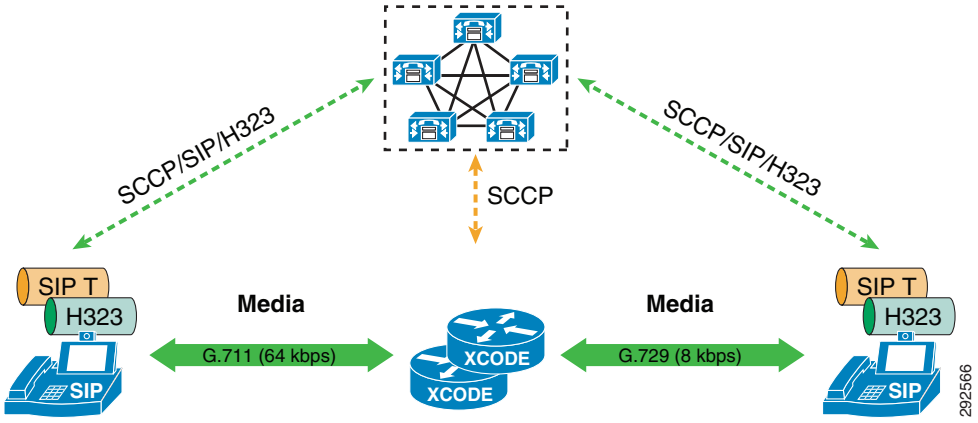
Settings after an upgrade to Cisco Unified CM 9.*x*:

- 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 Locations 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 Locations CAC bandwidth adjustment for MTPs and transcoders:

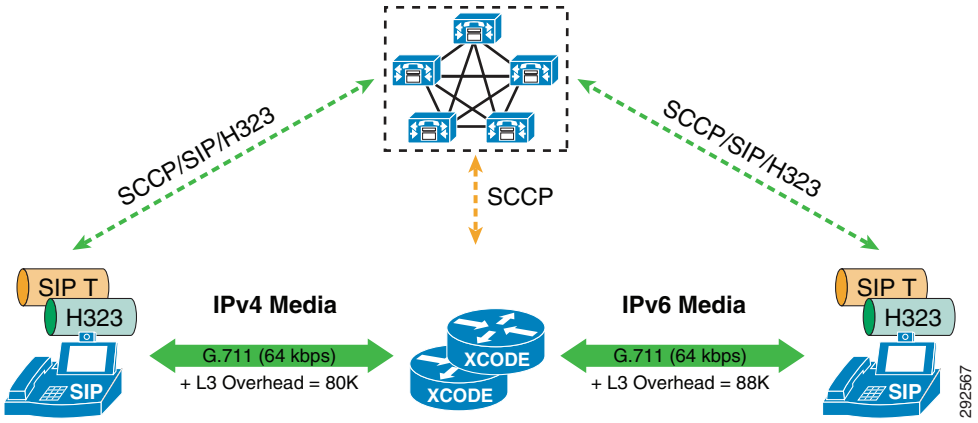
  For transcoding insertion, the bit rate is different on each leg of the connection. Figure 11-28 illustrates this.

Figure 11-28 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 11-29 illustrates the difference in bandwidth used for IPv4 and IPv6 networks with dual stack MTP insertion.

Figure 11-29 Bandwidth Differences for Dual Stack MTP Insertion



Enhanced Locations 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.

Cisco IOS Gatekeeper Zones

A Cisco IOS gatekeeper can provide call routing and call admission control between devices such as Cisco Unified CM, Cisco Unified Communications Manager Express (Unified CME), or H.323 gateways connected to legacy PBXs. It uses the H.323 Registration Admission Status (RAS) protocol to communicate with these devices and route calls across the network.

Gatekeeper call admission control is a policy-based scheme requiring static configuration of available resources. The gatekeeper is not aware of the network topology, so it is limited to simple hub-and-spoke topologies.

For a listing of the available Cisco IOS gatekeeper platforms and the features supported on each platform, refer to the *Cisco IOS H323 Gatekeeper Data Sheet* at

http://www.cisco.com/en/US/prod/collateral/voicesw/ps6788/vcallcon/ps4139/data_sheet_c78_561921.html

The call admission control capabilities of a Cisco IOS gatekeeper are based on the concept of gatekeeper *zones*. A zone is a collection of H.323 devices, such as endpoints, gateways, or Multipoint Control Units (MCUs), that register with a gatekeeper. There can be only one active gatekeeper per zone, and you can define up to 100 local zones on a single gatekeeper. A local zone is a zone that is actively handled by that gatekeeper – that is, all H.323 devices assigned to that zone register with that gatekeeper.

When multiple gatekeepers are deployed in the same network, a zone is configured as a local zone on only one gatekeeper. On the other gatekeepers, that zone is configured as a remote zone. This configuration instructs the gatekeeper to forward calls destined for that zone to the gatekeeper that "owns it" (that is, the gatekeeper on which that zone is configured as a local zone).

For details on configuring the gatekeeper, refer to the *Cisco IOS H.323 Configuration Guide* at

http://www.cisco.com/en/US/docs/ios/voice/h323/configuration/guide/15_0/vh_15_0_book.html

The bandwidth value deducted by the gatekeeper for every active call is double the bit-rate of the call, excluding Layer 2, IP, and RTP overhead. For example, a G.711 audio call that uses 64 kbps would be denoted as 128 kbps in the gatekeeper, and a 384-kbps video call would be denoted as 768 kbps. Table 11-6 shows the bandwidth values used by the gatekeeper feature for some of the most popular call speeds.

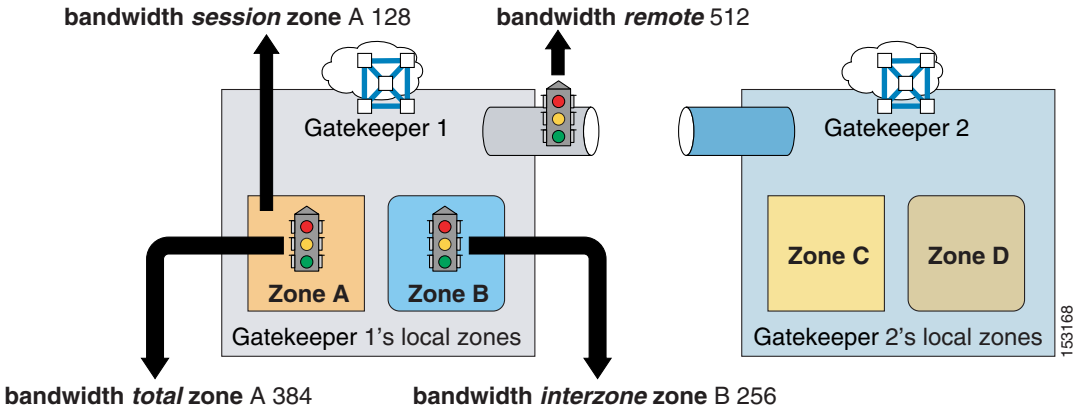
Table 11-6 Gatekeeper Bandwidth Settings for Various Call Speeds

Call Speed	Gatekeeper Bandwidth Value
G.711 audio call (64 kbps)	128 kbps
G.729 audio call (8 kbps)	16 kbps
128-kbps video call	256 kbps
384-kbps video call	768 kbps
512-kbps video call	1024 kbps
768-kbps video call	1536 kbps

Note Bandwidth calculations for the call Admission Request (ARQ) do not include compressed Real-Time Transport Protocol (cRTP) or any other transport overhead. See Bandwidth Provisioning, page 3-45, for details on how to provision interface queues.

To better understand the application of the **bandwidth** commands in a real network, consider the example shown in Figure 11-30.

Figure 11-30 Example of Cisco IOS Gatekeeper bandwidth Commands



Assuming that all calls are voice-only calls using the G.711 codec, and given the configuration commands shown in Figure 11-30, the following statements hold true:

- The maximum amount of bandwidth requested by any device in zone A for a single call is 128 kbps, which means that calls trying to use codecs with a higher bit-rate than 64 kbps will be rejected.
- The maximum amount of bandwidth used by all calls involving devices in zone A (either within the zone or with other zones) is 384 kbps, which means that there can be at most three active calls involving devices in zone A.
- The maximum amount of bandwidth used by all calls between devices in zone B and devices in any other zone is 256 kbps, which means that there can be at most two active calls between devices in zone B and devices in zones A, C, and D.
- The maximum amount of bandwidth used by all calls between devices registered with gatekeeper GK 1 and devices registered with any other gatekeeper is 512 kbps, which means that there can be at most four active calls between devices in zones A and B and devices in zones C and D.

Unified Communications Architectures Using Resource Reservation Protocol (RSVP)

This section covers the various Unified Communications architectures that implement Resource Reservation Protocol (RSVP) as the call admission control mechanism. The section begins with an introduction to RSVP and an overview of the protocol architecture, concepts of RSVP and Quality of Service, Application ID, and a summary of the infrastructure design considerations and recommendations.

Next this section discusses Unified CM RSVP-enabled locations in a single-cluster Unified CM environment. The discussion covers the components involved as well as the provisioning of those components, Unified CM's use of RSVP policy and Application ID, and a recommended migration strategy from call admission control based on Unified CM Enhanced Locations.

This section then covers distributed call processing architectures, beginning with RSVP SIP Preconditions, with an overview of the feature and how it works to synchronize the RSVP layer and call control layer between the various call control applications such as Unified CM, Unified CME, and SIP-TDM Cisco IOS Gateways. Then each call control application is discussed in further detail with regard to RSVP SIP Preconditions, including feature notes and design recommendations and considerations.

Resource Reservation Protocol (RSVP)

The Resource Reservation Protocol (RSVP) is the first significant industry-standard protocol for dynamically setting up end-to-end QoS across a heterogeneous network. RSVP, which runs over IP, was first introduced by the IETF in RFC 2205, and it enables an application to reserve network bandwidth dynamically. Using RSVP, applications can request a certain level of QoS for a data flow across a network. Because of its distributed and dynamic nature, RSVP is capable of reserving bandwidth across any network topology, therefore it can be used to provide topology-aware call admission control for voice and video calls.

This section focuses on the RSVP protocol principles and its interactions with the WAN infrastructure, specifically the QoS aspects, while the motivation and the mechanisms for call admission control based on RSVP are described in other sections of this chapter.

This section covers the following specific topics:

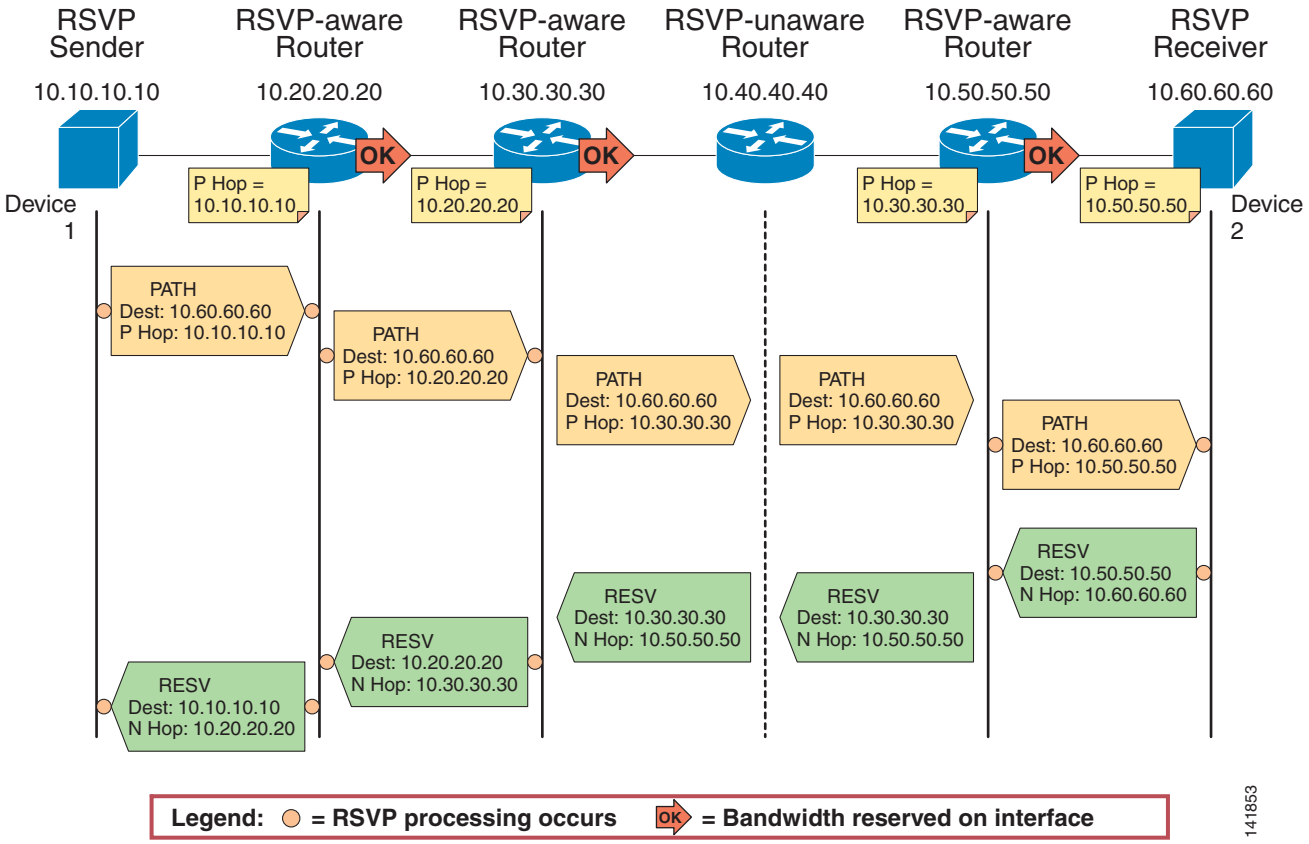
RSVP Principles

RSVP performs resource reservations for a given data flow across a network. RSVP reservations are unidirectional. Therefore, for a single audio call that contains two RTP streams, two RSVP reservations are generated, one for each RTP stream. The resource reservation is created by exchanging signaling messages between the source and destination devices for the data flow, and the messages are processed by intermediate routers along the path. The RSVP signaling messages are IP packets with the protocol number in the IP header set to 46, and they are routed through the network according to the existing routing protocols.

Not all routers on the path are required to support RSVP because the protocol is designed to operate transparently across RSVP-unaware nodes. On each RSVP-enabled router, the RSVP process intercepts the signaling messages and interacts with the QoS manager for the router's outbound interface involved in the data flow in order to "reserve" bandwidth resources. When the available resources are not sufficient for the data flow anywhere along the path, the routers signal the failure back to the application that originated the reservation request.

The principles of RSVP signaling can be explained by using the example shown in Figure 11-31. In this diagram, an application wishes to reserve network resources for a data stream flowing from Device 1, whose IP address is 10.10.10.10, to Device 2, whose IP address is 10.60.60.60.

Figure 11-31 Example of RSVP Path and Resv Message Flow



The following steps describe the RSVP signaling process for as single data flow, as shown by the example in Figure 11-31:

1. The application residing on Device 1 originates an RSVP message called Path, which is sent to the same destination IP address as the data flow for which a reservation is requested (that is, 10.60.60.60) and is sent with the "router alert" option turned on in the IP header. The Path message contains, among other things, the following objects:
   - The "session" object, consisting of destination IP address, protocol number, and UDP/TCP port, which is used to identify the data flow in RSVP-enabled routers.
   - The "sender T-Spec" (traffic specification) object, which characterizes the data flow for which a reservation will be requested. The T-Spec basically defines the maximum IP bandwidth required for a call flow using a specific codec. The T-Spec is typically defined using values for the data flow's average bit rate, peak rate, and burst size. Details of the T-Spec are discussed later in this chapter.
   - The "P Hop" (or previous hop) object, which contains the IP address of the router interface that last processed the Path message. In this example, the P Hop is initially set to 10.10.10.10 by Device 1.
2. By means of the "router alert" option, the Path message is intercepted by the CPU of the RSVP-aware router identified as 10.20.20.20 in Figure 11-31, which sends it to the RSVP process. RSVP creates a path state for this data flow, storing the values of the session, sender Tspec, and

P Hop objects contained in the Path message. Then it forwards the message downstream, after having replaced the P Hop value with the IP address of its outgoing interface (10.20.20.20 in this example).

3. Similarly, the Path message is intercepted by the CPU of the following RSVP-aware router, identified as 10.30.30.30 in Figure 11-31. After creating the path state and changing the P Hop value to 10.30.30.30, this router also forwards the message downstream.
4. The Path message now arrives at the RSVP-unaware router identified as 10.40.40.40 in Figure 11-31. Because RSVP is not enabled on this router, it just routes this message according to the existing routing protocols like any other IP packet, without any additional processing and without changing the content of any of the message objects.
5. Therefore, the Path message gets to the RSVP-aware router identified as 10.50.50.50, which processes the message, creates the corresponding path state, and forwards the message downstream. Notice that the P Hop recorded by this router still contains the IP address of the last RSVP-aware router along the network path, or 10.30.30.30 in this example.
6. The RSVP Receiver at Device 2 receives the Path message with a P Hop value of 10.50.50.50, and it can now initiate the actual reservation by originating a message called Resv. For this reason, RSVP is known as a receiver-initiated protocol. The Resv message carries the reservation request hop-by-hop from the receiver to the sender, along the reverse paths of the data flow for the session. At each hop, the IP destination address of the Resv message is the IP address of the previous-hop node, obtained from the path state. Hence, in this case Device 2 sends the Resv message with a destination IP address of 10.50.50.50. The Resv message contains, among other things, the following objects:
   - The "session" object, which is used to identify the data flow.
   - The "N Hop" (or next hop) object, which contains the IP address of the node that generated the message. In this example, the N Hop is initially set to 10.60.60.60 by Device 2.
7. When RSVP-aware router 10.50.50.50 receives the Resv message for this data flow, it matches it against the path state information using the received session object, and it verifies if the reservation request can be accepted based on the following criteria:
   - Policy control — Is this user and/or application allowed to make this reservation request?
   - Admission control — Are there enough bandwidth resources available on the relevant outgoing interface to accommodate this reservation request?
8. In this case, we assume that both policy and admission control are successful on 10.50.50.50, which means that the bandwidth provided by the Tspec in the path state for this session is reserved on the outgoing interface (in the same direction as the data flow, that is from Device 1 to Device 2), and a corresponding "reservation state" is created. Now router 10.50.50.50 can send a Resv message upstream by sending it as a unicast IP packet to the destination IP address stored in the P Hop for this session, which was 10.30.30.30. The N Hop object is also updated with the value of 10.50.50.50.
9. The Resv message now transits through the RSVP-unaware router identified as 10.40.40.40, which will route it toward its destination of 10.30.30.30 like any other IP packet. This mechanism allows RSVP signaling to work across a heterogeneous network where some nodes are not RSVP-enabled.
10. The RSVP-aware router identified as 10.30.30.30 receives the Resv message and processes it according to the mechanisms described in steps 7 and 8. Assuming policy and admission control are successful also at this hop, the bandwidth is reserved on the outgoing interface and a Resv message is sent to the previous hop, or 10.20.20.20 in this example.
11. After a similar process within the router identified as 10.20.20.20, the Resv finally reaches the RSVP sender, Device 1. This indicates to the requesting application that an end-to-end reservation has been established and that bandwidth has been set aside for this data flow in all RSVP-enabled routers across the network.

This example shows how the two main RSVP signaling messages, Path and Resv, travel across the network to establish reservations. Several other messages are defined in the RSVP standard to address error situations, reservation failures, and release of resources. In particular, the ResvErr message is used to signal failure to reserve the requested resources due to either policy control or admission control somewhere along the network. If, for example, admission control had failed at node 10.50.50.50 in Figure 11-31, this node would have sent a ResvErr message back to Device 2, specifying the cause of the failure, and the application would have been notified.

Another important aspect of the RSVP protocol is that it adopts a soft-state approach, which means that for each session both the path state and the reservation state along the network need to be refreshed periodically by the application by sending identical Path and Resv messages. If a router does not receive refresh messages for a given session for a certain period of time, it deletes the corresponding state and releases the resources reserved. This allows RSVP to react dynamically to network topology changes or routing changes due to link failures. The reservations simply start flowing along the new routes based on the routing protocol decisions, and the reservations along the old routes time-out and are eventually deleted.

RSVP in MPLS Networks

In some MPLS service-provider networks, the IP addresses used on the links between the customer edge (CE) and the provider edge (PE) are not distributed to the rest of the MPLS network, thus ensuring that the subnets stay local to the PE and are not advertised beyond the PE (because they are not unique and are being reused elsewhere). This creates a situation where RSVP is not able to forward RSVP messages because the P Hop (Previous Hop) value of the RSVP message is unknown in the network. Figure 11-32 illustrates this type of situation.

Figure 11-32 RSVP Over MPLS Without P Hop Overwrite

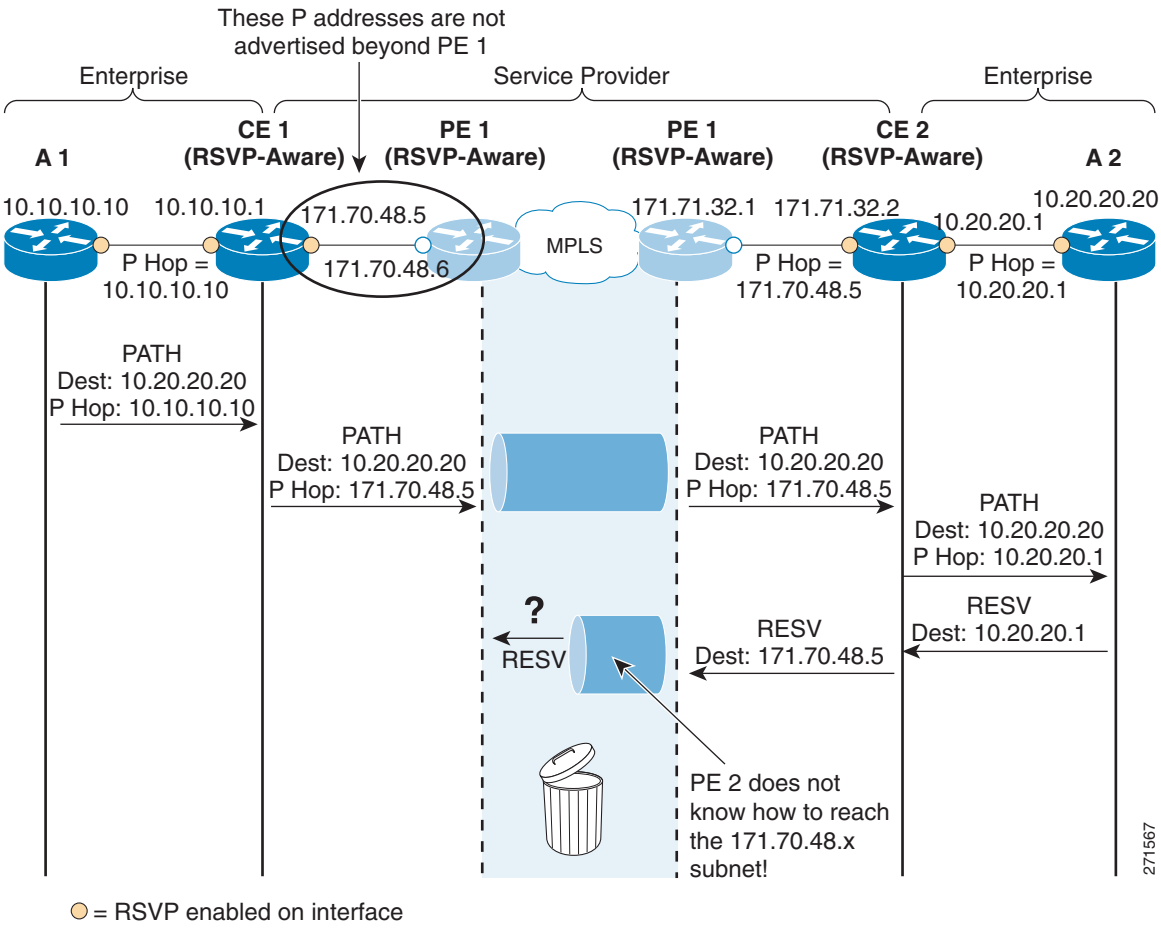


Figure 11-32 shows an enterprise network and a service provider MPLS network. CE1 and CE2 are RSVP-aware, and PE1 and PE2 are RSVP-unaware. The RSVP Path message contains a P Hop object. This object is rewritten at every RSVP hop. Its purpose is to enable an RSVP router (for example, CE1) to send a Path message to the next RSVP router (for example, CE2) to indicate that it (CE1) is the previous RSVP hop (or P Hop). This information is used by CE2 to forward the corresponding Resv message upstream hop-by-hop toward the sender.

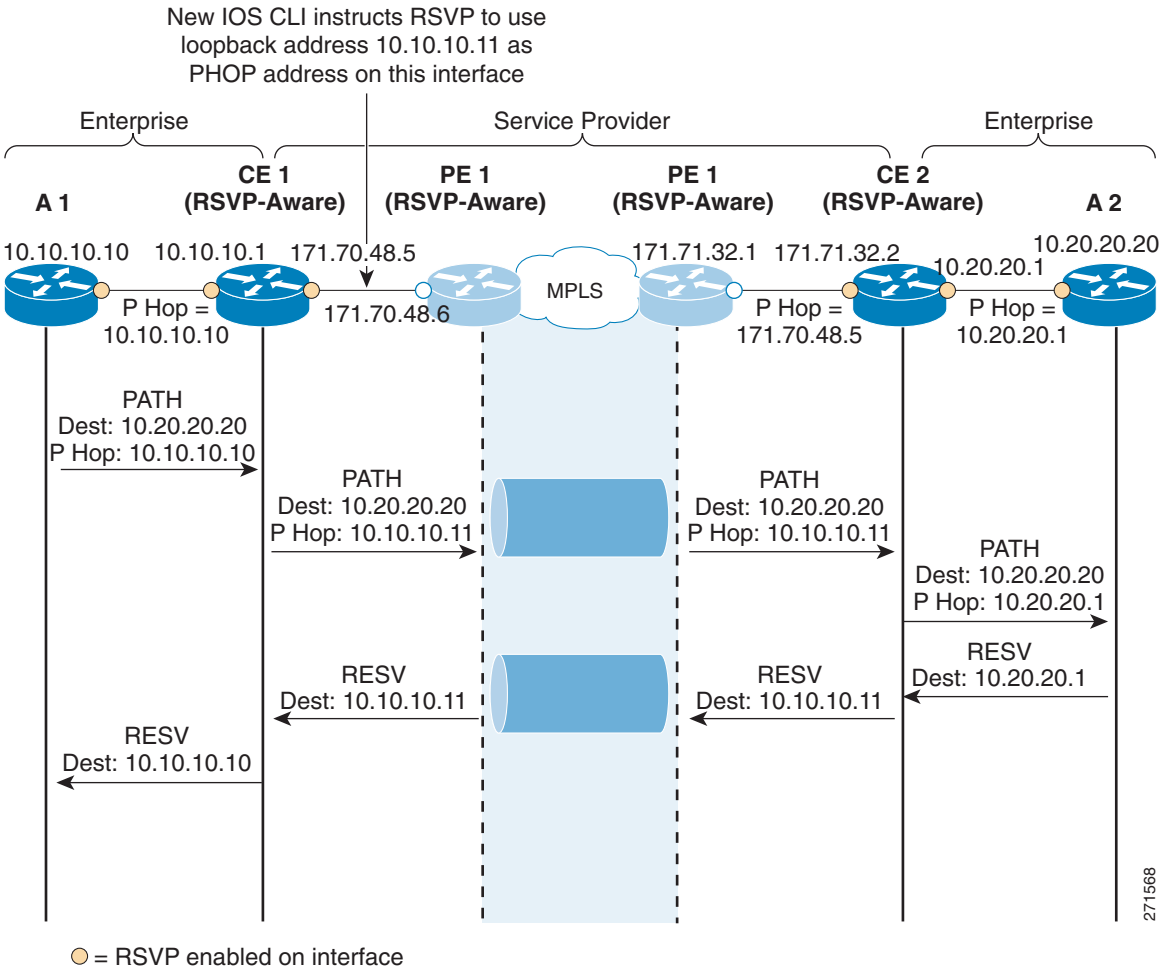
In Cisco IOS, the RSVP Router always sets the P Hop address to the IP address of the egress interface onto which it transmits the Path message. There are situations where, although some IP addresses of CE1 are reachable, the IP address of its egress interface is not reachable from a remote RSVP Router CE2. The result is that the corresponding Resv message generated by CE2 never reaches CE1, thus the reservation is never established.

When a call is made from A1 to A2, A1 tries to set up an RSVP session and starts by sending a Path message to CE1. A1 will populate the P Hop object in the Path message of its outgoing interface IP (in this case, 10.10.10.10). CE1 will then receive the Path message, process it, create the corresponding path state, update the P Hop field of the message with its egress interface IP address (171.70.48.5), which is not a routable IP address, and forward the Path message downstream. This Path message will be tunneled across the service provider network and will be processed by CE2. Upon reception of the Path message, CE2 records the IP address of the P Hop object (CE1's egress interface IP address) and forwards the Path message downstream to A1. A1 will record and process the Path message and initiate an RSVP message

to CE2. CE2 will process the RSVP message and send it's own RSVP message upstream to CE1. However, when CE2 replies with this Resv message, it will try to send it to the IP address that it had recorded earlier from the Path message received from CE1. Since this IP address (171.70.48.5) is not routable from CE2, the Resv message will fail, thus causing the reservation attempt to fail.

To resolve this behavior, a feature called Previous Hop Overwrite has been introduced in Cisco IOS Release 12.4.(20)T. P Hop Overwrite is a mechanism whereby the CE populates the Hop object in the Path message with an IP address from another interface on the router that is reachable in the customer VPN. In this way, the Resv message can find its way back toward the sender and reservations can be established. The P Hop Overwrite mechanism is illustrated in Figure 11-33.

Figure 11-33 RSVP P Hop Overwrite Feature in Cisco IOS 12.4(20)T



Describing Data Flow Characteristics in RSVP (TSpec)

RSVP was designed to support requesting Quality of Service (QoS) for any traffic flow, not just voice or video, across a wide range of Layer 2 technologies. To accomplish this, RSVP must be able to describe in detail the traffic flow for which it is requesting QoS, so that the intermediate routers can make admittance decisions correctly.

The bandwidth requirements for data flows for an RSVP session are characterized by senders in the TSpec (traffic specification) contained in Path messages and are mirrored in the RSpec (reservation specification) sent by receivers in Resv messages. The TSpec gets transported through the network to all intermediary routers and to the destination endpoint. The intermediate routers do not change this object, and the object gets delivered unchanged to the ultimate receiver(s).

The TSpec object contains the following elements:

- AverageBitRate (kbps)
- BurstSize (bytes)
- PeakRate (kbps)

Audio TSpec

For audio flows, the TSpec calculations specify the following measurements:

- AverageBitRate (kbps) — Including IP overhead
- BurstSize (bytes) — This value is calculated as the size of the packet times the number of packets in a burst. For audio flows, the burst usually specifies 1 to 2.
- PeakRate (bytes) — The peak rate, in bytes, refers to the maximum bytes per second that the endpoint transmits at any given time. If the burst is small, as is the case in audio streams, the peak rate can be calculated as 1.1 (or 1.2) times the tokenRate.

To avoid adjusting the bandwidth reservation upward when the call gets answered, Cisco Unified CM reserves the maximum bandwidth for each region codec at call setup time. Unified CM then modifies or adjusts the bandwidth based on the media capability of the connected parties when the call gets answered.

For more information on RSVP for Unified Communications, refer to the *Cisco Unified Communications Manager System Guide*, available at

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

Note This section focuses on providing an overview of RSVP principles and mechanisms. For more information on protocol behavior and extensions, complete message formats, and interactions with other protocols, refer to the numerous RFC documents related to RSVP, available at http://www.ietf.org.

RSVP and QoS in WAN Routers

RSVP has been supported in Cisco routers for many years, however most configurations recommended in this document are based on the RSVP Scalability Enhancements feature, which was first introduced in Cisco IOS Release 12.2(2)T.

By issuing the following Cisco IOS command in interface configuration mode on each Cisco IOS router interface, you can enable RSVP and define the maximum amount of bandwidth that it can control:

```
ip rsvp bandwidth [interface-kbps] [single-flow-kbps]
```

The *interface-kbps* parameter specifies the upper limit of bandwidth that RSVP can reserve on the given interface, while the *single-flow-kbps* parameter provides an upper bandwidth limit for each individual reservation (so that flows with higher bandwidth requests will be rejected even if there is bandwidth available on the interface).

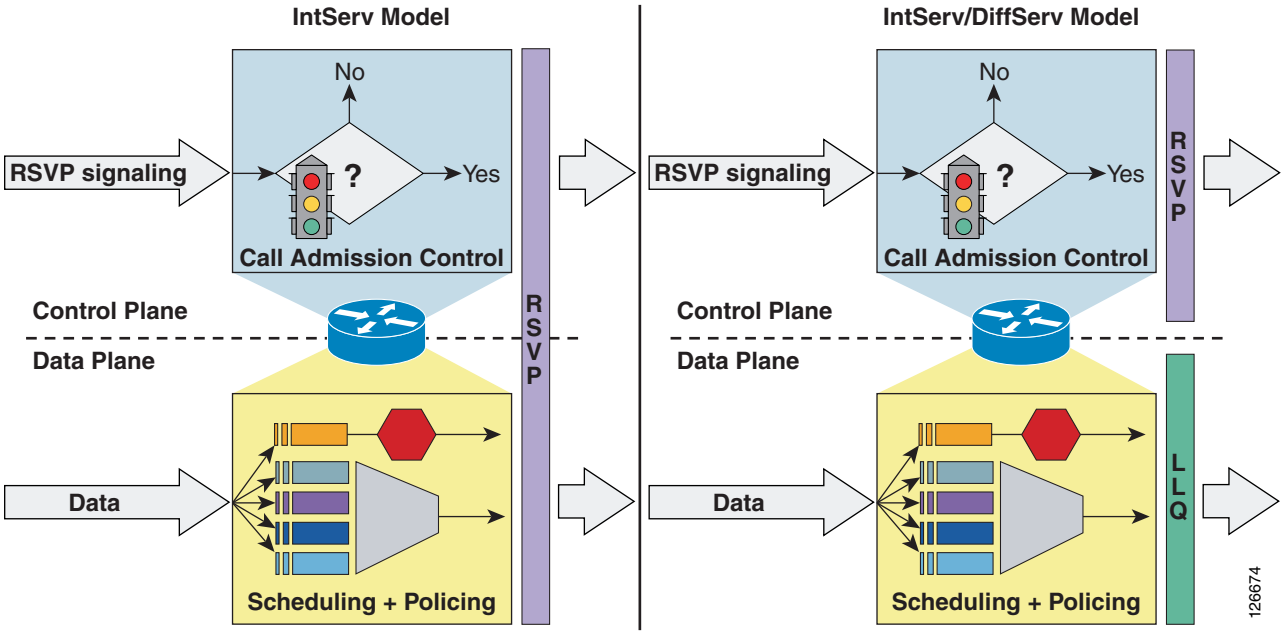
Note When RSVP is enabled on a router interface, all other interfaces in the router will drop RSVP messages unless they are also enabled for RSVP. To avoid dropping RSVP messages, enable RSVP on all interfaces through which you expect RSVP signaling to transit. If call admission control is not desired on an interface, set the bandwidth value to 75% of the interface bandwidth.

Within Cisco IOS, RSVP can be configured to operate according to two different models: the Integrated Services (IntServ) model, described in RFC 2210, or the Integrated Services/Differentiated Services (IntServ/DiffServ) model, described in RFC 2998. Both RFC documents are available on the IETF website at

http://www.ietf.org

Figure 11-34 shows the difference between these two approaches from the perspective of a Cisco IOS router.

Figure 11-34 The Two RSVP Operation Models: IntServ and IntServ/DiffServ

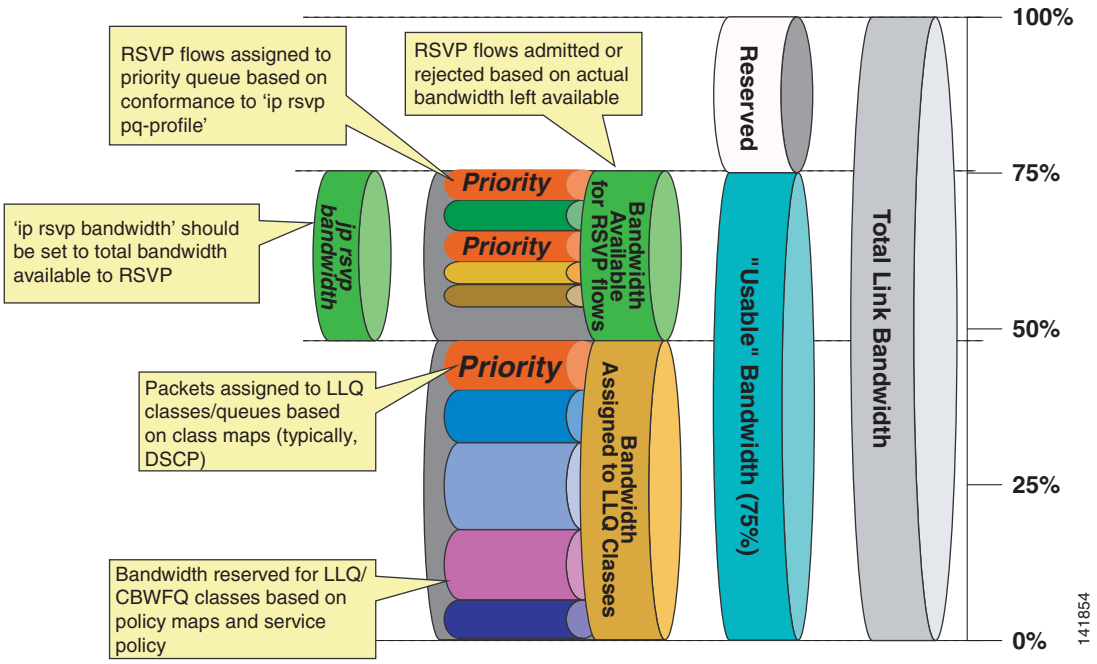


The IntServ Model

As shown on the left side of Figure 11-34, RSVP in the IntServ model involves both the control plane and the data plane. In the control plane, RSVP admits or denies the reservation request. In the data plane, it classifies the data packets, polices them based on the traffic description contained in the RSVP messages, and queues them in the appropriate queue. The classification that RSVP performs is based on the 5-tuple consisting of the source IP address, source port, destination IP address, destination port, and protocol number. In this model, all data packets transiting through the router must be intercepted by RSVP so that RSVP can inspect the 5-tuple and look for a match among the established reservations. If a match is found, the packets are scheduled and policed by RSVP according to the reservation's traffic specification.

As shown in Figure 11-35, when you combine the IntServ model with Low Latency Queuing (LLQ), the usable bandwidth is divided between RSVP and the predefined LLQ queues. RSVP controls the entrance criteria to the RSVP reserved bandwidth, while policy maps control the entrance criteria for the predefined queues.

Figure 11-35 Combining the IntServ Model with LLQ



To use the IntServ operation model on a Cisco IOS router, use the following commands in interface configuration mode:

```
ip rsvp resource-provider wfq [interface | pvc]
no ip rsvp data-packet classification
```

When these commands are active, RSVP admits or rejects new reservations, not only based on the upper bandwidth limit defined within the **ip rsvp bandwidth** command, but also based on the actual bandwidth resources available. For example, if there are LLQ classes with bandwidth statements, these amounts are deducted from the bandwidth pool that can be assigned to RSVP reservations. While LLQ classes statically allocate bandwidth at configuration time, RSVP does not allocate any amount until a reservation request is received. Therefore, it is important to ensure that an appropriate percentage of the available interface bandwidth is *not* allocated to LLQ classes, so that it can be used by RSVP as reservation requests are received.

Because the total maximum bandwidth that can be assigned to QoS mechanisms on a link is equal to 75% of the link speed, if you want to reserve 33% of the link bandwidth for RSVP-admitted flows, you have to make sure that the bandwidth assigned to LLQ classes does not exceed (75 - 33) = 42% of the link bandwidth.

Because RSVP is in control of assigning packets to the various queues within this model, it is possible to define a mechanism for RSVP to know whether or not to place flows in the Priority Queue (PQ) based on the data flow's T-Spec values by using the following Cisco IOS command in interface configuration mode:

```
ip rsvp pq-profile [r [b [p-to-r]]]
```

Cisco IOS RSVP uses the RSVP TSpec parameters *r*, *b*, and *p-to-r* to determine if the flow being signaled for is a voice flow that needs PQ treatment. These parameters represent the following values:

- *r* = the average traffic rate in bytes per second
- *b* = the maximum burst of a flow in bytes
- *p-to-r* = the ratio of peak rate to average rate, expressed as a percentage

If the traffic characteristics specified by the RSVP TSpec messages for a certain flow are less than or equal to the parameters in the Cisco IOS command, then RSVP will direct the flow into the PQ. If no parameters are provided with the command, the following values, representing the largest of the commonly used voice codecs (G.711), are used as default:

- *r* = 12288 bytes per second
- *b* = 592 bytes
- *p-to-r* = 110%

The IntServ/DiffServ Model

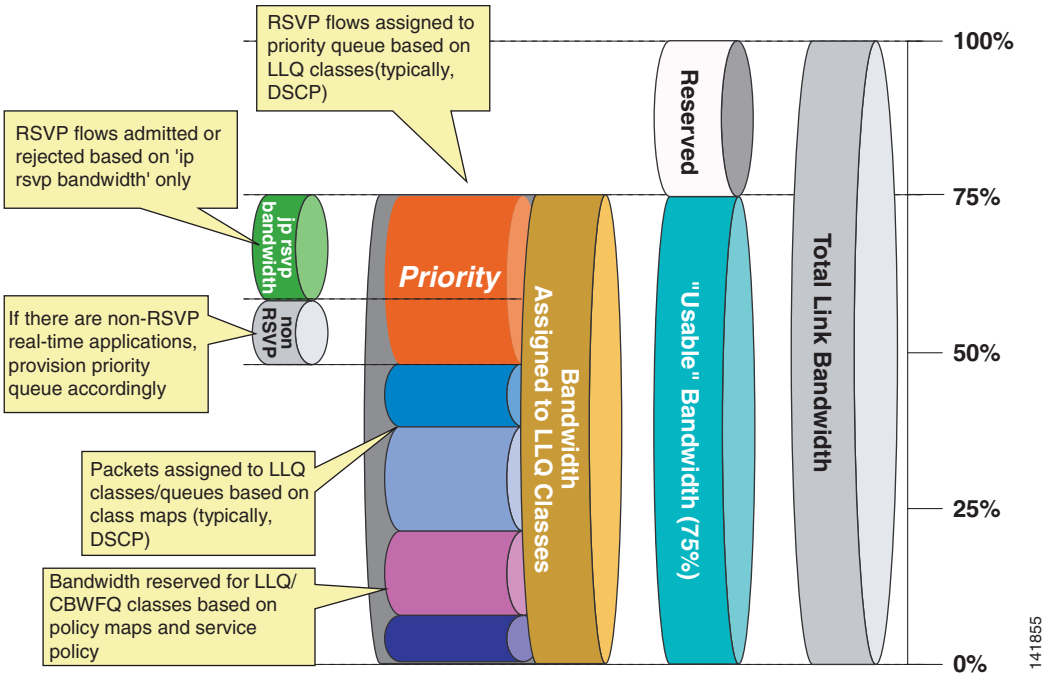
As shown on the right side of Figure 11-34, RSVP in the IntServ/DiffServ model involves only the control plane performing admission control but does not involve the data plane. This means that the call admission control function is separate from the scheduling and policing functions, which can be performed by the Low Latency Queuing (LLQ) algorithm according to predefined class maps, policy maps, and service policies.

With the IntServ/DiffServ model, it is therefore possible to add RSVP call admission control to a network that is already using a Differentiated Services approach to QoS. RSVP admits or rejects calls based on a preconfigured bandwidth amount, but the actual scheduling is based on the pre-existing LLQ criteria such as the DSCP value of each packet.

The entire usable bandwidth (75% of the link speed) can be assigned to LLQ classes, as shown in Figure 11-36, as it normally is today. The policy maps define the traffic that is admitted into each queue. RSVP is typically configured to admit flows up to the amount of bandwidth defined for priority traffic, but keep in mind that RSVP in this model does not adjust the scheduling, so any traffic admitted by RSVP in excess of the predefined priority queue may be dropped or remapped to other lower-priority queues.

If all applications that send priority traffic are RSVP-enabled, you may configure the RSVP bandwidth to match the size of the priority queue. If, on the other hand, there are non-RSVP applications that also need to send priority traffic (such as Unified CM locations-based CAC or a gatekeeper), as shown in Figure 11-36, the priority queue is divided into priority traffic that is controlled by non-RSVP mechanisms and priority traffic that is controlled by RSVP. The combined non-RSVP and RSVP admission control mechanisms must not use more bandwidth than is allocated to ensure that the priority queue is never over-subscribed.

Figure 11-36 LLQ Bandwidth Allocation with RSVP



To use the IntServ/DiffServ operation model on a Cisco IOS router, use the following commands in interface configuration mode:

```
ip rsvp resource-provider none
ip rsvp data-packet classification none
```

When these commands are active, RSVP admits or rejects new reservations uniquely based on the upper bandwidth limits defined within the **ip rsvp bandwidth** command, independently from the actual bandwidth resources available on the interface. Once admitted, the RSVP flows are subject to the same scheduling rules as all other non-RSVP traffic (for example, LLQ class and policy maps). Therefore, it is important to ensure that the RSVP-enabled traffic is marked with the appropriate DSCP value and that the bandwidth of the corresponding PQ or CBWFQ queues is provisioned to accommodate both RSVP-enabled traffic and all other traffic.

In this operating model, the **ip rsvp pq-profile** command is inactive because RSVP does not control the scheduling function.

RSVP Application ID

An application identity (app-id) is an RSVP object that can be inserted into the policy element of an RSVP message. This object is described in RFC 2872. This policy object helps to identify the application and associate it with the RSVP reservation request, thus allowing routers along the path to make appropriate decisions based on the application information.

The need for an app-id arises because RSVP is used to support multiple applications such as voice and video.

Without using an app-id, there is only one bandwidth value that is configurable per interface in RSVP. RSVP will admit requests until this bandwidth limit is reached. It does not differentiate between the requests and is not aware of the type of application for which the bandwidth is requested. As a result of

this, it is quite possible for RSVP to exhaust the allowed bandwidth by admitting requests representing just one type of application, thus causing all subsequent requests to be rejected due to unavailable bandwidth. In this way, a few video calls could prevent all or most of the voice calls from being admitted. For example, if an organization allocates 1000 units to RSVP, RSVP might exhaust a majority of this amount by admitting two 384-kbps video calls, thus leaving very little bandwidth for voice calls.

The solution to this problem is to configure separate bandwidth limits for individual applications or classes of traffic. Limiting bandwidth per application requires that an RSVP local policy matching the application bandwidth limit be applied to the router interface and that each reservation request flag the application to which it belongs so that it may be admitted against the appropriate bandwidth limit.

The app-id is not a single piece of information but multiple variable-length strings. As is described in RFC 2872, the object may include the following attributes:

- An identifier of the application (APP). This attribute is required.
- Global unique identifier (GUID). Optional.
- Version number of the application (VER). This attribute is required.
- Sub-application identifier (SAPP). An arbitrary number of sub-application elements can be included. Optional.

For example:

- APP = AudioStream
- GUID = CiscoSystems
- VER = 5.0.1.0
- SAPP = (not specified)

For more information on how Unified CM uses the Application ID, see RSVP Application ID and Unified CM, page 11-72.

Cisco IOS Features

This section describes new Cisco IOS features that apply to the design of deployments using Cisco Unified CM 8.5 and later releases. These features are new in Cisco IOS Release 15.1(3)T, and it is important to use the correct Cisco IOS version to obtain these features.

RSVP Ingress Call Admission Control

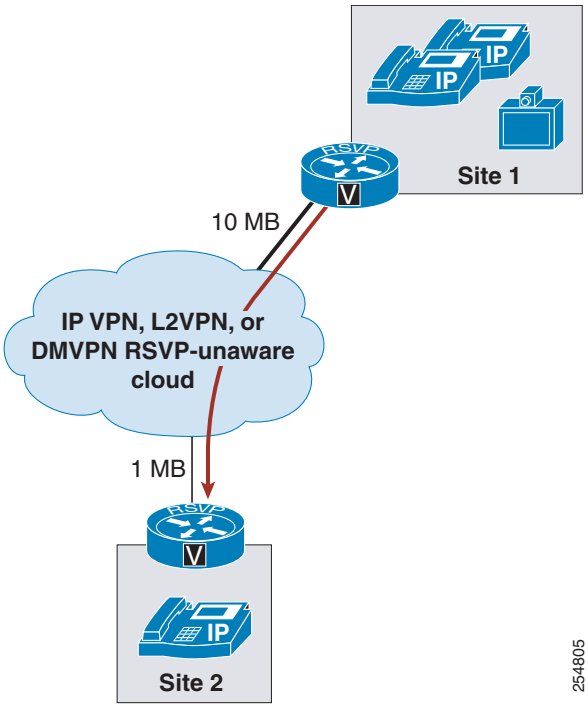
As indicated in the section on RSVP Principles, page 11-42, the RSVP protocol reserves the resources required by the source device to communicate with the destination device. The reservation is unidirectional. The source device signals the path message, which advertises the resource being requested. Upon receiving the path message, the destination device responds with the reservation message. The RESV message traverses the intermediate nodes between the source device and the destination device, hop-by-hop (only RSVP-aware hops), and determines if these nodes can allocate resources for the flow being requested. The reservation is checked against the resources on the outgoing interface only (egress with regard to the direction of the stream) while going downstream in the direction of the destination device.

In the following scenarios, the RESV message is not an indicator for guaranteed communication between the source device and destination device against the resource reserved:

Asymmetric Link Between Two RSVP-Aware Routers

In Figure 11-37, the path message enters the RSVP-unaware cloud through a 10 MB link and goes out of the RSVP-unaware cloud into a 1 MB link. For the stream flowing from Site 1 to Site 2, only the egress interface of the RSVP-aware router at Site 1 is taken into consideration. Downstream, the 1 MB link at Site 2 is not accounted for while making the reservation. In most scenarios this is not an issue because every call has two streams, and a stream in the opposite direction (from Site 2 to Site 1) will reserve the bandwidth on the Site 2 RSVP-aware router.

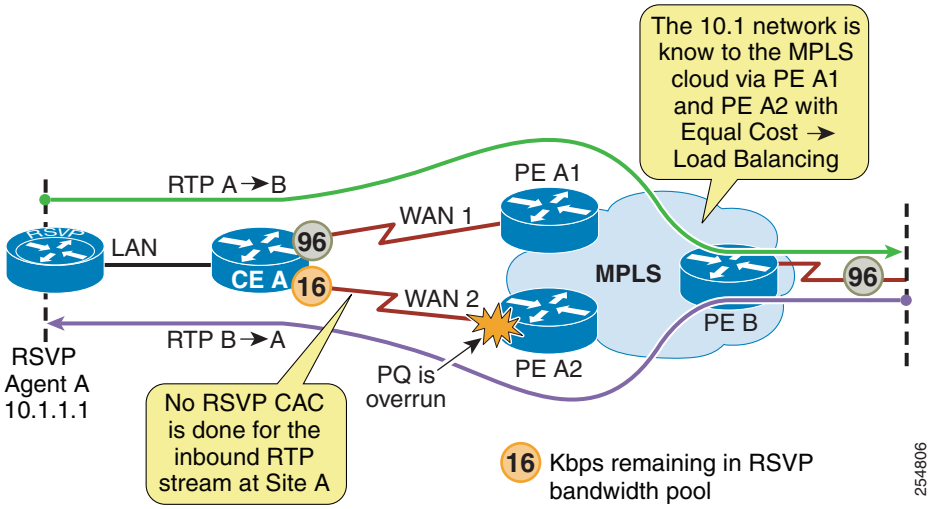
Figure 11-37 Asymmetric Link Between RSVP-Aware Routers



Asymmetric Routing Path Such as a Dual-Attached Customer Equipment (CE) with Load Balancing

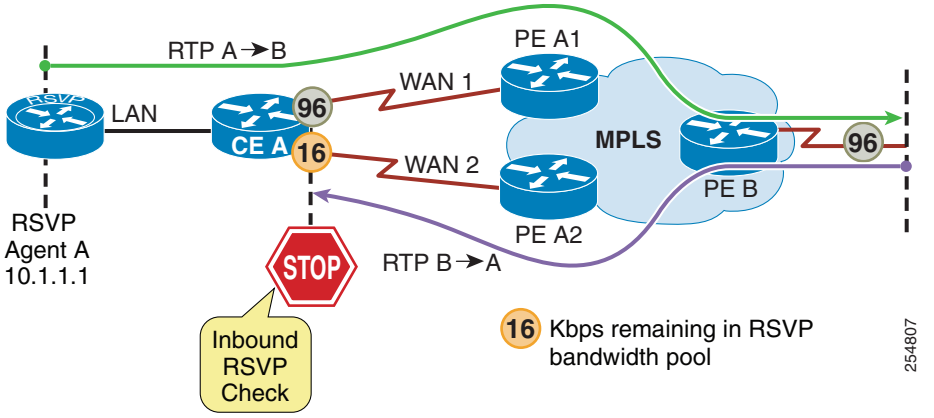
In Figure 11-38, an audio call (two streams, one in each direction) is made between RSVP Agent A and RSVP Agent B. One stream in direction A to B flows over WAN 1, and the other stream in direction B to A flows over WAN 2. This is referred to as an asymmetrically routed call and is common in dual-attached networks with load balancing from a service provider. When RSVP is accounting on only the egress interface into a WAN network that is RSVP unaware, as is the case with service provider clouds for example, there is the potential to overrun the WANs provisioned bandwidth when traffic is load balanced.

Figure 11-38 Asymmetric Routing for a Dual-Attached Customer Equipment (CE) with Load Balancing



To overcome the limitations posed by the above scenarios and to conform with RFC 2205, use the Ingress Call Admission Control feature, which is supported in Cisco IOS Release 15.1(3)T. Ingress Call Admission Control allows the reservation of an RSVP request to be validated against a bandwidth pool on ingress into the router instead of upon egress only. (See Figure 11-39.) Note that egress bandwidth validation will continue to function as usual.

Figure 11-39 RSVP Ingress Call Admission Control



RSVP VPN Tunnel Support

Dynamic Multipoint Virtual Private Network (DMVPN) allows users to scale large and small IPsec VPNs by combining GRE tunnels, IPsec encryption, and Next Hop Resolution Protocol (NHRP). The RSVP VPN Tunnel feature supports the following types of configurations:

- RSVP over manually configured generic routing encapsulation (GRE) and multipoint generic routing encapsulation (mGRE) tunnels
- RSVP over manually configured GRE and mGRE tunnels in an IPsec protected mode
- RSVP over GRE and mGRE tunnels (IPsec protected and IPsec unprotected) in a DMVPN environment

For more information on DMVPN and the RSVP VPN Tunnel feature, refer to the *Cisco IOS Quality of Service Solutions Configuration Guide, Release 15.1*, available at

http://www.cisco.com/en/US/docs/ios/qos/configuration/guide/15_1/qos_15_1_book.html

RSVP for Flexible Bandwidth Interfaces

As indicated in the section on RSVP Principles, page 11-42, when RSVP bandwidth is configured on an interface, the bandwidth value for that interface is fixed. This causes an issue with flexible interfaces (otherwise known as bundled interfaces) such as Multi-Link PPP, ATM IMA, FRF12, Gigabit EtherChannel (GEC), and so forth. The problem is that, when you configure a static RSVP bandwidth amount on a flexible bandwidth interface that contains bundles of links, if one or more of the links fail and the total bandwidth is reduced, the RSVP bandwidth remains fixed. This means that the ratio between the RSVP bandwidth and the total flexible interface bandwidth is no longer equal to the configured value, and this could cause oversubscription of that flexible bandwidth interface.

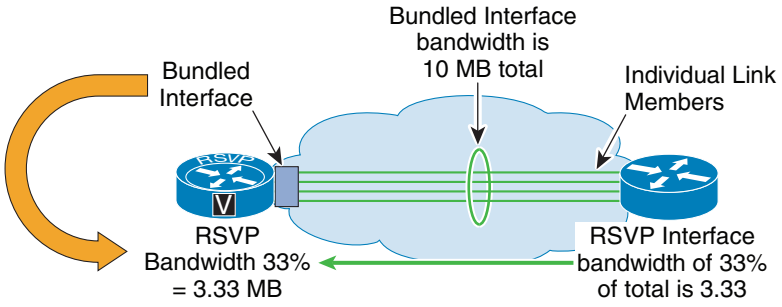
Note that Low Latency Queuing (LLQ) already allows for the Priority Queue and Class-Based Weighted Fair Queues to implement percentages; therefore, when applied to flexible bandwidth interfaces, LLQ parameters change in conjunction with the interface on which they are configured.

The Flexible Bandwidth Interfaces feature enhances the **ip rsvp bandwidth** command to allow for the configuration of a percentage of the interface bandwidth. This allows the RSVP bandwidth to change in parallel with the interface bandwidth, and it is applicable to interfaces that consist of a number of physical links that are bundled into one link.

For enterprise customers with sites that leverage bundled WAN interfaces either within their network or to a service provider, this feature allows them to fully maximize the bandwidth utilization for RSVP call admission control during complete up-time while also allowing them to use the same percentage of bandwidth on the bundle dynamically during link failures.

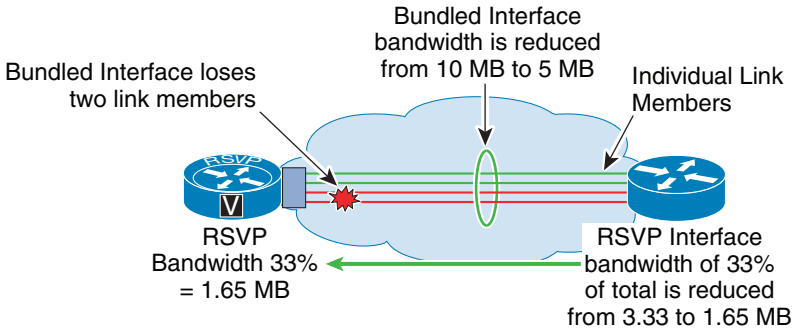
Figure 11-40 and Figure 11-41 illustrate the use of RSVP with flexible bandwidth interfaces.

Figure 11-40 Flexible Bandwidth Interfaces with 10 MB Total Bandwidth



ip rsvp bandwidth percent *rsvp-bandwidth [max-flow-bw* | **percent** *flow-bandwidth*]

Figure 11-41 Flexible Bandwidth Interfaces with Total Bandwidth Reduced to 5 MB



ip rsvp bandwidth percent *rsvp-bandwidth [max-flow-bw* | **percent** *flow-bandwidth*]

For more information on the use of RSVP with flexible bandwidth interfaces, refer to the *Cisco IOS Quality of Service Solutions Configuration Guide, Release 15.1*, available at

http://www.cisco.com/en/US/docs/ios/qos/configuration/guide/15_1/qos_15_1_book.html

RSVP Design Best Practices

When deploying RSVP in the IP WAN in conjunction with Unified CM, observe the following design best practices:

- Cisco recommends that you use the IntServ/DiffServ model if either of the following statements is true:
  - The only traffic destined for the Priority Queue (PQ) in the IP WAN interfaces is RSVP-enabled traffic.
  - All the non-RSVP traffic destined for the PQ can be deterministically limited to a certain amount by an out-of-band call admission control mechanism (such as Unified CM locations or a Cisco IOS gatekeeper).
- If all the PQ traffic is RSVP-enabled, the value specified in the **ip rsvp bandwidth** command and the **priority** command should match once Layer 2 overhead of the priority queue bandwidth has been taken into account.

- If RSVP is enabled on one or more interfaces of a router, all interfaces through which you expect RSVP signaling to transit should also be enabled for RSVP to ensure that RSVP messages do not get dropped. If call admission control is not desired on an interface, set the bandwidth value to 75% of the interface bandwidth.
- If some PQ traffic is not RSVP-enabled, you must ensure that the sum of the values specified in the **ip rsvp bandwidth** command and in the out-of-band call admission control mechanism do not exceed the bandwidth value specified in the **priority** command.
- Enable RSVP Application ID support if you need to limit the maximum amount of bandwidth used by voice and video calls. Application ID Support is introduced in Cisco IOS Release 12.4(6)T.
- Enable RSVP at the edge of the network, including the router WAN interfaces on both sides of the WAN link.
- Enable RSVP at all possible WAN congestion points, including redundant links of different speeds.
- If you do not have symmetric routing on load-balanced MPLS WAN links, ensure that ingress call admission control is configured (see RSVP Ingress Call Admission Control, page 11-53).
- RSVP is currently not available on most Catalyst Switching Platforms.

Bandwidth Provisioning for RSVP

This section discusses bandwidth provisioning as it relates to RSVP only. For a more general and complete discussion on bandwidth provisioning, see Bandwidth Provisioning, page 3-45.

Calculating RSVP Bandwidth Values for Use with Unified CM

At the time Unified CM instructs the Cisco RSVP Agent to make the initial reservation for the call flow, the endpoints that are involved in the call have not fully exchanged their codec capabilities. Without this information, Unified CM must rely on the region settings to determine how to describe the traffic flow. The size of the traffic flow is a function of two things, the codec bit-rate and the sampling rate (or packets per second). The region settings contain the maximum codec bit rate but do not describe the sampling rate. The preferred sampling rates for audio codecs are defined in the following cluster-wide service parameters:

- Preferred G722 millisecond packet size: 20 ms by default
- Preferred G711 millisecond packet size: 20 ms by default
- Preferred G729 millisecond packet size: 20 ms by default

However, the codec type and codec sampling rate are negotiated for every call and might not be the preferred settings because they are not supported on one or more of the endpoints. To avoid having to increase the reservation size once the capabilities are fully exchanged, possibly causing a post-ring failure, this initial reservation is for the worst-case scenario (the largest codec bit rate using the smallest packet size) for that codec. Once media capabilities have been exchanged between the endpoints, then the reservation is revised to the correct bandwidth allocation. In most cases, the default sampling rate is used, resulting in the reservation being reduced.

Note Unified CM does not include the SRTP overhead or the Layer 2 overhead in the RSVP Reservation. When compared to the RSVP T Spec bandwidth value, the Layer 3 IP RSVP bandwidth statement must take into account any SRTP traffic, and the Layer 2 priority queue value must also be over-provisioned if SRTP traffic is present. (See Table 3-10 and Table 3-11.)

Voice Bearer Traffic

Inter-region call with audio codec maximum set to G729, connecting using G.729:

- Initial request: 40 kbps using a 10 ms worst-case scenario
- Updated request: 24 kbps using the preferred sample size of 20 ms

Inter-region call with audio codec maximum set to G.728/iLBC, connecting using iLBC:

- Initial request: 48 kbps using a G.728 10 ms worst-case scenario
- Updated request: 31.2 kbps using iLBC with a preferred sample size of 20 ms

Inter-region call with audio codec set to G711, connecting using G.711:

- Initial request: 96 kbps using a 10 ms worst-case scenario
- Updated request: 80 kbps using the preferred sample size of 20 ms

Video Bearer Traffic

As with the audio stream, the initial reservation for the video stream will rely on the region settings because the endpoint codec capabilities will not be fully negotiated at the time of the reservation. The region settings for video calls include the bandwidth for the audio stream. (See IP Video Telephony, page 12-1, for more information.) Because the audio stream has its own reservation, the final reservation for the video stream will be the region setting minus the audio codec bit-rate. However, because these codecs have not been fully negotiated, the video stream reservation will be for the worst-case scenario, which assumes no audio stream. Once media capabilities have been exchanged between the endpoints, then the reservation will be revised to the correct bandwidth allocation.

Because video is inherently bursty, it is necessary to add some overhead to the stream requirements. (See Voice Bearer Traffic, page 3-46, for more information.) Unified CM uses the stream bandwidth to determine how to calculate the overhead, as follows:

- If the stream is < 256 kbps, then the overhead will be 20%
- If the stream is >= 256 kbps, then the overhead will be 7%

Inter-region video call, with G.729 audio codec and video setting of 384 kbps:

- Initial request: 384 * 1.07 = 410 kbps
- Updated request: (384 - 8) * 1.07 = 402 kbps

Inter-region video call, with G.711 audio codec and video setting of 384 kbps:

- Initial request: 384 * 1.07 = 410 kbps
- Updated request: (384 - 64) * 1.07 = 342 kbps

Configuration Recommendation

Because the initial reservation will be larger than the actual packet flow, over-provisioning the RSVP and LLQ bandwidth is required to ensure that the desired number of calls can complete.

When provisioning the RSVP bandwidth value for N calls, Cisco recommends that the Nth value be the worst-case bandwidth to ensure that the Nth call gets admitted.

For example:

- To provision four G.729 streams:

  (3 * 24) + 40 = 112 kbps

- To provision four G.711 streams:

  (3 * 80) + 96 = 336 kbps
- To provision four 384 kbps video streams (G.729 audio)

  (3 * (384 - 8) + 384) * 1.07 = 1618 kbps
- To provision four 384 kbps video streams (G.711 audio)

  (3 * (384 - 64) + 384) * 1.07 = 1438 kbps

Configuring Cisco IOS Application ID Support

RSVP Application ID feature support was introduced in Cisco IOS Release 12.4(6)T, and that is the minimum release required for the following examples.

Combined Priority Queue

To utilize the functionality allowed in Unified CM's implementation of Application ID support (that is, allowing voice calls to consume all the bandwidth available in the priority queue), we must modify the previous recommendations that voice and video priority queues be kept separate. (See RSVP Application ID and Unified CM, page 11-72.) To use this functionality, you should combine both the voice and video match criteria into one class-map. Because the requirements are to match either voice traffic or video traffic, be sure to make the class-map match criteria **match-any** instead of **match-all**, as follows:

```
class-map match-any IPC-RTP
 match ip dscp ef
 match ip dscp af41  af42
```

Configure the priority queue to support both the voice and video traffic. The following example configuration allocates 33% of the link bandwidth to the priority queue:

```
policy-map Voice-Policy
 class IPC-RTP
  priority percent 33
```

Mapping Application ID to RSVP Policy Identities

The RSVP Local Policy provides the mechanism for controlling a reservation based on an Application ID. Application IDs are mapped to RSVP Local Policies through the **ip rsvp policy identity** command. RSVP Local Policy identities are defined globally and are available to each interface for policy enforcement. Each identity can have one policy locator defined to match an Application ID.

To give the user as much flexibility as possible in matching application policy locators to local policies, the RSVP local policy command line interface (CLI) accepts application ID match criteria in the form of Unix-style regular expressions for the policy locator. Regular expressions are already used in the CLI for existing Cisco IOS components such as Border Gateway Protocol (BGP). Refer to the follow documentation for more information on how regular expressions are used in Cisco IOS:

- *Access and Communication Servers Command Reference*

  http://www.cisco.com/en/US/docs/ios/11_0/access/command/reference/arbook.html
- *Using Regular Expressions in BGP*

  http://www.cisco.com/en/US/tech/tk365/technologies_tech_note09186a0080094a92.shtml
- *Regex Engine Performance Enhancement*

  http://www.cisco.com/en/US/docs/ios/12_3t/12_3t4/feature/guide/gt_rexpe.html

RSVP Policy Identities for Matching the Default Unified CM Application IDs

```
ip rsvp policy identity rsvp-video policy-locator .*VideoStream.*
ip rsvp policy identity rsvp-voice policy-locator .*AudioStream.*
```

Interface RSVP Local Policies

Whether configuring Application ID support or not, for an interface to support RSVP, you must configure the **ip rsvp bandwidth** <*value*> command on that interface. This value cannot be exceeded by any one RSVP reservation or the sum of RSVP reservations on that interface, regardless of Application ID support. In fact, if a reservation passes the local policy check, it still must pass the interface RSVP bandwidth check before it is reserved.

Local policies based on Application ID are applied to an interface using the **ip rsvp policy local identity** command.

For reservations that match its policy locator value, a local policy has the ability to perform the following functions:

- Define the maximum amount of bandwidth the reservations can reserve as a group or as a single sender
- Forward or not forward RSVP messages
- Accept or not accept RSVP messages
- Define the maximum bandwidth the group or sender can reserve

For example, to limit the amount of video bandwidth to 384 kbps on a Serial T1, use the following commands:

```
interface Serial0/0/1:0
 ip rsvp bandwidth 506
 ip rsvp policy local identity rsvp-video
  maximum bandwidth group 384
  forward all
```

There is also a catch-all local policy called the default local policy. This local policy will match any RSVP reservation that did not match the other RSVP local policies configured on the link. The default local policy can be used to match reservations that are not tagged with an Application ID or reservations that are tagged with an Application ID that you want to treat as untagged traffic.

Example

The following example supports both voice and video calls using the model discussed in How Unified CM Uses the Application ID, page 11-72. The voice calls are guaranteed 352 kbps of bandwidth while video calls are limited to 154 kbps of bandwidth. Voice calls can use all of the available RSVP bandwidth.

```
interface Serial0/0/1:0
 ip address 10.2.101.5 255.255.255.252
 service-policy output Voice-Policy
 ip rsvp bandwidth 506
 ip rsvp data-packet classification none
 ip rsvp resource-provider none
 ip rsvp policy local identity rsvp-voice
  maximum bandwidth group 506
  forward all
 ip rsvp policy local identity rsvp-video
  maximum bandwidth group 154
  forward all
 ip rsvp policy local default
  no accept all ! Will not show in the configuration
```

```
no forward all! Will not show in the configuration
```

In this example, if an RSVP reservation is received that does not have an Application ID or its Application ID does not match the two configured options, the reservation will fail. This configuration works if RSVP traffic originates only from Cisco RSVP Agents controlled by Unified CM. However, if there is intercluster RSVP traffic via an IP-IP gateway or if RSVP messages from a controller other than Unified CM are traversing this link, then the default local policy should be configured to accept and forward the reservations and a maximum bandwidth value should be configured on the policy. Note that it is possible to oversubscribe the RSVP bandwidth via the use of multiple RSVP local policies (if the sum of the policies is greater than the RSVP interface bandwidth), but reservations then become first-come, first-serve.

Provisioning for Call Control Traffic with RSVP and Centralized Call Processing

This section discusses bandwidth provisioning for call control traffic when RSVP is used as the call admission control mechanism in a centralized call processing deployment. For a more general discussion of bandwidth provisioning when RSVP is not used, see .Provisioning for Call Control Traffic with Centralized Call Processing, page 3-49

Considerations for Calls Using RSVP

In systems where call admission control uses RSVP, there is additional SCCP call control traffic between Unified CM and the Cisco RSVP Agents located at the branch when IP calls are placed across the WAN. To compute the associated bandwidth, use the following equation:

Bandwidth (bps) = (21 ∗ CHW) ∗ (Number of IP phones and gateways in the branch)

Where CHW represents the number of calls placed across the IP WAN per hour per phone, including calls between IP phones at different branches as well as calls made through gateways located in a different site. For example, in a site where 20 phones each make 10 calls per hour, if 20% of the calls are placed across the IP WAN, then CHW = 2. The equation thus yields: (21∗2)∗20 = 840 bps.

The bandwidth calculated by this equation should be added to the required bandwidth for phone call control.

Unified CM RSVP-Enabled Locations

Cisco Unified CM provides a topology-aware call admission control mechanism based on the Resource Reservation Protocol (RSVP), which is applicable to any network topology and which eases the restriction of a traditional hub-and-spoke topology. The Cisco RSVP Agent is a Cisco IOS feature that enables Unified CM to perform the RSVP-based call admission control. For information on which Cisco IOS platforms support the Cisco RSVP Agent, refer to the *Cisco RSVP Agent Data Sheet*, available at

http://www.cisco.com/en/US/partner/products/ps6832/products_data_sheets_list.html

The Cisco RSVP Agent registers with Unified CM as either a media termination point (MTP) or a transcoder device with RSVP support. When an endpoint device makes a call in need of a bandwidth reservation, Unified CM invokes a Cisco RSVP Agent to act as a proxy for the endpoint to make the bandwidth reservation.

Figure 11-42 shows the signaling protocols used between Unified CM and various other devices, as well as the associated RTP streams for calls across the WAN in a given location. For any calls across the WAN, Unified CM directs the endpoint devices to send the media streams to their local Cisco RSVP Agent, which originates another call leg synchronized with an RSVP reservation to the Cisco RSVP Agent at the remote location. Figure 11-42 illustrates the following signaling protocols:

- Cisco RSVP Agents register to Unified CM via Skinny Client Control Protocol (SCCP).
- IP phones register with Unified CM via SCCP or Session Initiation Protocol (SIP).
- PSTN gateways register with Unified CM via Media Gateway Control Protocol (MGCP), SIP, or H.323 protocol.

Figure 11-42 Protocol Flows for RSVP-Enabled Locations

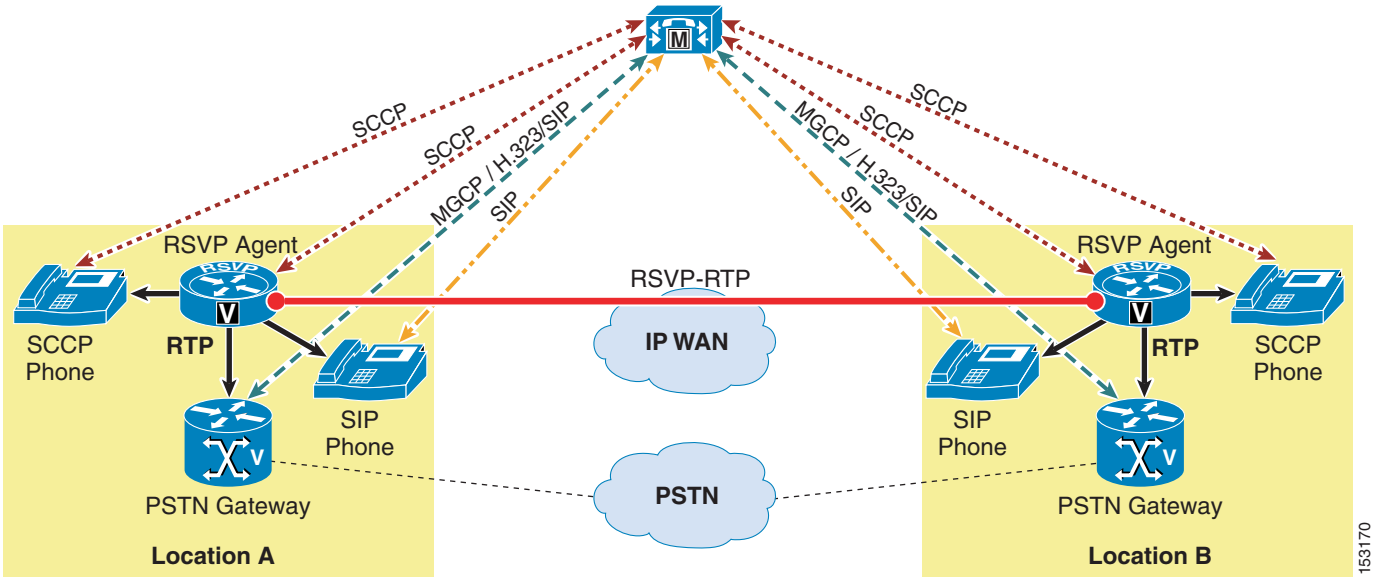
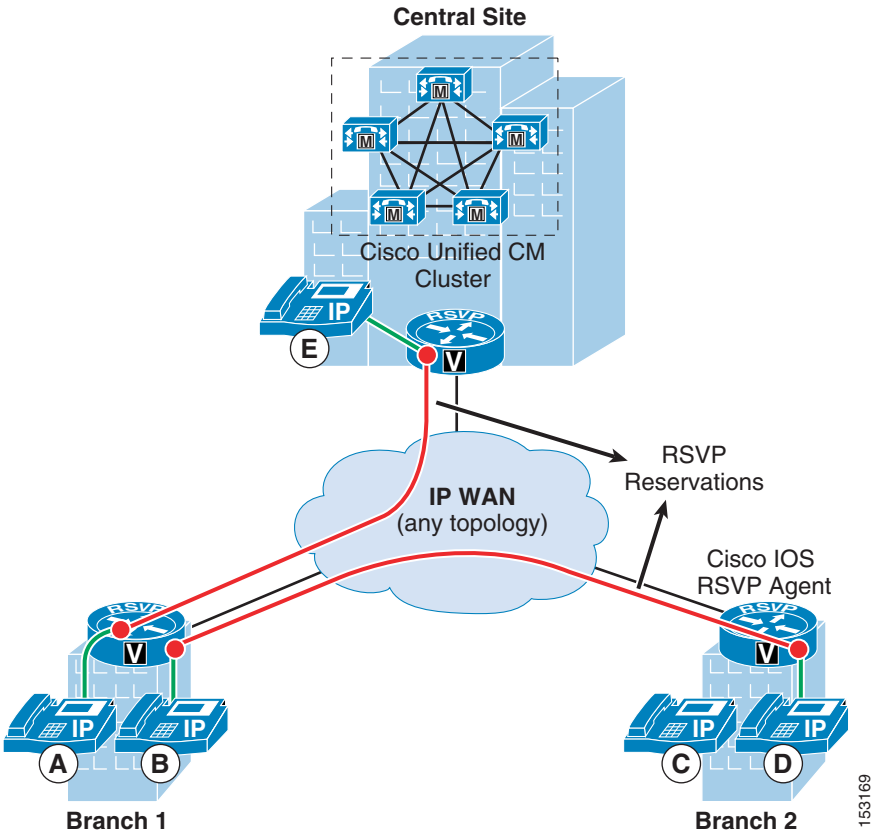


Figure 11-43 shows a typical Cisco RSVP Agent deployment within a Unified CM cluster, which includes three locations: Central Site, Branch 1, and Branch 2. The IP WAN connecting the three locations can be of any topology type and is not restricted to the hub-and-spoke topology. For any call between two locations that requires an RSVP reservation in the media path, a pair of Cisco RSVP Agents is invoked dynamically by Unified CM. The Cisco RSVP Agent acts as a proxy to make an RSVP reservation for the IP phone in the same location with the Cisco RSVP Agent. For example, when phone A in Branch 1 calls phone E in the Central Site, an RSVP reservation (illustrated as the red line in Figure 11-43) is established between Cisco RSVP Agents in the Branch 1 and Central Site locations.

There are three call legs for the media streams of this call. The first call leg is between phone A and the Branch 1 Cisco RSVP Agent, the second call leg is between the Branch 1 and Central Site Cisco RSVP Agents, and the third call leg is between the Central Site Cisco RSVP Agent and phone E. By the same token, when phone B in Branch 1 calls phone D in Branch 2, the RSVP reservation is established between the Branch 1 and Branch 2 Cisco RSVP Agents. Note that the media streams of a call between two branch locations are not sent through the Central Site in this case, which is different from a call made over the traditional hub-and-spoke topology.

Note While RSVP-enabled locations and the use of Cisco RSVP Agent introduce support for arbitrary WAN topologies, they are based on static assignment of devices to a location, which means that every time a device is moved from one physical site to another, its configuration in Unified CM needs to be updated. Device Mobility can be used to update site-specific device configuration information automatically when the device moves to a new physical site. For more information, see the section on Device Mobility, page 25-14.

Figure 11-43 Cisco RSVP Agent Concept



Cisco RSVP Agent Provisioning

The capacity of Cisco RSVP Agent in terms of simultaneous calls (also referred to as sessions) depends on the following factors:

- For software-based MTP functionality, the session capacity is determined by the router platform and the relative CPU load. (Refer to the *Cisco RSVP Agent Data Sheet*, available at http://www.cisco.com/en/US/products/ps6832/products_data_sheets_list.html.)
- For hardware-based MTP and transcoder functionality, the session capacity is limited by the number of DSPs available. (See Media Resources, page 17-1, for DSP sizing considerations.)

For more information on supported platforms, requirements, and capacities, refer to the *Cisco RSVP Agent Data Sheet*, available at:

http://www.cisco.com/en/US/products/ps6832/products_data_sheets_list.html

For software-based MTP functionality, the *Cisco RSVP Agent Data Sheet* provides guidelines for session capacity based on a router dedicated to the Cisco RSVP Agent and 75% CPU utilization. These numbers apply to specific Cisco IOS releases and should be considered as broad guidelines. Different combinations of specific services, configurations, traffic patterns, network topologies, routing tables, and other factors can significantly affect actual performance for a specific deployment and hence reduce the number of concurrent sessions supported. Cisco recommends careful planning and validation testing prior to deploying a multi-service router in a production environment.

Cisco RSVP Agent Registration

The Cisco RSVP Agent registers with Unified CM as an MTP or transcoder device with RSVP support. The Cisco RSVP Agent does not have transcoding capability when registering as an MTP device. To have transcoding capability, the Cisco RSVP Agent must register with Unified CM as a transcoder device.

Registration Switchover and Switchback

If the primary Unified CM fails, the Cisco RSVP Agent switches over to the secondary Unified CM. When the primary Unified CM recovers from the failure, the Cisco RSVP Agent switches its registration back to the primary Unified CM. Use the following commands to configure the Cisco RSVP Agent registration switchover and switchback:

```
sccp ccm group
 switchover method immediate
 switchback method guard timeout 7200
!
gateway
  timer receive-rtp 180
```

- The **switchover method immediate** command specifies the immediate registration switchover to the secondary Unified CM server after failure of the primary Unified CM server is detected. The available DSP resources become available immediately for new calls after the switchover has completed.
- The **switchback method guard timeout 7200** command specifies the registration switchback mechanism after the primary Unified CM recovers from its failure. With this command configured, the Cisco RSVP Agent starts to switch its registration gracefully back to the primary Unified CM after the last active call disconnects. If the graceful registration switchback has not initiated by the time the guard timer expires, the Cisco RSVP Agent will use the immediate switchback mechanism and register with the primary Unified CM right away. The default value of the guard timer is 7200 seconds, and it can be configured statically in the range of 60 to 172800 seconds.
- The **timer receiver-rtp** command in the gateway configuration mode defines the RTP clean-up timer for RSVP reservations. If a failure occurs, the RSVP reservation for the existing call will stay in place until the RTP clean-up timer expires. The default value of this timer is 1200 seconds. Cisco recommends that you configure this timer with its lowest allowed value, which is 180 seconds.

Maximum Sessions Support

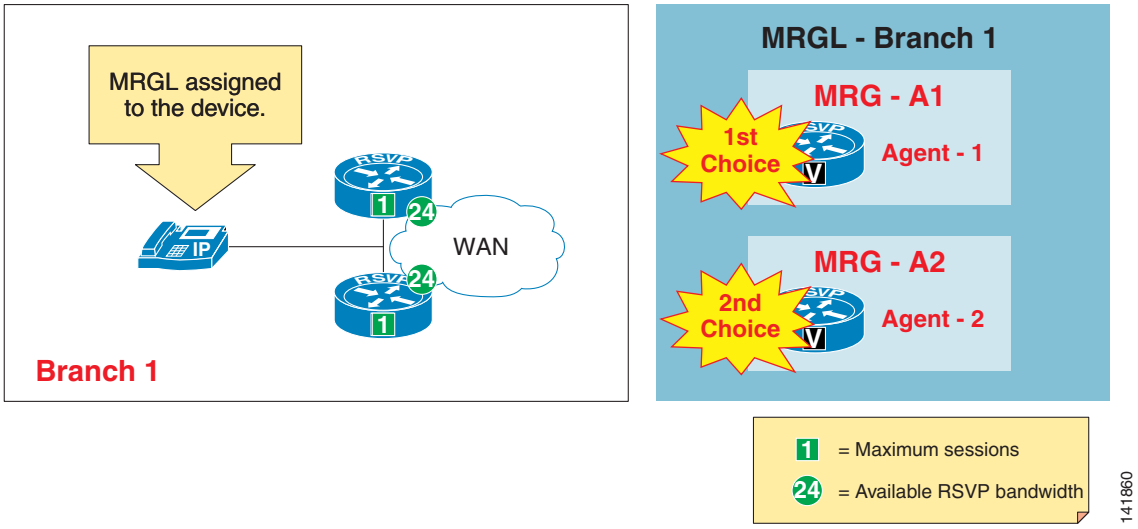
The Cisco RSVP Agent supports a maximum number of calls or sessions, based on the software-based (CPU) and hardware-based (DSP) resources equipped on the Cisco RSVP Agent router. The **maximum sessions** command in the **dspfarm profile** configuration mode specifies the maximum number of calls

that the Cisco RSVP Agent is able to handle. The Cisco RSVP Agent notifies Unified CM of its session capacity based on this configuration. The maximum number of sessions decreases by one for every call going through the Cisco RSVP Agent. When the counter reaches zero, the Cisco RSVP Agent is regarded as having no resources available, and Unified CM skips that Cisco RSVP Agent for any subsequent calls.

Figure 11-44 shows a branch site with dual Cisco RSVP Agents. The Cisco RSVP Agents are co-resident with the WAN routers, and Cisco RSVP Agent redundancy is achieved by assigning two Cisco RSVP Agents to different MRGs in the same MRGL. If Agent-1 in MRG-1 is unavailable or out of session capacity, Unified CM will try to allocate Agent-2 in MRG-2 for RSVP calls to or from Branch 1. To ensure that Agent-2 is selected when Agent-1's capacity is reached, Cisco recommends configuring the maximum number of sessions to match exactly the number of calls supported by the **ip rsvp bandwidth** configured on the WAN interface of the Cisco RSVP Agent. In this example, both Cisco RSVP Agents need to be configured with **maximum sessions 1**. This recommendation is made on the assumption that all calls going across the WAN will use the same type of codec so that an accurate calculation of the number of calls across the WAN can be derived, which is done by dividing the total available RSVP bandwidth by the bandwidth requested per call.

Note If the maximum number of sessions is higher than the number of calls supported by the **ip rsvp bandwidth** configuration, Unified CM will still send the call to the Cisco RSVP Agent but the RSVP reservation will fail because there is no bandwidth available, and Unified CM will follow the usual behavior for call admission control failure (that is, it will deny the call or invoke the AAR feature).

Figure 11-44 Configuring Maximum Sessions on the Cisco RSVP Agent



Pass-Through Codec

The pass-through codec enables a Cisco IOS Enhanced MTP device to terminate an RTP media stream received from an endpoint without knowing the media encoding of the stream. That is, the UDP packets of the media stream flow through the MTP without being decoded. This method enables the MTP to support every audio, video, and data codec that is defined in Unified CM. Because the MTP does not decode the media stream, the pass-through codec can also be used with encrypted (SRTP) media streams.

In fact, for video and SRTP media streams to use an MTP, it must support the pass-through codec. When configured with the pass-through codec, the Cisco RSVP Agent will substitute its own IP address for the source IP address in the IP/UDP headers of the packets and let them flow through.

The Cisco RSVP Agent will use the pass-through codec only if all of the following conditions are met:

- The two endpoint devices involved in the call have matching audio codec capability, and the region configuration permits the matching codec to be used for the call. In other words, no transcoder device needs to be inserted in the call.
- **MTP Required** is not configured for either endpoint device.
- All intermediate resource devices support the pass-through codec.

Note If the Cisco RSVP Agent registers as an MTP device and a transcoder device needs to be inserted in the call, the codec configured in the Cisco RSVP Agent dspfarm MTP profile must match the inter-region bit rate configured in Unified CM Administration. For example, if 8 kbps (G.729) is the inter-region bit rate configured in Unified CM Administration, then the G.729 codec must also be configured in the dspfarm MTP profile.

The following example shows an Cisco RSVP Agent configuration on a Cisco IOS 2900 Series platform:

```
interface Loopback0
 ip address 10.11.1.100 255.255.255.255
!
sccp local Loopback0
sccp ccm 20.11.1.50 identifier 1 priority 1 version 7.0+
sccp ccm 20.11.1.51 identifier 2 priority 2 version 7.0+
sccp
!
sccp ccm group 1
 associate ccm 1 priority 1
 associate ccm 2 priority 2
 associate profile 1 register RSVPAgent
 switchover method immediate
 switchback method guard timeout 7200
!
dspfarm profile 1 mtp
 codec pass-through
 codec g729ar8
 rsvp
 maximum sessions software 100
 associate application SCCP
```

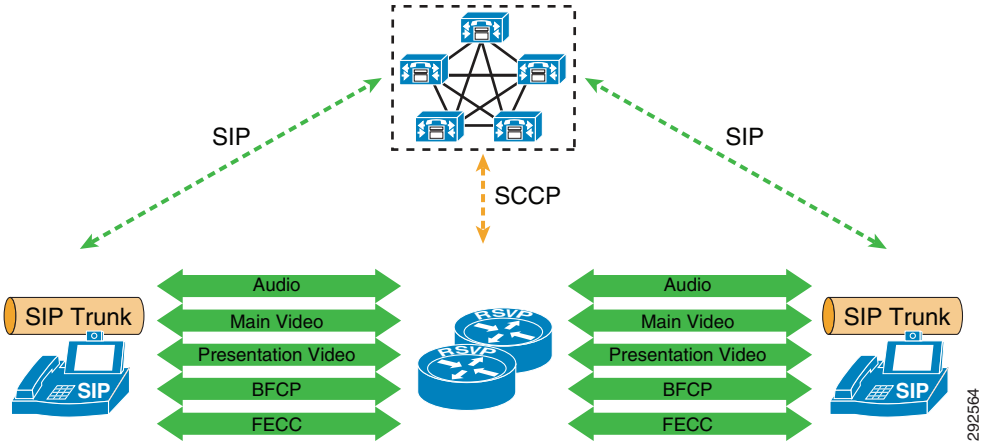
RSVP Agent Support for RTCP, BFCP and FECC Negotiation

As mentioned, RSVP Agent supports a pass-through codec that does not decode the media but, as the name implies, passes the media through yet terminating and re-originating Layer 3 headers. This allows RSVP Agent to support any codec defined or used in Unified CM. With Cisco Unified CM 9.x and Cisco IOS Release 15.2.1T and later releases, Unified CM supports RTCP, BFCP, and FECC negotiation and pass-through as described in the following sections.

BFCP and FECC Pass-Through

RSVP Agent (Cisco IOS MTP and transcoder) and Unified CM support Binary Flow Control Protocol (BFCP) and Far End Camera Control (FECC) pass-through. Previously this was not possible due to lack of media port support by the RSVP Agent that was limited to three media ports. With more media port support, BFCP and FECC negotiation now works end-to-end through the RSVP Agent. Figure 11-45 illustrates BFCP and FECC support in the RSVP Agent.

Figure 11-45 RSVP Agent Support for BFCP and FECC Pass-Through



When Unified CM negotiates a video call with presentation sharing using BFCP and FECC, the RSVP Agent passes through the BFCP control channel, the FECC channel if negotiated, and the a second video channel associated to the presentation sharing controlled through BFCP. However, RSVP Agent reserves the bandwidth only for the bit rate of the main video channel. Endpoints using BFCP down-speed the main video to allow the presentation video, so that the main video and presentation video do not use more bandwidth than requested for the main video. If there is no main video channel, then Unified CM reserves bandwidth only for the presentation video negotiated through BFCP. The latter scenario is quite rare; typically there will be a main video channel negotiated, and thus the bandwidth reservation made by RSVP Agent is associated to that.

When BFCP and FECC are negotiated with RSVP Agent, Unified CM requests ports from the RSVP Agent as they are available. This port request has a priority order to it that is hard-coded in the Unified CM request logic. The order is as follows:

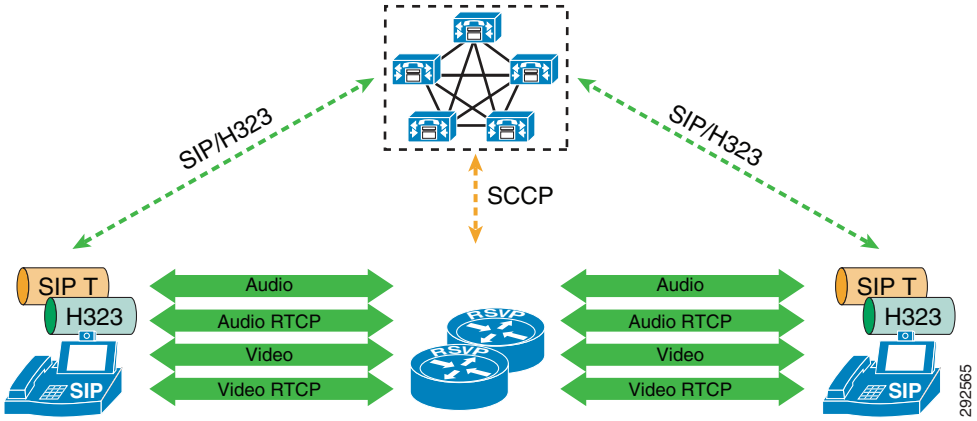
1. Audio
2. Main video
3. BFCP control channel
4. Second video or presentation video channel
5. FECC channel

An example of where this priority comes into play is in scenarios where the RSVP Agent can provide fewer ports than requested, in which case certain functions will be excluded from the negotiation. An example of this is if the RSVP Agent can provide only four ports for a video call request with presentation sharing and FECC. In this case FECC is last in the priority, so it will not get a channel because five channels are requested but only four are available.

RTCP Pass-Through

In the same way that RSVP Agent (Cisco IOS MTP and transcoder) supports BFCP and FECC, it also supports RTCP pass-through. RTCP is a highly utilized protocol negotiated between endpoints, and it can be critical for higher definition video calls to ensure audio and video synchronization. Figure 11-46 illustrates a video call with RTCP pass-through.

Figure 11-46 Video Call Using RTCP Pass-Through



When Unified CM negotiates a video call over a trunk or endpoint and both RTCP and RSVP are enabled, the RSVP Agent opens a new RTCP channel for each media stream. Figure 11-46 illustrates a video call where both audio and video have independent RTCP channels.

For more information on function, design, and deployment of Cisco IOS media resources (RSVP Agent, MTP, and transcoder) with BFCP, FECC, and RTCP support, see the chapter on Media Resources, page 17-1.

RSVP Policy

Unified CM can apply different RSVP policies to different location pairs. The RSVP policy can be configured in Unified CM Administration. The RSVP policy defines whether or not Unified CM will admit the call if the RSVP reservation attempt fails. The following RSVP policy settings can be configured between any two locations:

- No Reservation

  No RSVP reservation attempt is made and, if enabled, Enhanced Locations call admission control is performed by Unified CM.

- Mandatory

  Unified CM does not ring the terminating endpoint device until the RSVP reservation succeeds for the audio stream and, if the call is a video call, for the video stream as well.

- Mandatory (Video Desired)

  A video call can proceed as an audio-only call if a reservation for the video stream cannot be reserved but the reservation for the audio stream succeeds.

- Optional (Video Desired)

  A call can proceed as a best-effort audio-only call if it fails to obtain reservations for both its audio and video streams. The Cisco RSVP Agent re-marks the media packets as best-effort.

- Use System Default

  The RSVP policy for the location pair matches the clusterwide RSVP policy. The default clusterwide RSVP policy is No Reservation. To change the default RSVP policy in Unified CM Administration, select **System > Service Parameters > Cisco CallManager Service**.> **Default Inter-location RSVP Policy**

Note With the Optional (video desired) policy, IP WAN calls may proceed as best-effort not only if the RSVP reservation fails but also if the Cisco RSVP Agent is not available. In this case, Unified CM instructs SCCP and MGCP devices to re-mark their traffic as best-effort. However, this re-marking is not possible with H.323 and SIP devices, which will keep sending their traffic with the default QoS marking. To prevent over-subscribing the priority queue in the latter case, Cisco recommends configuring an access control list (ACL) on the IP WAN router to permit only packets marked with DSCP EF or AF41 if the source IP address is that of the Cisco RSVP Agent.

Cisco recommends configuring the RSVP policy as **Mandatory** or **Mandatory (Video Desired)** because those settings guarantee the bandwidth reservation and the voice quality of the call. The most efficient method for setting the clusterwide RSVP policy is to configure the **Default Inter-location RSVP Policy** in the RSVP clusterwide parameters of the Cisco CallManager Service Service Parameter Configuration, and leave the RSVP configuration in the location configuration set to **Use System Default**.

In the clusterwide RSVP parameters configuration, there is a service parameter named **Mandatory RSVP mid call error handle option**. If the RSVP policy is configured as **Mandatory** or **Mandatory (Video Desired)**, this parameter specifies how Unified CM will treat an existing RSVP call based on the failure of a mid-call RSVP reservation attempt. The mid-call RSVP reservation attempt can be triggered by (but is not limited to) a network convergence after a WAN failure or by an existing voice-only call becoming a video call. A network convergence makes the Cisco RSVP Agent not only start to send the media streams over the newly converged path but also to try to make a new RSVP reservation over the new path.

The default setting of the **Mandatory RSVP mid call error handle option** is **Call Becomes Best Effort**. With the default option configured, Unified CM will maintain the existing call even though the mid-call RSVP reservation attempt fails, but the RTP streams will be marked as best effort (DSCP 0). Cisco recommends configuring this parameter with the **Call Fails Following Retry Counter Exceeded** option. With this option configured, Unified CM will fail the call if the RSVP reservation attempt keeps failing after a certain number of retries. The default value of the retry counter is 1, which is defined by the **RSVP Mandatory mid-call retry counter** service parameter, and the default value of **RSVP retry timer** is 60 seconds. Cisco recommends having both the retry counter and the retry timer service parameters configured with their default values. With both set to their default values, Unified CM will wait for 60 seconds before it disconnects the call if the RSVP mid-call retry fails. During this period, users might experience degraded voice quality because no RSVP reservation is in place and the RTP streams are marked as best effort.

Migrating to RSVP Call Admission Control

To migrate to RSVP-based call admission control, Cisco recommends configuring and deploying RSVP in the network, configuring and deploying RSVP Agents in the branch locations and in Unified CM, and when all RSVP configurations are complete, using the Unified CM clusterwide RSVP service parameter **Default inter-location RSVP Policy** to switch all locations directly over to RSVP CAC. This method

allows the administrator to completely deploy RSVP in both the network infrastructure and the Unified Communications infrastructure while continuing to use Enhanced Locations CAC until the switch is ready to be made. It also allows the administrator to easily switch back to Enhanced Locations CAC in the event of a misconfiguration.

Note that Unified CM first checks if RSVP is enabled and then checks locations and links through the LBM. This simultaneous functioning of CAC mechanisms allows for an easier migration and the ability to revert back to Enhanced Location CAC in the event that there is a misconfiguration.

The following is a short list of events that occur for an intra-cluster call when both RSVP and Enhanced Locations CAC are enabled:

1. Unified CM first checks the location pair policy or the clusterwide **Default inter-location RSVP Policy** of the locations of the devices in the call. If RSVP is enabled between the locations, Unified CM allocates RSVP Agents from the MRGL of each device in the call and makes an RSVP reservation request.
2. When RSVP Agent returns the reservation request result, Unified CM checks to see if Enhanced Locations CAC is enabled (LBM is active). If it is, Unified CM requests the bandwidth from LBM over the effective path (end-to-end location and link path).
3. LBM returns the results of the path request and, if the request is successful, allows the call.
4. If LBM is not enabled or available, Unified CM checks the Cisco CallManager service parameter **Call Treatment When No LBM Available**. If this parameter is set to allow the call, then the call will complete; if it is set to reject the call, then the call will fail.

Observe the following recommendations when planning an RSVP migration:

- Deploy RSVP in the WAN network infrastructure. See the section on RSVP and QoS in WAN Routers, page 11-48.
- Set up Cisco RSVP Agent in each branch location and assign each RSVP Agent to the MRG and MRGL associated to the IP phones and devices in each applicable branch. Ensure that the phones and devices in the branch use the RSVP Agent that is located in the same local branch.
- Ensure that each branch location RSVP setting between any pair of locations is configured with **Use System Default**.
- Ensure that the Cisco CallManager service parameter **Call Treatment When No LBM Available** is set to **allow call**.
- Once RSVP configuration and deployment is completed, change the Cisco CallManager clusterwide RSVP service parameter of **Default inter-location RSVP Policy** from **No Reservation** to the desired RSVP policy, such as **Mandatory** or **Mandatory (Video Desired)**.
- After ensuring that RSVP is being engaged for calls between locations, you can disable the Cisco Locations Bandwidth Manager (LBM).
- If inter-cluster RSVP support is required, enable RSVP over SIP Preconditions as outlined in the section on Migration from Enhanced Locations Call Admission Control to RSVP SIP Preconditions, page 11-78.
- If for any reason there is a need to return to Enhanced Locations CAC, enable LBM services on the Unified CM servers running the Cisco CallManager service and change the Cisco CallManager clusterwide RSVP service parameter **Default inter-location RSVP Policy** to **No Reservation**.

RSVP Application ID and Unified CM

The RSVP Application ID is a mechanism that enables Unified CM to add an identifier to both the voice and video traffic so that the Cisco RSVP Agent can set a separate bandwidth limit on either traffic based on the identifier received. To deploy the RSVP Application ID in the network, you must use a minimum version of Cisco IOS Release 12.4(6)T or higher on the Cisco RSVP Agent router. The RSVP Application ID strings can be configured via two service parameters in the clusterwide RSVP parameter configuration: **RSVP Audio Application ID** and **RSVP Video Application ID**.

Unified CM uses SCCP to convey the RSVP Application ID to the Cisco RSVP Agent. The Cisco RSVP Agent also inserts the RSVP Application ID into the RSVP signaling messages (such as the RSVP PATH and RESV messages) and sends those messages to the downstream or upstream RSVP routers.

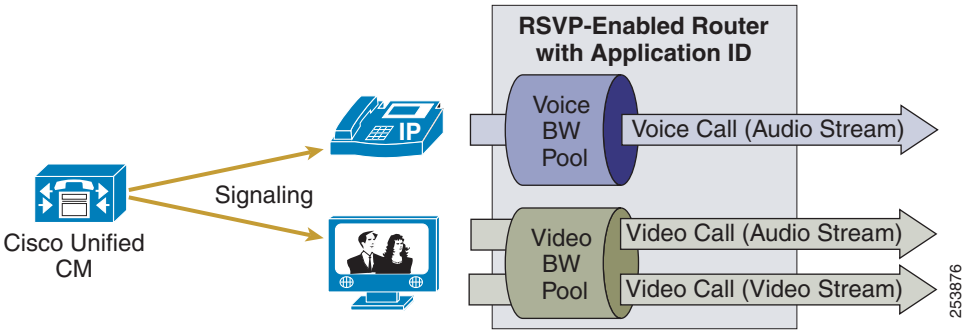
How Unified CM Uses the Application ID

Unified CM has two cluster-wide service parameters that define the Application ID used to tag audio and video call reservations using RSVP:

- RSVP Audio Application ID (Default = AudioStream)
- RSVP Video Application ID (Default = VideoStream)

Figure 11-47 shows how Unified CM tags voice and video calls with an Application ID in RSVP.

Figure 11-47 Unified CM and RSVP Application ID



How Voice Calls Are Tagged

When a voice call is made between locations with an RSVP policy, the resulting reservations for the audio stream will be tagged with the RSVP Audio Application ID.

How Video Calls Are Tagged

When a video call is made between locations with an RSVP policy, the resulting reservations for the audio and video streams will both be tagged with the RSVP Video Application ID. A video call has both audio and video associated to the Video Application ID.

RSVP Application ID Design Considerations and Best Practices

- The AudioStream Application ID is used for audio streams of audio-only calls.
- The VideoStream Application ID is used for both the audio and video streams of a video call.

- The Application ID does not currently differentiate between various types of video, such as telepresence video versus other video. All video in an RSVP session will be marked with the Video Application ID and the Video DSCP value.
- Unified CM currently has separate settings for both the Application ID and the DSCP values of the signaling and media streams. These are managed separately; however, Cisco recommends using the default values because they are configured to work in conjunction with one another.
- When video escalation occurs, the RSVP reservation for the audio stream is readmitted with the Video Application ID and configured DSCP value (PHB of AF41, by default). If the readmission for the audio stream fails due to insufficient bandwidth, the audio stream will continue as best-effort with a Video Application ID until the reservation into the Video Application ID pool succeeds.
- When video de-escalation occurs, the RSVP reservation for the audio stream is readmitted with the Audio Application ID and configured DSCP value (PHB of EF, by default). If the readmission for the audio stream fails due to insufficient bandwidth, the audio stream will continue as best-effort with an Audio Application ID until the reservation into the Audio Application ID pool succeeds.

Video Escalation Example with Application ID

An audio-only call is set up with the AudioStream Application ID, and the DSCP for the stream is set to a PHB value of EF. When the call is escalated video, the video streams are set up with the VideoStream Application ID. If the video stream reservation fails, the call will stay as an audio-only call with the AudioStream Application ID. However, if the video stream reservation succeeds, the audio stream will be readmitted from AudioStream Application ID to VideoStream Application ID. If the readmission succeeds, then both the video and audio streams will have the VideoStream Application ID set to a PHB value of AF41. If the readmission fails, then the video stream will have the VideoStream Application ID with a PHB value of AF41 while audio stream will have the VideoStream Application ID with a PHB set to 0 (video fail DSCP value).

See Unified CM Video Calls with RSVP SIP Preconditions, page 11-80, for information on video escalation and de-escalation in distributed Unified CM environments with RSVP SIP Preconditions.

RSVP SIP Preconditions

RSVP SIP Preconditions is based on SIP Preconditions as defined in RFC 3312 and RFC 4032, and it offers the ability for Cisco call processing products to negotiate a level of Quality of Service and perform call admission control using the RSVP protocol. The term RSVP SIP Preconditions is used to identify the passing of policy information elements, or preconditions, over SIP signaling to negotiate a Quality of Service (QoS) policy. The actual RSVP message is not signaled over the SIP trunk; only the policy-related information elements are. The RSVP messages are then carried over by the RSVP Agent or RSVP-capable router. This use of SIP preconditions extends the negotiation of RSVP Quality of Service policy across Unified CM clusters as well as to Unified CM Express and SIP-TDM Cisco IOS gateways to allow for synchronization of the RSVP layer and call control layer between these various call control applications.

Overview of SIP Preconditions

As mentioned, SIP Preconditions provide for the negotiation of RSVP policy information across call control applications, thus allowing synchronization between these call control applications of the RSVP Layer for resource reservation and the call control layer for call setup and establishment.

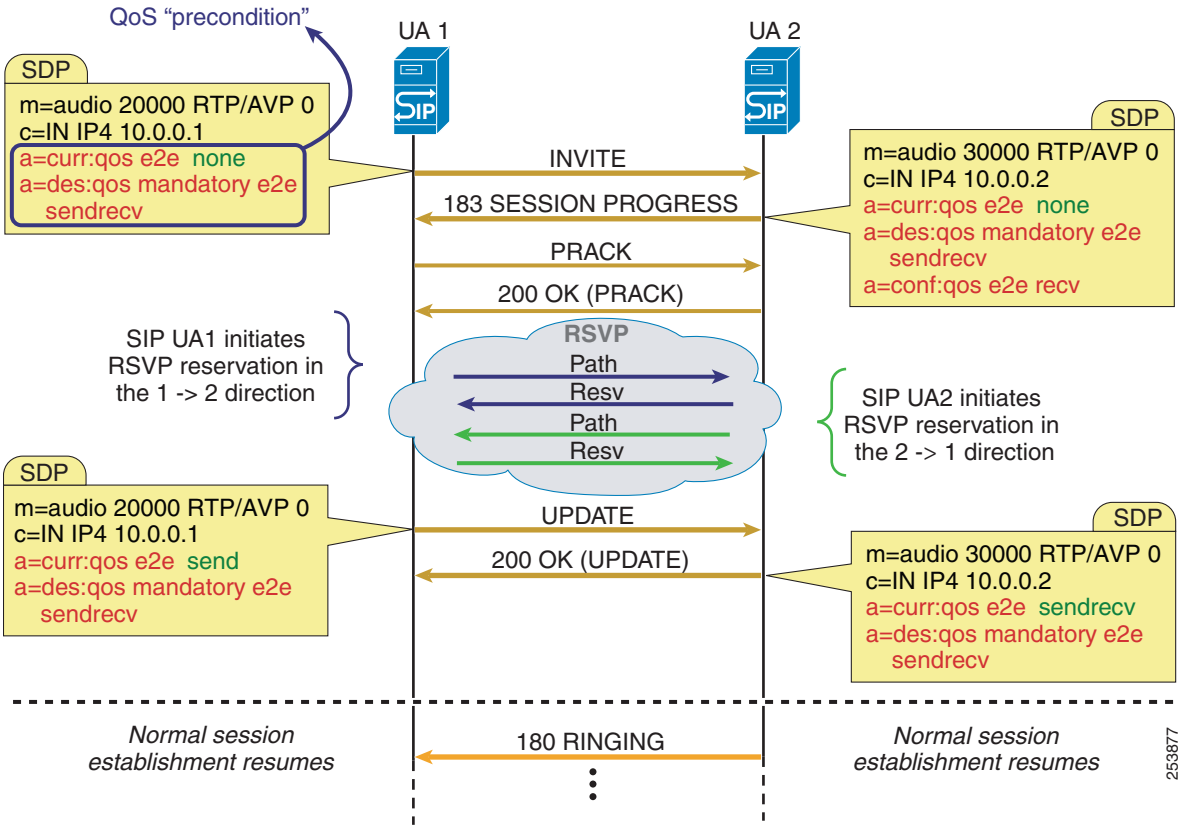
The concept of a precondition in SIP signaling also avoids the potential for what is referred to as "ghost rings" across independent call control applications. Ghost rings can occur at session establishment time if the called party is rung without having first reserved the resources necessary to establish the media

between the callers. In order to minimize ghost rings, network resources for the session must be reserved before the called party is alerted. However, the reservation of network resources frequently requires learning the IP address, port, and session parameters from the called party. This information is obtained as a result of the initial offer and answer exchange carried in SIP. This exchange normally causes the phone to ring, thus introducing a loop dilemma: resources cannot be reserved without performing an initial offer and answer exchange, but the initial offer and answer exchange cannot be done without performing resource reservation.

RSVP SIP Preconditions solves this dilemma by setting SIP Preconditions or constraints about the session that are introduced in the offer. The recipient of the offer generates an answer but does not alert the user or otherwise proceed with session establishment. Proceeding occurs only when the preconditions are met. This knowledge can be obtained through a local event, such as a confirmation of a resource reservation, or through a new offer sent by the calling party.

Figure 11-48 illustrates how these SIP Preconditions work in a generic SIP signaling call flow.

Figure 11-48 SIP with RSVP Between Call Agents



In Figure 11-48, a SIP user agent (SIP UA 1) starts the call by sending a SIP Invite message. The preconditions are in the SIP Invite in the Session Description Protocol (SDP), where the calling party's IP address and port number are identified. The preconditions stipulate a current QoS policy (a=curr:qos) and a desired QoS policy (a=des:qos). In this example, SIP UA 1 sends an invite to SIP UA 2 with a current QoS policy for the audio line set to **none** and the desired QoS policy is set to **mandatory e2e sendrecv**. This tells the receiver that an RSVP reservation is mandatory before offering the call (ringing the end device). The SIP UA 2 receiving the Invite then responds with a 183 session progress message with SDP stipulating a response to the preconditions sent. In this example, SIP UA 2 sends a current QoS

policy as **none**, a desired QoS policy of **mandatory e2e sendrecv**, and a configured QoS policy (a=conf:qos) of **e2e recv**, indicating that it has received the request and will initiate a reservation using RSVP. At this point both user agents negotiate RSVP to reserve bandwidth for the media as described in the SDP. If this reservation succeeds, the UAs update one another with the updated QoS policy preconditions and then proceed with the call by ringing the end user. In the example, SIP UA 2 then responds with a 180 Ringing message and the call can continue with normal establishment. If the reservation fails, then either SIP UA can terminate the call prior to ringing the called party. This avoids any "ghost ringing" condition.

Unified Communications Manager and RSVP SIP Preconditions

RSVP SIP Preconditions for Unified CM provides the functionality of intercluster call admission control in distributed Unified CM deployments. If you deploy RSVP SIP Preconditions in Unified CM, Cisco recommends having local RSVP-enabled locations-based call admission control fully functional prior to enabling RSVP SIP Preconditions. This approach is also recommended for migration purposes. For further details on enabling intra-cluster RSVP call admission control, see Unified CM RSVP-Enabled Locations, page 11-62.

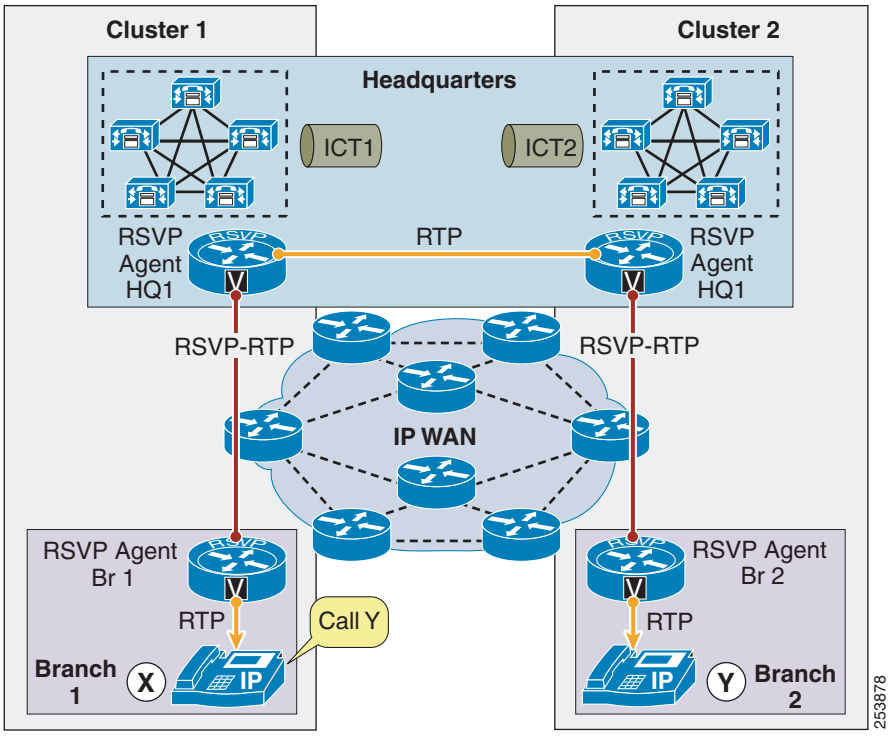
RSVP SIP Preconditions has two modes of configuration, local RSVP and end-to-end RSVP. These modes are configured on the SIP Trunk Profile in Unified CM administration pages.

Local RSVP

Local RSVP does not support reservations between two RSVP agents that are located in separate clusters. It uses two RSVP agents per cluster and does not use RSVP across the trunk that connects the clusters. This is the default configuration of the SIP Trunk Profile.

Figure 11-49 illustrates local RSVP in a distributed Unified CM deployment.

Figure 11-49 Local RSVP in a Distributed Unified CM Deployment



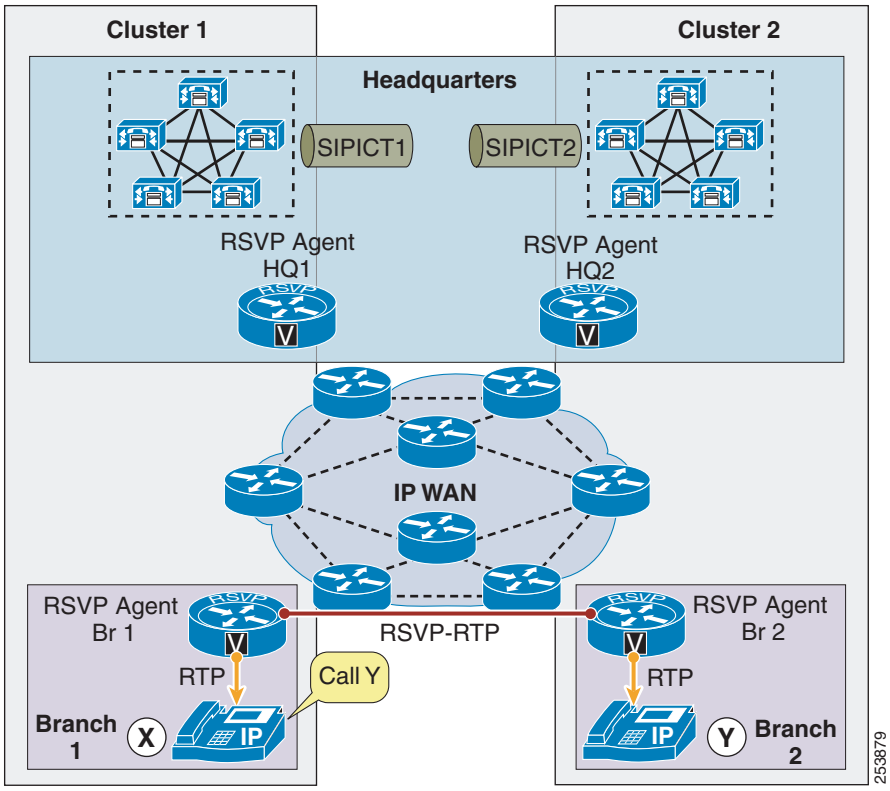
In Figure 11-49, X indicates an endpoint in cluster 1, Y indicates an endpoint in cluster 2, and ICT1 and ICT2 indicate the intercluster trunks configured in clusters 1 and 2 respectively. The RSVP agents associated with the respective devices are also indicated. In this scenario, Cisco Unified CM cluster 1 controls the reservation between AgentBr1 and AgentHQ1, and Cisco Unified CM cluster 2 controls the reservation between AgentBr2 and AgentHQ2.

End-to-End RSVP

End-to-end RSVP configuration is available if the clusters are connected by a SIP trunk. End-to-end RSVP uses RSVP on the entire connection between the RSVP agents, and it uses only one RSVP agent per cluster.

Figure 11-50 illustrates end-to-end RSVP in Unified CM.

Figure 11-50 End-to-End RSVP



In Figure 11-50, X indicates an endpoint in cluster 1, Y indicates an endpoint in cluster 2, and ICT1 and ICT2 indicate the intercluster trunks configured in clusters 1 and 2 respectively. The RSVP agents associated with the respective devices are also indicated. In this scenario, Cisco Unified CM establishes an end-to-end RSVP connection directly between AgentBr1 and AgentBr2.

RSVP SIP Preconditions and Fallback to Local RSVP

Unified CM can be configured to fall back from end-to-end RSVP to local RSVP by configuring **Fall back to local RSVP** on the SIP Trunk profile. This fallback occurs only when the terminating side of the SIP trunk returns a SIP 420 response (Bad Extension), which indicates that the terminating side does not understand the preconditions. Fallback does not occur when a response such as a SIP 580 response (Precondition Failed) is returned. In the case where an end-to-end RSVP SIP Preconditions failure occurs with a SIP 420 (Bad Extension) response during call establishment, Unified CM will invoke local RSVP. If this behavior is desired, a media resource group list with an RSVP Agent association must be assigned to the SIP intercluster trunk. If fallback to local RSVP is not configured, then Unified CM will continue down the route group and route list to another configured trunk or gateway (if configured), otherwise the call will fail.

This feature could be used in designs where a single SIP trunk could terminate to multiple destinations, where both SIP preconditions are supported and where they are not supported. An example might be with the Unified Proxy Server, where there is a single SIP trunk that is configured to a SIP proxy and the returned destination could be a terminating cluster that understands the SIP preconditions or a terminating cluster or SIP server that does not understand the SIP preconditions. Because there is only a single SIP trunk in this case, it would be enabled for RSVP SIP preconditions with fallback enabled. In cases where the terminating side does not understand the SIP preconditions, an RSVP agent can be

associated to the SIP trunk in fallback mode so that, when a SIP 420 message (Bad Extension) is received and fallback occurs, a new SIP Invite will go out without the SIP preconditions. In cases where SIP preconditions are supported, the call will continue as explained in the Overview of SIP Preconditions, page 11-73.

Cisco does *not* recommend enabling local RSVP fallback. Instead, a different route should be configured to reach the destination. Cisco recommends using a function such as Local Route Group or a similar function to reroute calls that fail RSVP call admission control to a gateway that is local to the calling device in order to extend the call over the PSTN.

Migration from Enhanced Locations Call Admission Control to RSVP SIP Preconditions

When migrating from Enhanced Locations call admission control to RSVP SIP Preconditions, it is important to first follow the migration recommendations in the section on Migrating to RSVP Call Admission Control, page 11-70. Once migration of locations-based call admission control to local RSVP call admission control is complete, RSVP SIP Preconditions can be enabled on the SIP intercluster trunk.

The following steps are required to enable RSVP SIP Preconditions:

Step 1 Configure a SIP intercluster trunk in each cluster, and direct it to the other cluster.

Step 2 Place the SIP intercluster trunks into their own location. All devices must be in a separate location from the SIP intercluster trunk location and have an inter-location RSVP policy of **Mandatory** or **Mandatory (Video Desired)**. The inter-location policy determines the RSVP policy that is sent over the SIP trunk in the preconditions. (See Table 11-7, which lists the Unified CM inter-location policy that corresponds to the equivalent SIP audio and video precondition attributes.)

Step 3 Configure the intra-location RSVP policy of the SIP intercluster trunk to **Mandatory** or **Mandatory (Video Desired)**. Intra-location RSVP policy is accomplished by setting an inter-location RSVP policy of the specified location to itself. This is necessary for calls that are transferred back to the cluster on the same SIP intercluster trunk so that transfer does not fail.

Step 4 Configure the SIP profile of the SIP intercluster trunks on each Unified CM cluster by setting **RSVP Over SIP** to **E2E**, the **Fall back to local RSVP** field to your preference, and the **SIP Rel1XX Options** to **Send PRACK if 1XX contains SDP**.

Note For the SIP trunk configuration, IPv6 is not supported on RSVP SIP Preconditions. Therefore, ensure that the IPv6 enablement checkbox **Enable ANAT for early offer calls** is not checked because it is not supported with RSVP SIP Preconditions.

Note Unified CM ignores the **MTP Required** and **Use TRP** check boxes on the SIP trunk when it is configured for end-to-end RSVP.

As mentioned, the RSVP SIP Preconditions feature allows Unified CM endpoints to establish direct RSVP agent-to-agent reservations across clusters. Figure 11-51 shows the components involved in a call made with RSVP SIP Preconditions.

Figure 11-51 RSVP SIP Preconditions, Distributed Unified CM Deployment Dual Cluster Design

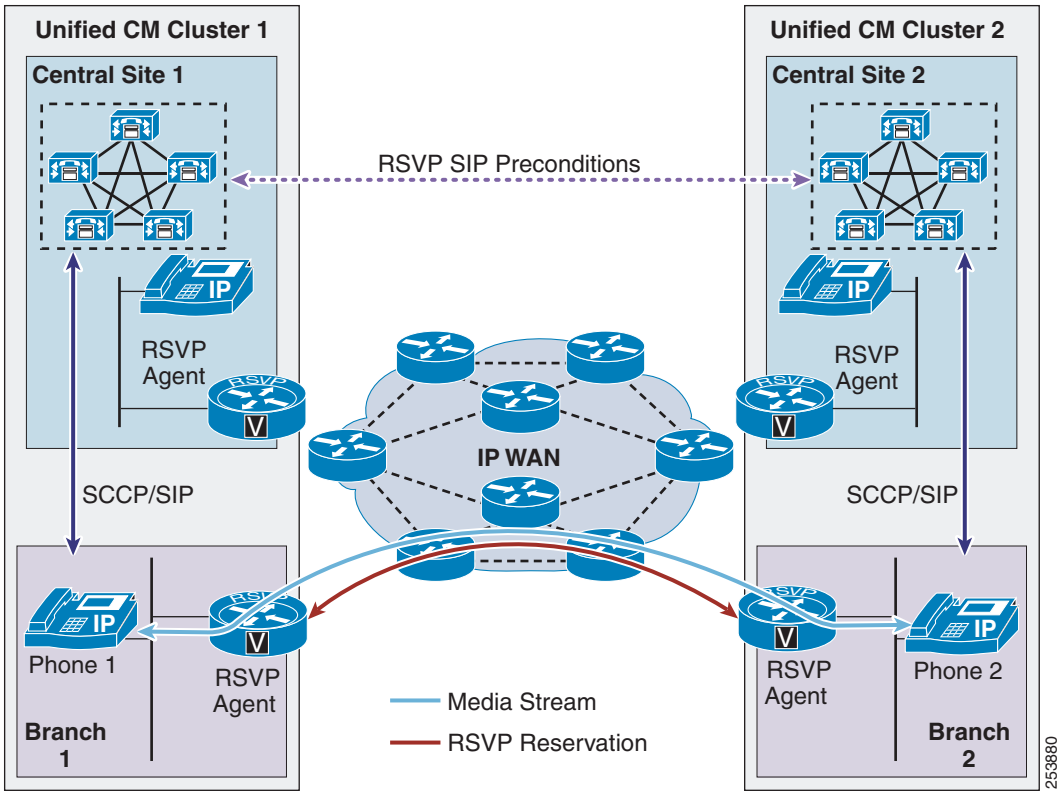


Figure 11-51 illustrates a typical dual cluster deployment with RSVP SIP Preconditions enabled. It includes four locations: Central Site 1, Branch 1, Central Site 2, and Branch 2. The IP WAN connecting the locations can be of any topology type and is not restricted to the hub-and-spoke topology. For any call between two clusters that requires an RSVP reservation in the media path, a Cisco RSVP Agent is invoked dynamically by each Unified CM cluster. The Cisco RSVP Agent acts as a proxy to make an RSVP reservation for the IP phone in the same location with the Cisco RSVP Agent. For example, when phone 1 in Branch 1 calls phone 2 in Branch 2, an RSVP reservation (illustrated as the red line in Figure 11-51) is established between Cisco RSVP Agents in the Branch 1 and Branch 2 locations. This is similar to the media stream setup of a single cluster RSVP-enabled locations solution. The difference here is that the SIP trunk is passing the RSVP policy negotiation between the two Unified CM clusters so that only a single RSVP Agent is allocated per cluster location associated with the respective phones.

There are three call legs for the media streams of this call. The first call leg is between phone 1 and the Branch 1 Cisco RSVP Agent, the second call leg is between the Branch 1 and Branch 2 Cisco RSVP Agents, and the third call leg is between the Branch 2 Cisco RSVP Agent and phone 2. Note that the media streams of a call between two branch locations are not sent through the central site in this case, which is different from a call made over the traditional hub-and-spoke topology using call admission control based on static locations.

There are five call legs for the signaling of this same call. The first call leg is between phone 1 and Unified CM Cluster 1; the second leg is between the Branch 1 Cisco RSVP Agent and Unified CM Cluster 1; the third call leg is between Unified CM Cluster 1 and Unified CM Cluster 2; the fourth is between Unified CM Cluster 2 and the Branch 2 Cisco RSVP Agent; and the fifth and last call leg is between Unified CM Cluster 2 and phone 2.

In Figure 11-51, Phone 1 in Cluster 1 Branch 1 calls a Phone 2 in Cluster 2 Branch 2. The call signaling between the phones and Unified CM could be SCCP or SIP, and the signaling between Unified CMs will be SIP with the RSVP SIP Preconditions feature enabled. When Phone 1 initiates a call to Phone 2, the Cluster 1 server allocates an RSVP Agent to Phone 1 based on the RSVP Agent located in Phone 1's media resource group and list, and it then extends the call over the SIP trunk to Cluster 2 with SIP preconditions (RSVP Policy). The preconditions that are advertised in the SIP INVITE to Cluster 2 are a derivative of the inter-location policy configured between the locations of Phone 1 and the SIP trunk in Cluster 1. So in this case, on Cluster 1 the inter-location policy between locations Branch 1 and HQ is set to Mandatory (Video Desired). For details about Unified CM policy, see RSVP Policy, page 11-69. This inter-location policy determines the policy set on the outbound SIP INVITE to Cluster 2. At this point, Cluster 2 receives a SIP INVITE from Cluster 1 with preconditions set to Mandatory. Cluster 2 then allocates an RSVP Agent to Phone 2 based on its media resource group and list, and also checks the configured locations between the SIP trunk on Cluster 2 and Phone 2 in Branch 2. If this policy is also Mandatory, then Cluster 2 responds with a 183 SESSION PROGRESS message (followed by a PRACK) and starts the RSVP negotiation between the two RSVP Agents in Branch 1 and Branch 2. Once the RSVP Agents have successfully negotiated a reservation for the call, they will inform their respective clusters and the SIP signaling will progress through to ringing stage.

Table 11-7 compares the Unified CM inter-location policy to the equivalent SIP audio and video precondition attribute. (For details about Unified CM RSVP Policy, see RSVP Policy, page 11-69.)

Table 11-7 Unified CM RSVP Policy and Equivalent SIP Preconditions

Unified CM RSVP Policy	SIP Precondition (Audio Call)	SIP Precondition (Video Call)
No Reservation	audio = none	audio = none video = none
Optional (Video desired)	audio = optional	audio = optional video = optional
Mandatory	audio = mandatory	audio = mandatory video = mandatory
Mandatory (Video desired)	audio = mandatory	audio = mandatory video = optional

Unified CM Video Calls with RSVP SIP Preconditions

Unified CM supports video escalation and de-escalation across Unified CM clusters with RSVP SIP Preconditions. Video escalation occurs when an ongoing audio-only call is escalated to video or when a video stream is added to the audio-only call. Conversely, de-escalation is the downgrading or de-escalating of a video call to an audio-only call.

In order to support video escalation and de-escalation across clusters with RSVP SIP Preconditions, Unified CM signals two media lines (or m-lines) in the SIP Preconditions within the SIP Session Description Protocol (SDP), one for the audio stream and one for the video stream. Having separate media lines for both audio and video allows each stream to have its own RSVP policy and status in SIP signaling. Because the audio and video streams have their own precondition attributes Unified CM, RSVP policies can be mapped easily into the preconditions. This function allows Unified CM to pass the successful status of an audio stream reservation while simultaneously passing the failed status of the video stream reservation, the potential of a Mandatory (Video Desired) policy, thus allowing the call to be downgraded from a video call to an audio-only call, without rejecting the call entirely.

The video bandwidth reserved for RSVP SIP Preconditions is set to the value configured between the region pair. In this case that would be the region of the endpoint and the SIP intercluster trunk region. Video bandwidth is then adjusted after the video channel is established. Cisco recommends ensuring a video bandwidth value that is greater than or equal to the expected negotiated bit rate of any two endpoints between the region pairs.

For mid-call video escalation, the video stream will be set up only after having video bandwidth reserved.

During hold/resume of a video call, video and audio bandwidth will continue to be reserved while connecting to music on hold.

For other supplementary services such as transfer, Unified CM triggers video reservation and video stream setup after the audio stream completes setup (this is also known as delayed video escalation).

Example 11-1 Delayed Video Escalation: Call transfer from audio-only to video call with RSVP SIP Preconditions

Video device A in cluster A calls audio device B in cluster B through a SIP trunk with RSVP SIP Preconditions enabled. The call is set up as an audio call whose audio streams are allocated to the audio pool with AudioStream Application ID and PHB (Per Hop Behavior) of EF.

Device B transfers the call to a video device C in cluster B. Audio streams between A and C are first established in the audio pool with AudioStream Application ID and PHB of EF.

Delayed video escalation happens between A and C only after the audio media connection is successful. The video streams are allocated to the video pool with VideoStream Application ID. If the video stream allocation fails, the call will stay as an audio-only call with AudioStream Application ID and PHB of EF. If the video stream reservation succeeds, the audio stream will be re-admitted from the audio pool to the video pool with the VideoStream Application ID. If the re-admission succeeds, then video and audio streams will have the VideoStream Application ID with a PHB of AF41. However, if the re-admission fails, then the video stream will have the VideoStream Application ID with a PHB of AF41 while the audio stream will have the VideoStream Application ID with PHB of 0 (video fail best effort value).

Unified CM and RSVP SIP Preconditions Best Practices and Design Considerations:

- The SIP trunk should always have both an inter-location and an intra-location RSVP policy. The inter-location policy ensures that the correct RSVP policy is set for inbound and outbound calls. The intra-location policy ensures that calls hair-pinned on the same trunk (due to intercluster forward and transfer operations) are ensured an end-to-end RSVP policy.
- Cisco recommends configuring a **Mandatory** or **Mandatory (Video Desired)** RSVP policy because those settings guarantee the bandwidth reservation and the voice quality of the call.
- Cisco recommends configuring the SIP trunk profile with the **SIP Rel1XX Options** field set to **Send PRACK if 1XX contains SDP**. A SIP PRACK message is required for RSVP SIP Preconditions operation, but only for 1XX messages containing SDP.
- Ensure the configuration of each cluster in an RSVP SIP Preconditions deployment is standardized across the solution so that the RSVP cluster service parameters, inter-location policies, and codecs used across the WAN as well as on the RSVP Agent are the same. It is important to ensure there is no mismatch in capabilities or configuration across the clusters when call establishment is being attempted.

- Unified CM with RSVP SIP Preconditions supports termination to shared lines across clusters, subject to the following guidelines and restrictions:
  - When setting up a call across clusters to a shared line, the RSVP reservation occurs between the calling device's RSVP Agent and the first RSVP Agent allocated for the shared line device. (This is not controllable by programming.) All other devices with this shared line in separate locations will only allocate an RSVP Agent and not establish a reservation.
  - One RSVP Agent is allocated to each location where one or more devices of the shared line exist.
  - If the device that had the first RSVP Agent allocated is the device that answers the call, then the call establishment will take place and the RSVP Agents that were allocated to other shared line devices in other locations will be released.
  - If a device that did not have a reservation established answers the call, then a new reservation will be initiated between the calling devices RSVP Agent and the one allocated for the answering device with an optional RSVP policy, and the RSVP Agents will continue to try the reservation for the duration of the call until a reservation is successful. During the time that the call is under an optional policy and the **Mandatory RSVP mid call error handle option** is set to **Call becomes best effort** (default), then the media stream between the two devices will be marked best-effort until a reservation succeeds, in which case the media will be re-marked to a PHB (Per Hop Behavior) value of EF (audio) or AF41 (video).
  - If the device that had the first RSVP Agent allocated fails the RSVP reservation with a Mandatory policy, then neither that device nor any device in that location will be rung. However, the shared line devices in all other location will be rung.
- Based on the above shared line limitations, Cisco recommends restricting a shared line to a group of devices within the same location.
- Unified CM with RSVP SIP Preconditions supports termination to Mobile Connect destinations (remote destinations), subject to the following guidelines and restrictions:
  - Local RSVP: For remote destinations to devices, gateways, or trunks that are remote to the calling device, gateway, or trunk, apply the same rules as explained above in the shared line support.
  - End-to-end RSVP: Remote destinations for any single line should not point to more than one RSVP SIP Preconditions destination. Unified CM supports only one RSVP SIP Preconditions call per line for remote destinations.
- If the MoH server is in the same location as the holding party (the party placing another party on hold), the initial reservation is reused and no new reservation is made.
- Hold/Resume functionality with RSVP SIP Preconditions will break the media streams across endpoints and RSVP agents, but the reservation will still be preserved.
- sRTP is supported with RSVP SIP Preconditions and is negotiated during media setup and after RSVP reservation. Unified CM does not signal RTP/SAVP and crypto attributes during the precondition phase.

- T.38 is supported with RSVP SIP Preconditions and negotiated from SIP, H.323, and MGCP endpoints supporting T.38 fax transmission. Unified CM will negotiate an initial reservation using the inter-region audio bandwidth (between the endpoint and SIP intercluster trunk). After call establishment and upon T.38 switchover, the bandwidth usage will be readjusted to 80 kbps if it is not already set.
  - Limitation: If the inter-region bit rate is set to less than 80 kbps, then after T.38 switchover occurs, the RSVP reservation will be readjusted to 80 kbps. This can cause failure if the new adjusted bandwidth cannot be reserved. In such cases, if the reservation fails after switchover, the call will continue because Unified CM will not signal this failure over the SIP intercluster trunk.
  - For the above reason, when deploying T.38 fax with RSVP SIP Preconditions, Cisco recommends using 80 kbps as the inter-region audio bit rate between T.38 endpoints and the intercluster trunk enabled for RSVP SIP Preconditions.
- To support RSVP SIP Preconditions for supplementary services such as hold/resume, transfer, and conference, media resources such as the music on hold servers, annunciators, and conference bridges must have a local RSVP agent assigned to their respective device pools' media resource group list (MRGL).

Note In various call flows for supplementary services such as hold/resume or transfer and conference, different media resources are brought into the RSVP SIP Preconditions call. Those media resources such as conference bridges, music on hold servers, and annunciators also require an RSVP Agent association, just as any other device would when invoked into a RSVP SIP Preconditions or RSVP-enabled locations call. These media resources obtain an RSVP resource from the media resource group list associated to the configured device pool.

Architecture and Considerations for Extension Mobility Cross Cluster

Extension Mobility Cross Cluster (EMCC) enables users in one cluster to log into IP phones of another cluster. For more detailed information about Extension Mobility Cross Cluster with regards to feature function, high availability, and scalability, see Extension Mobility Cross Cluster (EMCC), page 19-10. The rest of this section covers the EMCC feature as it applies to call admission control. It also assumes an understanding of the information covered in Extension Mobility Cross Cluster (EMCC), page 19-10.

EMCC and RSVP-Enabled Environments

In Unified CM RSVP-enabled deployments with either RSVP-enabled locations (for single cluster or intra-cluster) or RSVP SIP Preconditions (for distributed clusters or inter-cluster), a local RSVP Agent must be invoked into the call flow by Unified CM to accomplish the RSVP signaling on behalf of the IP phone. This is accomplished in EMCC environments by the passing of call control information between the Unified CM clusters to enable an EMCC user logged in remotely to make both intra-cluster and inter-cluster calls with RSVP.

In an EMCC deployment, there are always two clusters for any given login or registration interaction. From the EMCC user's perspective, this would be a home cluster and a visiting cluster (see Figure 11-52). The home cluster is the user's originating cluster, and it is where the user information is stored. The visiting cluster is the phone's originating cluster, where an EMCC user roaming between clusters would log in and where the device information is stored.

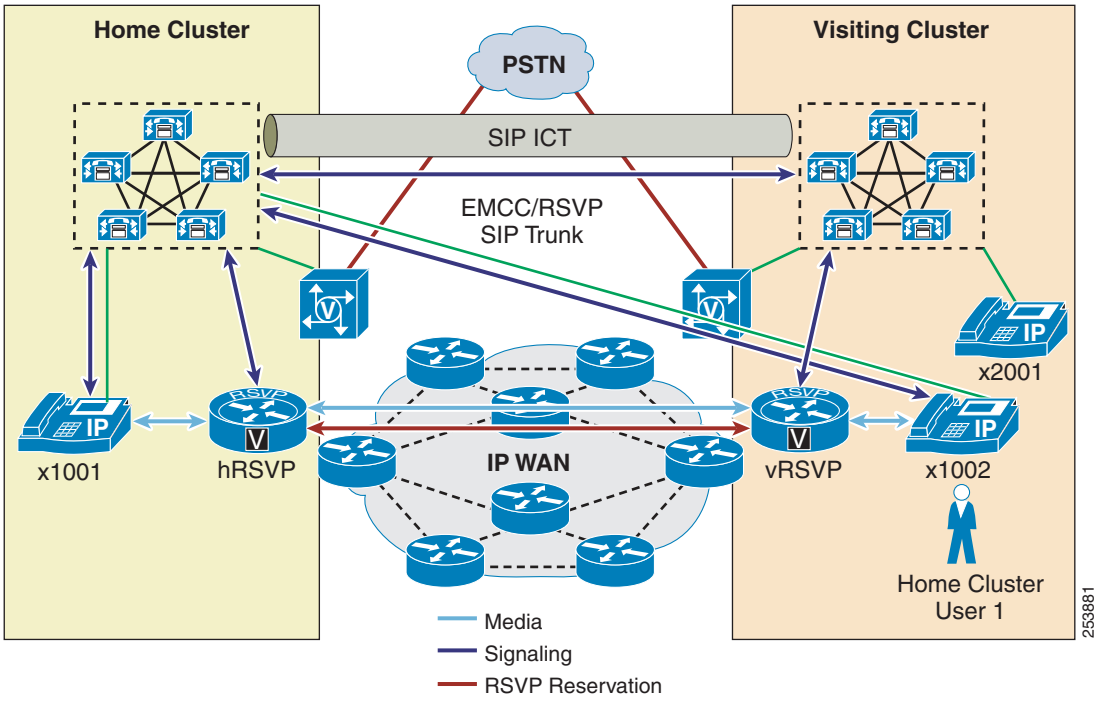
When a user logs into a visiting clusters phone (visiting phone), that phone in turn registers directly with the EMCC user's home cluster. All calls that are then made from that user and visiting phone are made from the call control of the home cluster. The home cluster thus manages the visiting phone and provides this visiting phone with an EMCC roaming device pool. (See Extension Mobility Cross Cluster (EMCC), page 19-10, for further information on EMCC roaming device pools.)

The home cluster makes requests to the visiting cluster for an RSVP Agent when required, and it does this over the EMCC-enabled SIP trunk between the two clusters (specifically in SIP REFER messages). An RSVP Agent is requested from the visiting cluster only when the home cluster has determined that the endpoint requires an RSVP Agent for a call. This is determined by the home cluster from the inter-location RSVP policy between the visiting phone's location (from the EMCC roaming device pool) and the called party's location (from the called party device, gateway, or trunk).

Once the RSVP policy is determined, an RSVP Agent is requested from the visiting cluster for the visiting phone. In this request for the RSVP Agent, the home cluster sends the device name (sep*xxxxxxxxxxxx*) so that the visiting cluster can do a look up on the device name to determine the RSVP Agent (derived from the MRGL on the device itself or on the device pool). Once the home cluster has the information for the RSVP Agent to associate to the visiting phone, it can start the procedures to establish a local RSVP call (RSVP-enabled locations within a cluster) or an RSVP SIP Preconditions call (between clusters).

Figure 11-52 illustrates the signaling and media connections between the various components involved in an EMCC call using RSVP over a SIP-enabled trunk.

Figure 11-52 EMCC Call Using RSVP over a SIP-Enabled Trunk



Best Practices

- Set the location policy between the EMCC roaming device pool location and all other locations to **Mandatory (Video Desired)**.

- RSVP-enabled locations-based call admission control should be functioning prior to enabling EMCC in conjunction with RSVP SIP Preconditions.
- Any IP phone to which an EMCC user can log in must have a local RSVP Agent associated to it.

Unified CM Interoperability and Feature Considerations

This section discusses interoperability considerations between Unified CM and Cisco IOS Gateways and Unified CME.

Cisco IOS Gateway and Unified CME

Both Cisco IOS SIP/TDM gateways and Cisco Unified Communication Manager Express (Unified CME) support RSVP SIP Preconditions. This support enables audio-only calls to be established between Unified CM and the Cisco IOS SIP/TDM gateway or Unified CME signaling RSVP policy over SIP signaling.

For further information regarding SIP RSVP features in Cisco IOS and restrictions for the RSVP SIP Preconditions feature on SIP/TDM Cisco IOS gateways and Unified CME, refer to the *Cisco IOS SIP Configuration Guide*, available at

http://www.cisco.com/en/US/docs/ios/voice/sip/configuration/guide/12_4t/sip_12_4t_book.html

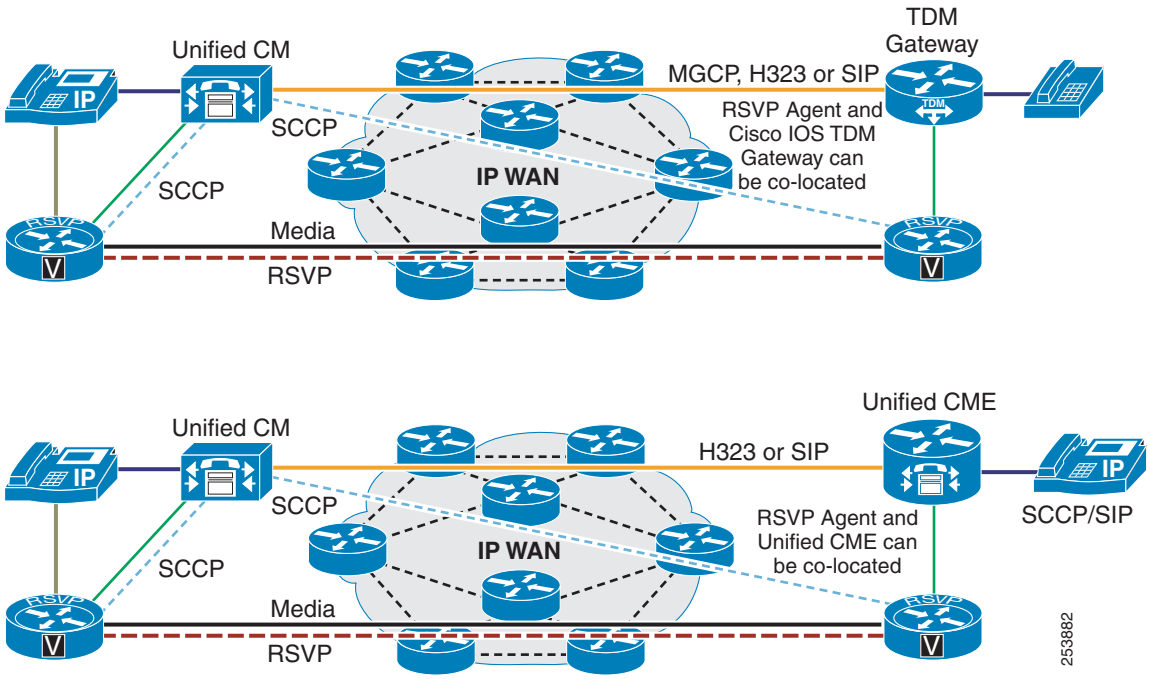
Unified CM has two modes of configuration when interoperating with Cisco IOS gateways and Unified CME in RSVP deployments: local RSVP supporting MGCP, H.323, and SIP call signaling, and end-to-end RSVP supporting SIP signaling only. When interoperating with Cisco IOS gateways and Unified CME, Unified CM can support both of these methods of operation. There are, however, implications when using one method or the other, as described in the following sections.

Unified CM and Local RSVP with Cisco IOS Gateways and Unified CME

In local RSVP mode, Unified CM supports interoperating with Cisco IOS TDM gateways over MGCP, H.323, or SIP call signaling protocols and with Unified CME over H.323 or SIP. In this mode, Unified CM allocates an RSVP agent to the Cisco IOS TDM gateway for calls established to or from the gateway, and it does not signal preconditions or RSVP policy to the Cisco IOS TDM gateway. This is the default configuration for MGCP, H.323, and SIP in Unified CM.

Figure 11-53 illustrates local RSVP integration of Unified CM with a Cisco IOS TDM gateway and with Unified CME.

Figure 11-53 Local RSVP Integration of Unified CM with a Cisco IOS TDM Gateway and Unified CME



Advantages

This model provides the following advantages:

- Support for a wide variety of Cisco IOS gateway signaling protocols (MGCP, H.323, SIP).
- Support for both SIP and SCCP Unified CME endpoints.
- Centralized administration of RSVP policy and Application ID from Unified CM.
- With MGCP for a Cisco IOS TDM gateway, the media path is optimized in call transfer and forward supplementary service scenarios. For calls transferred or forwarded from the local system (Cisco IOS TDM gateway), both media and signaling are torn down and re-established to the transferred or forwarded party.

Disadvantages

This model has the following disadvantages:

- It uses RSVP Agent sessions (software or hardware, depending on session requirements and functionality such as session transcoding).
- With H.323 and SIP integrations for Cisco IOS TDM gateways and Unified CME, the media path is not optimized in transfer and forward supplementary service scenarios. This means that calls transferred or forwarded from the local system (Cisco IOS TDM gateway or Unified CME) hairpin both media and signaling on the local system, which results in double bandwidth consumption.

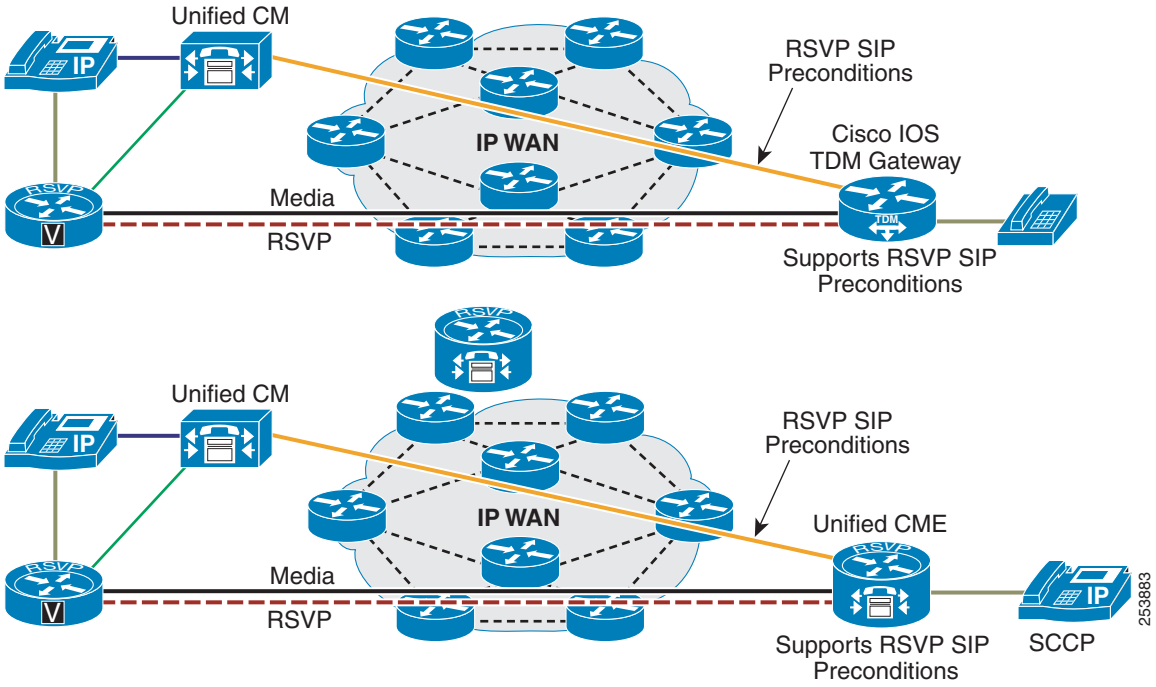
In this method, intra-cluster call admission control functions as explained in the section on Unified CM RSVP-Enabled Locations, page 11-62.

Unified CM and End-to-End RSVP or RSVP SIP Preconditions with Cisco IOS Gateway and Unified CME

In end-to-end RSVP mode, Unified CM supports interoperating with Cisco IOS gateways and Unified CME using RSVP SIP Preconditions signaling. In this mode, Unified CM does not allocate an RSVP agent. The Cisco IOS gateway or Unified CME natively supports RSVP. This method reduces the usage of RSVP agent software sessions on a Cisco Integrated Services Router (ISR).

Figure 11-54 illustrates an RSVP SIP Preconditions integration between Unified CM and a Cisco IOS TDM gateway or Unified CME.

Figure 11-54 RSVP SIP Preconditions Integration Between Unified CM and Cisco IOS Gateway or Unified CME



Advantages

This model provides the following advantages:

- Support for RSVP SIP Preconditions.
- Does not use RSVP Agent resources for SIP Cisco IOS TDM gateways or Unified CME.

Disadvantages

This model has the following disadvantages:

- It supports only SCCP Unified CME endpoints.
- It supports only SIP trunk implementations.
- The media path is not optimized in transfer and forward supplementary service scenarios. This means that calls that are transferred from the local system (SIP Cisco IOS TDM gateway or Unified CME) hairpin both media and signaling on the local system, which results in double bandwidth consumption.

Design Considerations for Unified CM Interoperability with SIP Cisco IOS TDM Gateway and Unified CME

When choosing between local RSVP and end-to-end RSVP deployments, determine the best option based on following criteria:

- The desired call signaling protocol (H.323, MGCP, or SIP). This could be based on many requirements outside of the scope of call admission control, such as dial-plan, PBX interoperability, and call signaling features, to name a few.
- Required supplementary services of call transfer and forward to destinations remote to the local system (SIP Cisco IOS TDM gateway and Unified CME). For example, these services might be required for forwarding of calls over the WAN to centralized voice messaging environments.
- Administration of the solution. Decide between centralized or distributed management of RSVP policy and application ID.
- Resource utilization. Consider the utilization of RSVP Agent sessions versus native RSVP. In some cases the number of sessions might require a dedicated platform and thus cannot reside on the SIP Cisco IOS TDM gateway or Unified CME.
- The SIP trunk configured on Unified CM pointing to the SIP Cisco IOS TDM gateway or Unified CME supporting RSVP SIP Preconditions should always have both an inter-location and an intra-location RSVP policy. The inter-location policy ensures that the correct RSVP policy is set for inbound and outbound calls. The intra-location policy ensures that calls hair-pinned on the same trunk (due to forward and transfer operations) are ensured an end-to-end RSVP policy.
- In Unified CM, Cisco recommends configuring a single separate location that can be applied to all Cisco IOS TDM gateways and Unified CMEs configured on a single cluster. That location should have an inter-location RSVP policy set to **Mandatory** or **Mandatory (Video Desired)** with all other locations including itself. An RSVP policy is required for the correct functioning of RSVP SIP Preconditions in these environments.

Note Even IP phones that are in the same physical LAN as the SIP Cisco IOS TDM gateway require an RSVP policy between their location and the location on the SIP trunk. This will utilize RSVP Agent resources for the IP phones but will not deduct bandwidth over the WAN because the RTP stream remains local.

- Ensure that the RSVP policy configured on Unified CM matches the policy configured on the Cisco IOS TDM gateway. Use the following options under the **dial-peer** configuration when enabling RSVP reservations for the SIP Cisco IOS TDM gateway or Unified CME:

```
req-qos guaranteed-delay audio
acc-qos guaranteed-delay audio
```

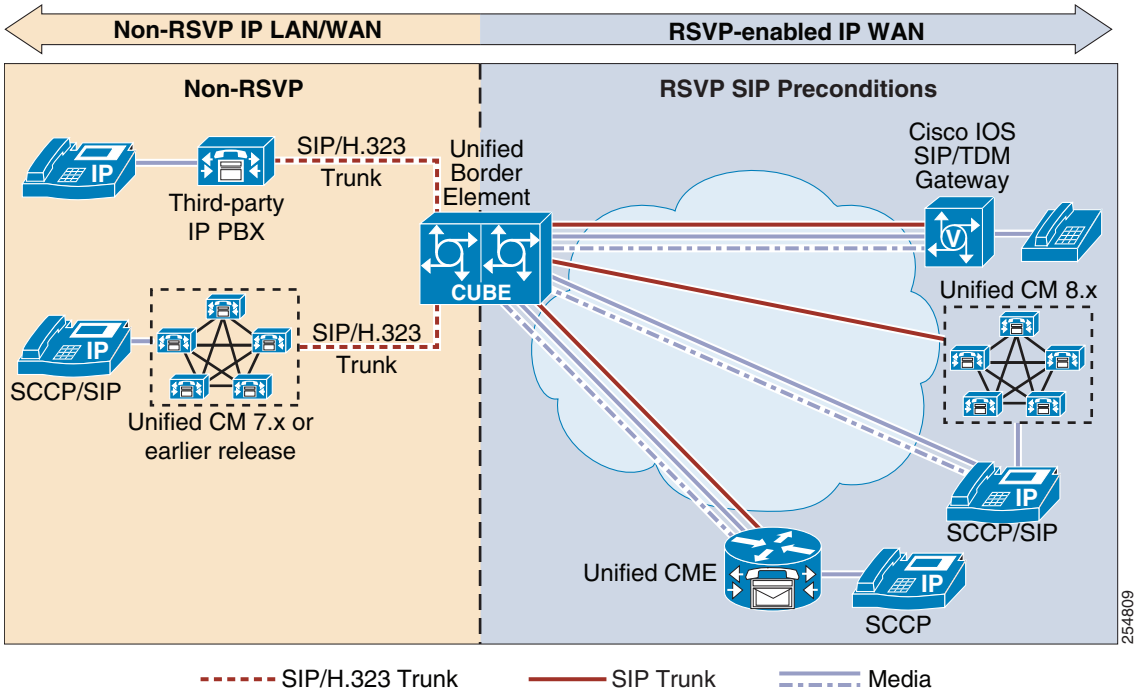
  This configuration ensures that, for each voice call, the SIP Cisco IOS TDM gateway will request an RSVP reservation using the guaranteed delay service. The fact that both the requested QoS and the acceptable QoS specify this RSVP service means that the RSVP reservation is mandatory for the call to succeed (that is, if the reservation cannot be established, the call will fail).
- Ensure that the Application ID configured in Unified CM matches the Application ID configured on the Cisco IOS TDM gateway and Unified CME.
- Ensure that inbound and outbound dial peers are correctly matched to ensure that the appropriate dial peers configured with SIP preconditions are utilized. For further information, refer to the *Cisco IOS SIP Configuration Guide*, available at

  http://www.cisco.com/en/US/docs/ios/voice/sip/configuration/guide/12_4t/sip_12_4t_book.html

Cisco Unified Border Element and RSVP SIP Preconditions

With Cisco Unified Communications System 8.5 and later releases, Cisco Unified Border Element offers audio-only call support for RSVP SIP Preconditions. This support enables enterprises to use the Unified Border Element to integrate non-RSVP call control applications into an RSVP SIP Preconditions infrastructure. On the non-RSVP side of the call control, the Unified Border Element supports integrations with both H.323 and SIP. On the RSVP side of the call, SIP can be used with RSVP Preconditions to integrate with RSVP SIP Preconditions call control such as Unified CM, Unified CME, and SIP-TDM Cisco IOS Gateways. Figure 11-55 illustrates this type of interworking.

Figure 11-55 Cisco Unified Border Element with RSVP over SIP Trunks



For SIP integrations on the non-RSVP side of call control, the Unified Border Element provides support for either Early Offer or Delayed Offer; and for H.323 integrations, either Fast Start or Slow Start is also supported. On the RSVP SIP Preconditions side of call control, SIP Early Offer is always sent in order to provide the preconditions required to negotiate the RSVP policy across call control applications; therefore, RSVP SIP Preconditions is always Early Offer.

For more information, refer to the *Cisco IOS SIP Configuration Guide*, available at

http://www.cisco.com/en/US/docs/ios/voice/sip/configuration/guide/12_4t/sip_12_4t_book.html

Service Advertisement Framework (SAF) and Call Control Discovery (CCD) Considerations

The Cisco Service Advertisement Framework (SAF) enables networking applications to advertise and discover information about networked services within an IP network. Call Control Discovery (CCD) uses SAF to distribute and maintain information about the availability of internal directory numbers (DNs) hosted by call control agents such as Unified CM and Unified CME. CCD also distributes the corresponding number prefixes that allow these internal directory numbers to be reached via the PSTN ("To PSTN" prefixes).

This section discusses SAF CCD as it relates to RSVP SIP Preconditions deployments. For more information on the Service Advertisement Framework and Call Control Discovery, refer to the chapters on Network Infrastructure, page 3-1, Unified Communications Deployment Models, page 5-1, and Dial Plan, page 9-1.

SAF CCD and RSVP SIP Preconditions work together to provide an easy to manage dynamic dial plan for moves, adds, and changes and a topology-aware call admission control method for complex, multi-homed, multi-tiered networks, thus providing a dynamic replacement for a static gatekeeper infrastructure for both dial plan resolution and call admission control that reacts to changes in the network.

SAF CCD works together with RSVP SIP Preconditions call admission control to ensure that there is an alternate route to reach the destination in case of reservation failure. This function is referred to as Call Control Discovery automatic PSTN failover.

Call Control Discovery Automatic PSTN Failover

SAF CCD is different than standard call routing in that only a single IP route can be chosen for any given call; whereas with standard call routing, multiple IP paths may be defined and consecutively attempted for a single call by using route lists and route groups. For any call made using SAF learned routes, the following options exist:

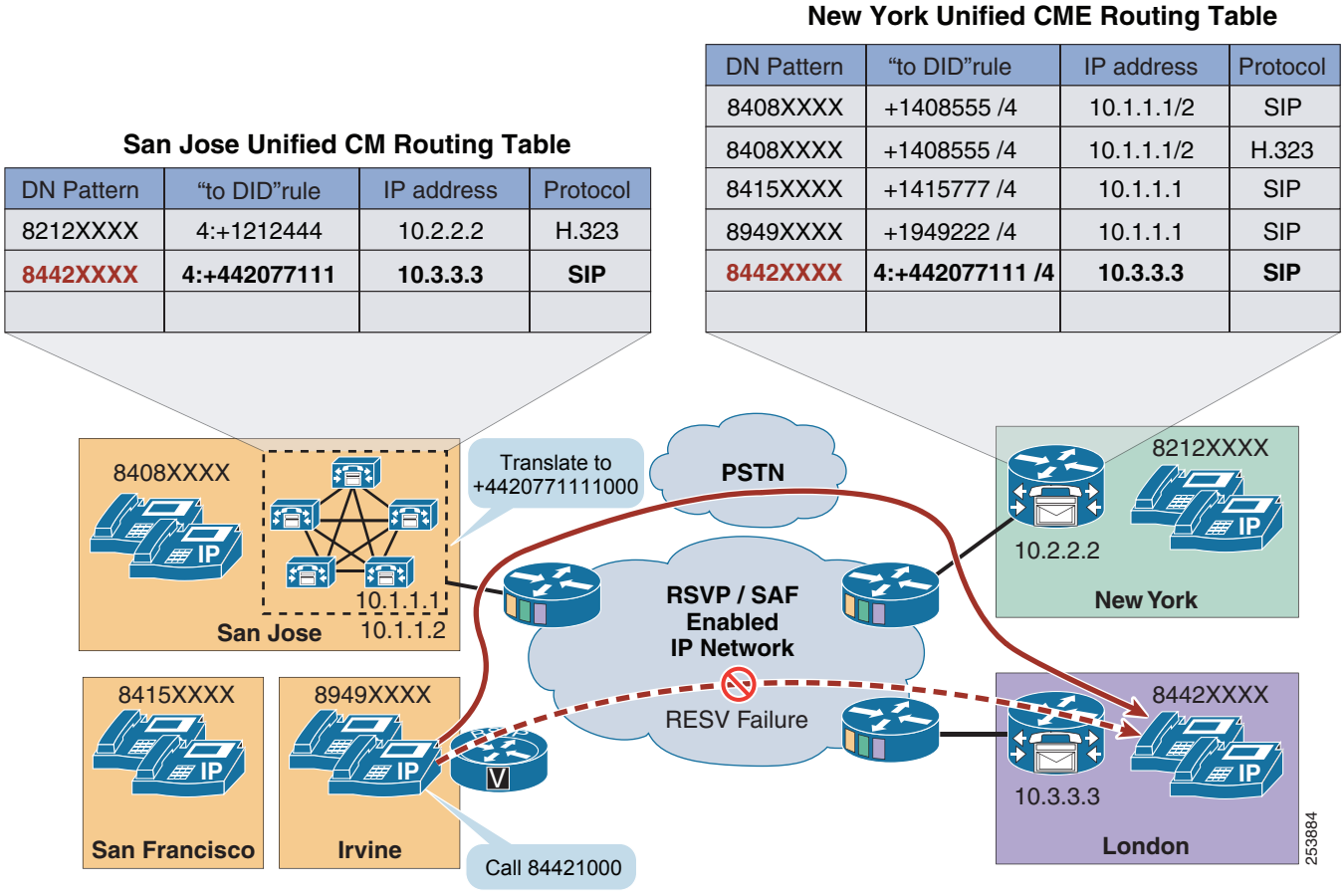
- Take the selected IP path to reach the called number.
- If the IP path is not available, use the PSTN prefix to modify the called number as well as the automatic alternate routing (AAR) calling search space (CSS) from the calling device and route the call via the PSTN.

When RSVP SIP Preconditions is used on a SAF learned route, and if the reservation succeeds, then the call will be established according to normal call establishment on the IP path with RSVP. However, if the reservation fails over the IP route with a returned call termination cause code of Precondition Failure or Reservation Failure (SIP message 580) or Precondition Unsupported (SIP message 420), CCD automatic PSTN failover will occur. (It can also occur on other call termination cause codes, as described below.) CCD automatic PSTN failover is similar to automated alternate routing (see Automated Alternate Routing, page 9-117) in that, when a SAF CCD IP route fails due to a call admission control failure, CCD automatic PSTN failover occurs much like AAR would occur in an intracluster call admission control failure scenario. However, CCD automatic PSTN failover is different from AAR in that it can also occur for other routing failures apart from call admission control. CCD automatic PSTN failover will occur on any call to a learned pattern that fails prior to the alerting phase of the call and that has a call termination cause code other than normal call clearing, user busy, destination out of order, unallocated number, or geo-location mismatch.

CCD automatic PSTN failover uses the AAR CSS as well as the "To PSTN" prefixes (distributed by CCD) to reroute the call. This allows the administrator to leverage the same Class of Service used for AAR call rerouting for local call admission control as for CCD automatic PSTN failover. A key difference is that CCD automatic PSTN failover uses a prefix provided by SAF CCD distribution ("To PSTN" prefix) and not the AAR prefixes.

Figure 11-56 illustrates CCD automatic PSTN failover after RSVP SIP Preconditions call admission control failure.

Figure 11-56 CCD Automatic PSTN Failover with RSVP SIP Preconditions



There is also a difference in functionality between a CCD automatic PSTN failover from a SAF CCD learned route and a reroute by route list and route group functionality with a static route pattern. With a static route pattern pointing to a route group and route list, when a RSVP SIP Preconditions reservation failure occurs, Unified CM routes the call to the next trunk or gateway configured in the route group and list.

In either case (using SAF or a static route pattern), Cisco recommends ensuring that the next choice after call admission control failure is to route the call to a local route group. (See Local Route Group, page 9-13, for more information on local route groups.) This has to be done in the constructs of the CCD automatic PSTN failover function. It is important to ensure that the correct calling search space is configured in the AAR CSS of the calling party to ensure that, when the CCD automatic failover occurs, the call is directed to a route pattern that will engage the local route group function and route the call to the local gateway. This route pattern can be a catch-all pattern used specifically for all CCD automatic failover conditions to ensure the routing of calls out the gateway local of the calling party.

Cisco Unified SIP Proxy Considerations

The Cisco Unified SIP Proxy is a high-performance, highly available stateless Session Initiation Protocol (SIP) server for centralized routing and SIP signaling normalization. By forwarding requests between call control domains, the Cisco Unified SIP Proxy provides the means for routing sessions within enterprise and service provider networks. The main purpose for the Unified SIP Proxy in Cisco Unified Communications deployments is the aggregation of SIP signaling, SIP normalization, and dial plan centralization. For more information on the Cisco Unified SIP Proxy and its features and functions, refer to the documentation at

http://www.cisco.com/en/US/prod/collateral/modules/ps2797/data_sheet_c78-521390_ps2797_Products_Data_Sheet.html

In a RSVP SIP Preconditions environment, the Unified SIP Proxy simply passes along the preconditions that are contained the SDP portion of various SIP messages and does not modify the preconditions in any way.

Adaptive Security Appliance (ASA) Considerations

When deploying a Cisco ASA between any of the Cisco Unified Communications call processing applications such as Unified CM, Unified CME, Unified SIP Proxy, or Cisco IOS SIP/TDM gateway with RSVP SIP Preconditions, both of the following inspections are required:

- SIP Inspection

  The SIP inspection allows the SIP signaling from any Cisco SIP signaling product to traverse the ASA. The ASA subsequently opens the appropriate media pinholes that are recorded in the SDP of the various SIP messages. This is important when the ASA is in the media path between RSVP Agents that are reserving bandwidth and sourcing media flows for Unified Communications endpoints.

- IP Options Inspection

  The IP Options inspection allows the RSVP signaling from RSVP Agent to RSVP Agent to traverse the ASA. All RSVP messages have the IP Router Alert Option set in the IP header of every packet. The ASA drops these packets by default unless the IP Router Alert Option is allowed in the IP Options inspection so that these packets are allowed to flow through the ASA.

Support for both SIP inspection and IP Options inspection specific to RSVP SIP Preconditions implementation is provided in ASA Software Release 8.3. Therefore, for compatibility reasons you must use ASA 8.3 or later software release in any RSVP SIP Preconditions deployment where the ASA is required to inspect the SIP signaling and/or pass RSVP packets.

For information on configuring the ASA for the SIP and IP Options inspections, refer to the *Cisco ASA 5500 Series Configuration Guide*, available at

http://www.cisco.com/en/US/products/ps6120/products_installation_and_configuration_guides_list.html

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 Locations 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 Locations 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.

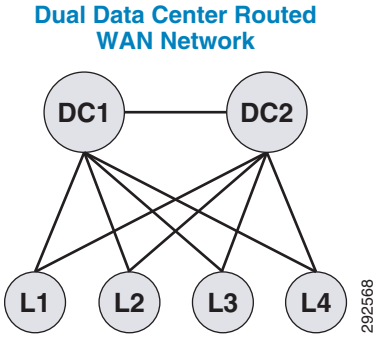
To learn how Enhanced Locations CAC functions and how to design and deploy Enhanced location CAC, see the section on Unified CM Enhanced Locations Call Admission Control, page 11-12.

In this section explores a few typical topologies and explains how Enhanced Locations CAC can be designed to manage them.

Dual Data Center Design

Figure 11-57 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.

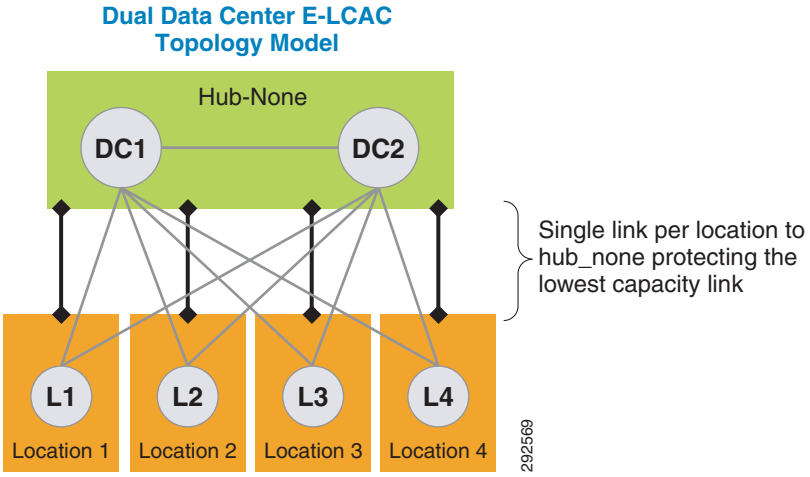
Figure 11-57 Dual Data Center WAN Network



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 Locations 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 Locations CAC and configure a single link to each remote site location. Figure 11-58 illustrates an Enhanced Locations (E-L) CAC locations and links overlay.

Figure 11-58 Enhanced Locations 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 E-L 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 11-59, 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 11-59 Single MPLS Cloud

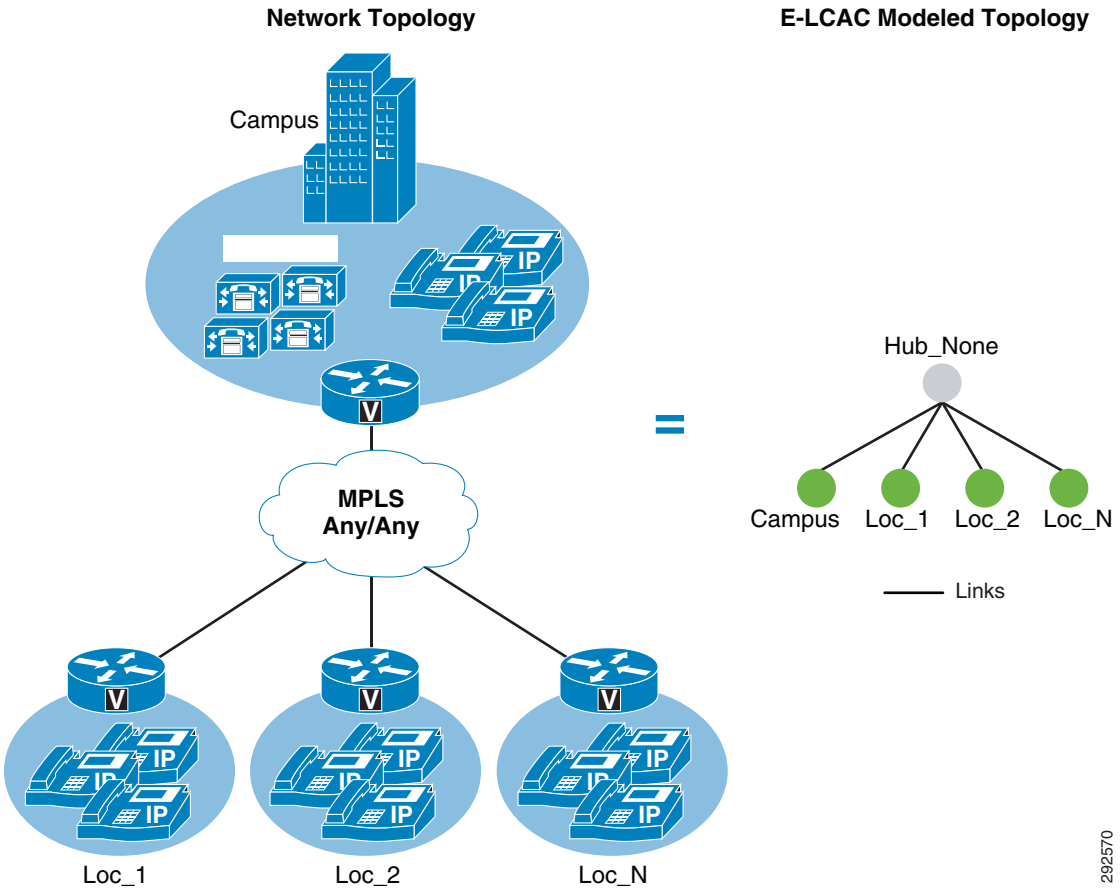


Figure 11-60 shows two MPLS clouds that serve as transit locations interconnecting the campus location where servers, endpoints, and devices are located, with remote locations where only endpoints and devices are located. The campus also connects to both clouds. Each link to the MPLS cloud from the remote 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 11-60 Separate MPLS Clouds

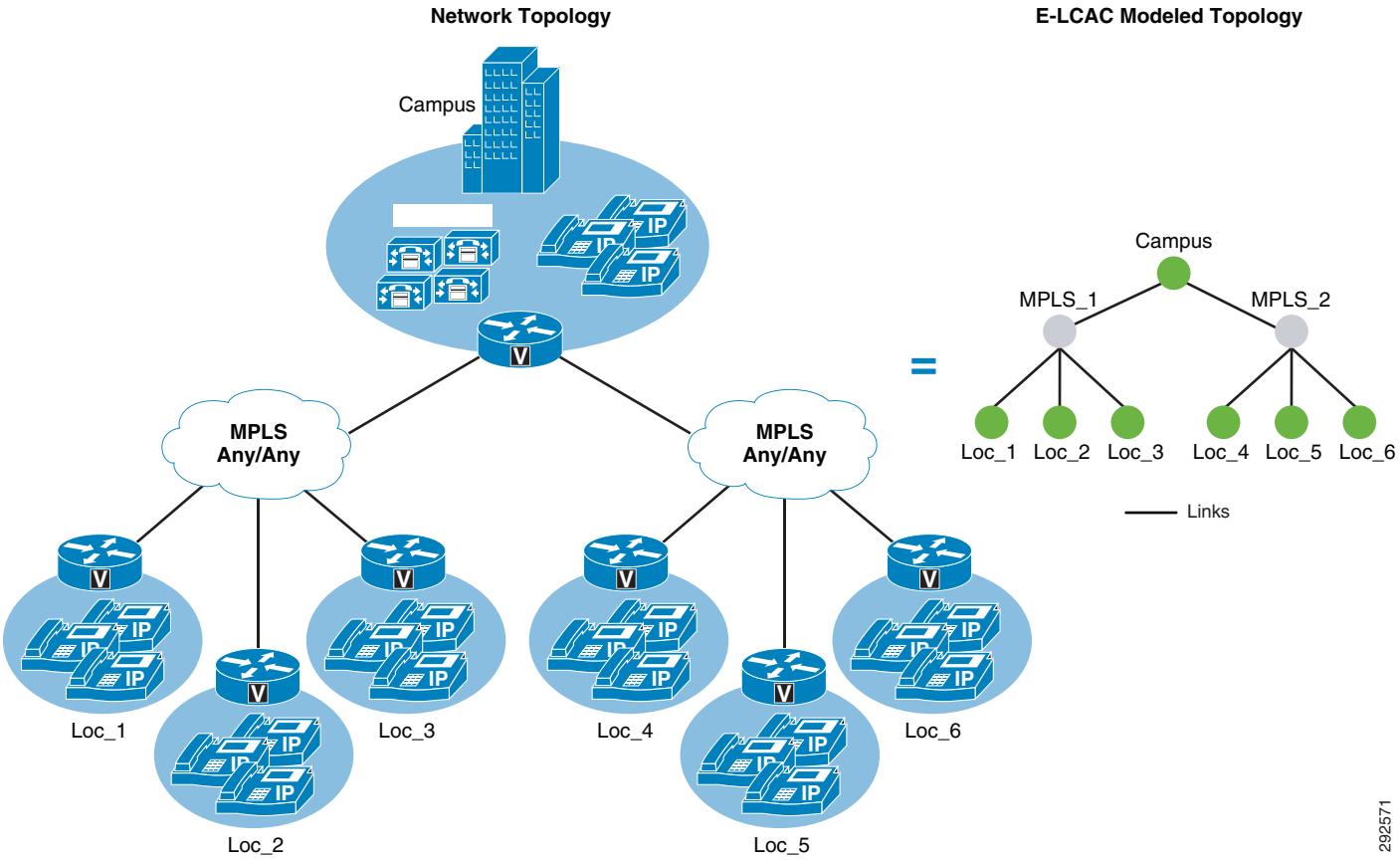
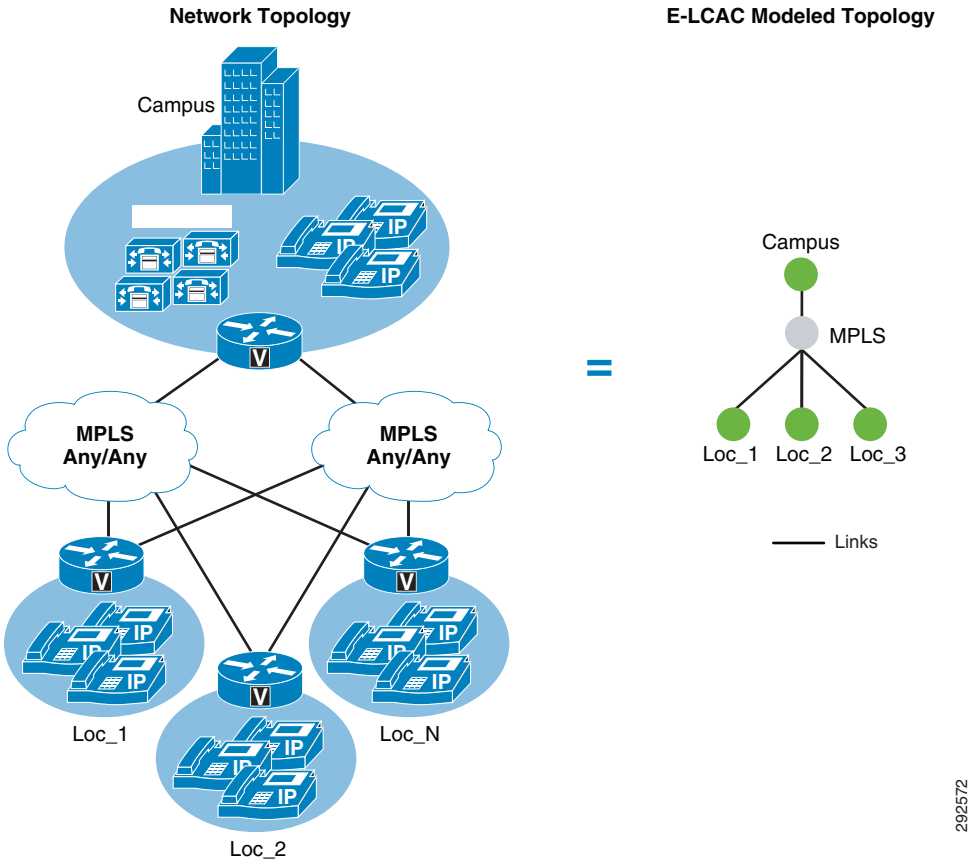


Figure 11-61 shows multiple MPLS clouds from different providers, where each site has one connection to each cloud and uses the MPLS clouds in either an equal-cost load-balanced manner or in a primary/backup scenario. In any case, this design is equivalent to the dual data center design where a single location represents both clouds and a single link represents the lowest capacity link of the two.

Figure 11-61 Remote Sites Connected to Dual MPLS Clouds



Design Recommendations

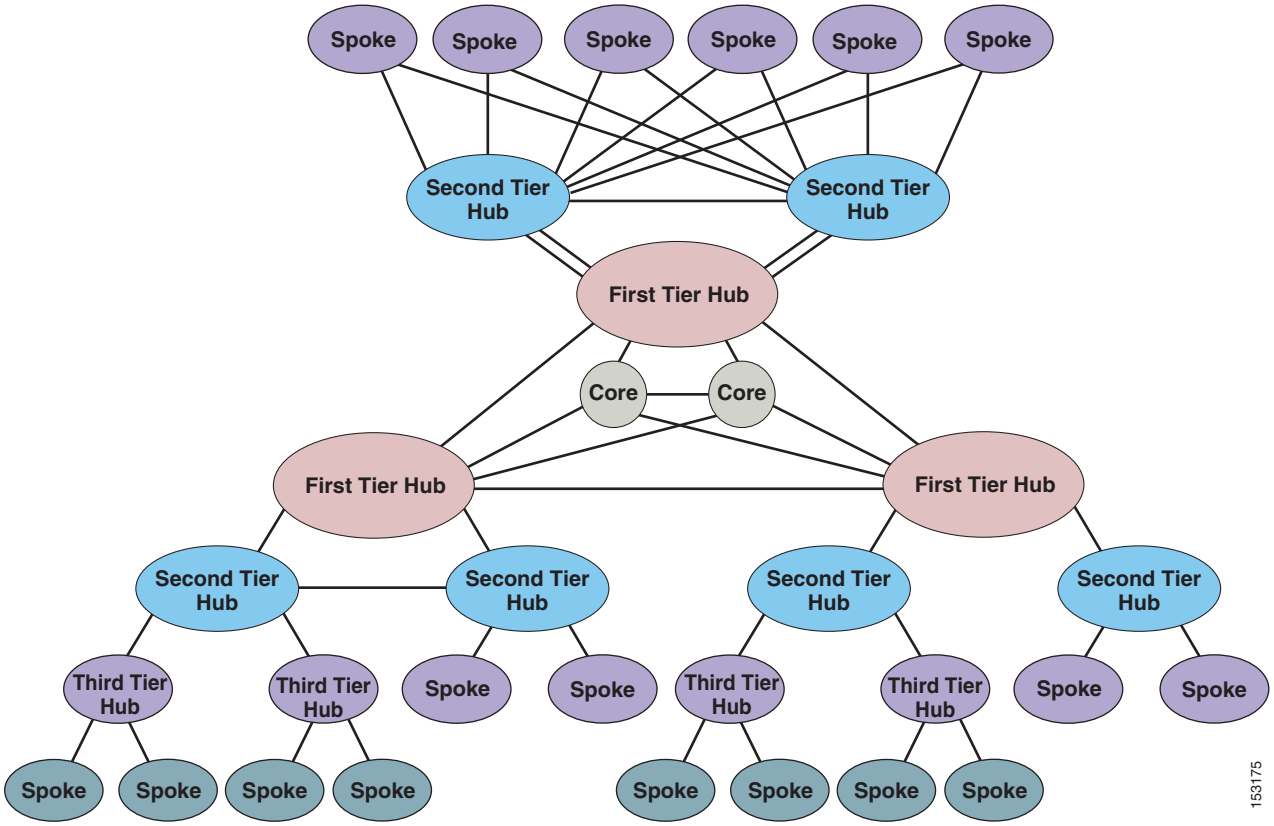
- The MPLS cloud should be configured as a location that does not contain any endpoints but is used as a hub to interconnect locations.
- The MPLS cloud serves as a transit location for interconnecting multiple variable-sized bandwidth WAN uplinks to other remote locations.
- Remote sites with connectivity 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.

Generic Topologies

In the context of this chapter, a generic topology is a network topology that cannot be reduced to a simple hub-and-spoke, a two-tier hub-and-spoke, or a simple MPLS-based network.

As Figure 11-62 illustrates, a generic topology can present full-mesh features, hub-and-spoke features, partial-mesh features, or possibly all of them combined in a single network. It may also present dual connections between sites, as well as multiple paths from one site to another.

Figure 11-62 A Generic Topology



The complex nature of these networks requires the adoption of topology-aware call admission control mechanisms based on RSVP. In particular, these mechanisms can properly control bandwidth in presence of any of the following topology aspects:

- Remote sites dual-homed to different hub sites
- Multiple IP WAN links between any two sites, either in a primary/backup configuration or in an active/active load-balanced configuration
- Redundant hubs or data centers with a dedicated connection
- Fully-meshed core networks
- Multiple equal-cost IP paths between any two sites
- Multi-tiered architectures

The remainder of this section contains design best practices for generic network topologies according to the Unified CM deployment model adopted:

- Centralized Unified CM Deployments, page 11-99

  One or more Unified CM clusters are located at a given site, but only endpoints and gateways are located at all other sites.

- Distributed Mixed Call Processing Deployments, page 11-104

  Call control applications are distributed in various topologies.

Centralized Unified CM Deployments

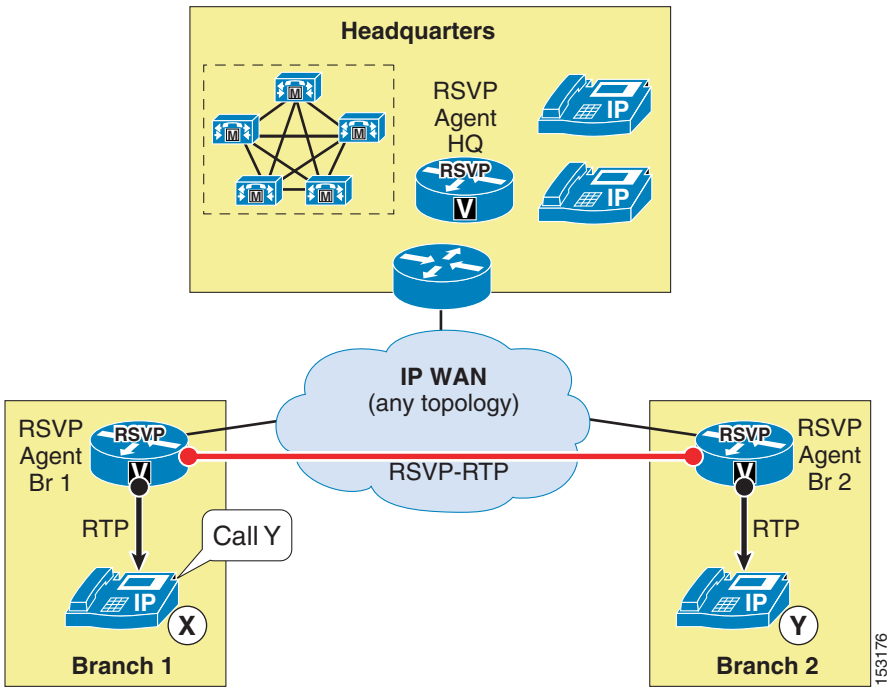
Centralized Unified CM deployments using a generic topology can be categorized into two sub-types:

- Single Unified CM Cluster, page 11-99
- Co-Located Unified CM Clusters, page 11-100

Single Unified CM Cluster

The recommendations in this section apply to a single Unified CM cluster deployed in a generic network topology, as illustrated in Figure 11-63.

Figure 11-63 A Single Unified CM Cluster in a Generic Topology



The following guidelines apply to this type of deployment:

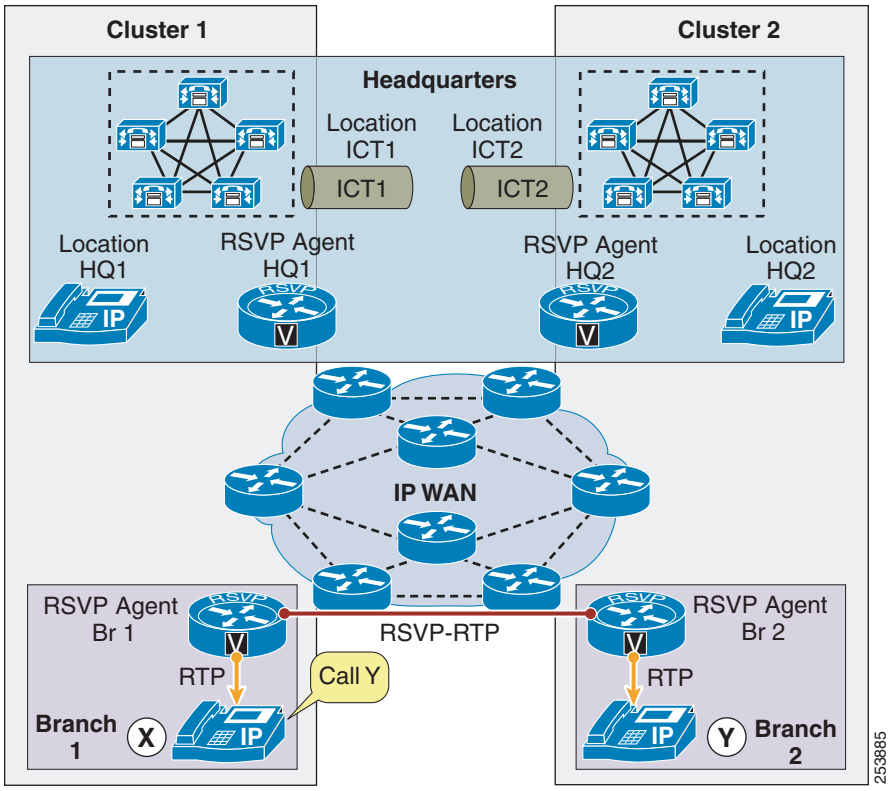
- Enable the Cisco IOS RSVP Agent feature on a Cisco IOS router at each site, including the central site where Unified CM resides. At smaller sites, this router may coincide with the IP WAN router and PSTN gateway, while at larger sites they may be different platforms.
- In Unified CM, define a location for each site, and leave all bandwidth values as **Unlimited**.
- Assign all devices located at each site to the appropriate location (this includes endpoints, gateways, conferencing resources, and the Cisco RSVP Agents themselves).
- Ensure that each Cisco RSVP Agent belongs to a media resource group (MRG) contained in the media resource group list (MRGL) of all devices at that site.
- In the Unified CM service parameters, set the **Default inter-location RSVP Policy** to **Mandatory** or **Mandatory (video desired)** and set the **Mandatory RSVP mid-call error handle option** to **Call fails following retry counter exceeded**.
- Enable RSVP on every WAN router interface in the network where congestion might occur, and configure the RSVP bandwidth based on the provisioning of the priority queue. (See RSVP Design Best Practices, page 11-57.)
- If you need to provision bandwidth separately for voice and video calls, also configure an RSVP application ID on the same WAN router interfaces.
- If the Cisco RSVP Agent is not co-resident with the IP WAN router, enable RSVP on the LAN interfaces connecting the agent to the WAN router.

Co-Located Unified CM Clusters

The recommendations in this section apply to deployments where multiple Unified CM clusters are located on the same LAN or MAN. However, the same considerations may also be valid if the sites where the Unified CM clusters reside are connected via a lower bandwidth link. Due to the design, any call cluster-to-cluster will engage an RSVP Agent for an endpoint in each cluster.

Figure 11-64 illustrates a deployment with two Unified CM clusters located at a given site (HQ) and a number of remote sites with endpoints and gateways, which are controlled either by Cluster 1 (for example, Branch 1) or Cluster 2 (for example, Branch 2).

Figure 11-64 Co-Located Unified CM Clusters in a Generic Topology



The following guidelines apply to the deployment in Figure 11-64:

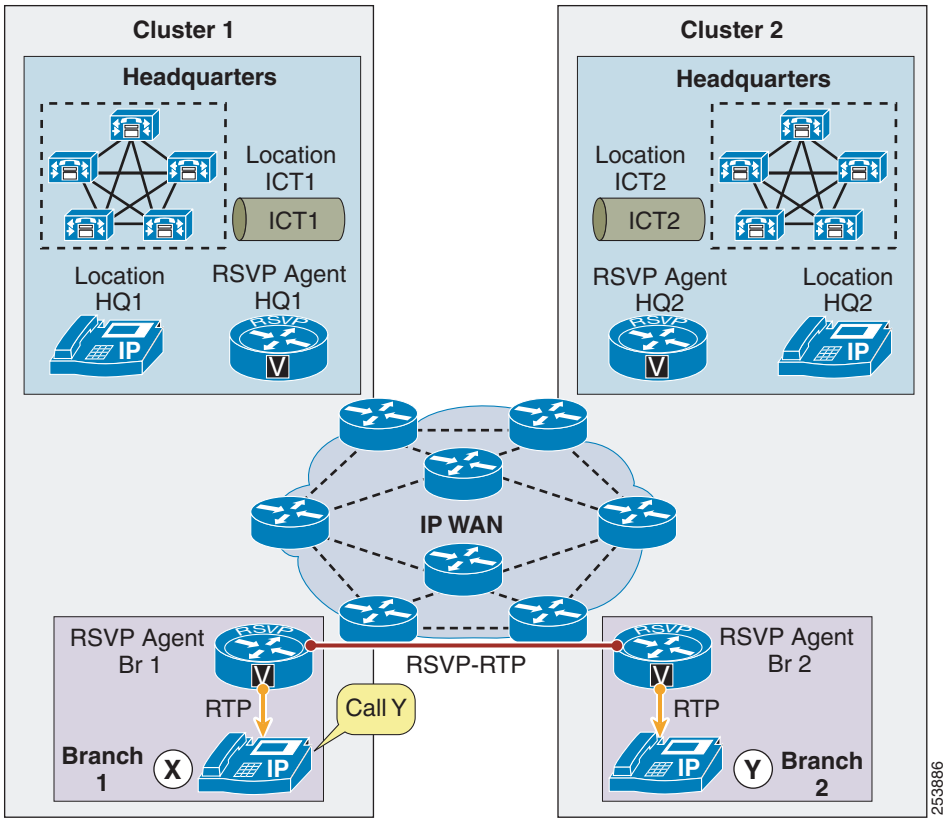
- Enable the Cisco IOS RSVP Agent feature on a Cisco IOS router at each site, including the central site where Unified CM resides. At smaller sites, this router may coincide with the IP WAN router and PSTN gateway, while at larger sites they may be different platforms.
- Depending on the amount of call traffic between the central site and remote sites and between clusters, consider co-locating the RSVP Agents for both clusters at the central site on a single or multiple Cisco Integrated Services Routers (ISR). A single ISR can host multiple RSVP Agents controlled by different clusters.
- In each Unified CM cluster, define a location for each site, and leave all bandwidth values as unlimited.
- Assign all devices located at each site to the appropriate location. This includes endpoints, gateways, conferencing resources, and the Cisco RSVP Agents themselves.
- Ensure that each Cisco RSVP Agent belongs to a media resource group (MRG) contained in the media resource group list (MRGL) of all devices at that site.
- In the Unified CM service parameters, set the **Default inter-location RSVP Policy** to **Mandatory** or **Mandatory (video desired)** and set the **Mandatory RSVP mid-call error handle option** to **Call fails following retry counter exceeded** on both clusters.
- Enable RSVP on every WAN router interface in the network where congestion might occur, and configure the RSVP bandwidth based on the provisioning of the priority queue. (See RSVP Design Best Practices, page 11-57.)

- If you need to provision bandwidth separately for voice and video calls, also configure an RSVP application ID on the same WAN router interfaces.
- If the Cisco RSVP Agent is not co-resident with the IP WAN router, enable RSVP on the LAN interfaces connecting the agent to the WAN router.
- Ensure that RSVP SIP Preconditions are enabled on the SIP intercluster trunks (see Migration from Enhanced Locations Call Admission Control to RSVP SIP Preconditions, page 11-78, for the steps).
- Ensure that an inter-location RSVP policy of **Mandatory** or **Mandatory (video desired)** is set between the SIP intercluster trunk location and all locations, including itself (see intra-location RSVP policy below).
- Ensure that an intra-location RSVP policy of **Mandatory** or **Mandatory (Video Desired)** is set on the SIP intercluster trunk. An intra-location RSVP policy is set by selecting the location and configuring a policy for itself. This effectively ensures that calls within this location will engage RSVP call admission control. This is important for calls hairpinned on the trunk from call transfers or forwards back to the originating cluster.
- For calls from cluster to cluster within the central site, RSVP Agents will be engaged. This is due to the design and the ability for supplementary services to function across clusters, keeping intact end-to-end RSVP across the clusters. There are a few variations that can be supported. Consult with your Cisco account team to determine the best possible call admission control design for co-located clusters.

Distributed Unified CM Deployments

RSVP SIP Preconditions provides call admission control for distributed deployments of Unified Communication Manager clusters in a generic network topology. This section contains an example of a dual-cluster deployment with RSVP SIP Preconditions support between the clusters (see Figure 11-65). Hybrids and variations of this model are expected, and this is only a high-level example of a simple design, with best practices and design considerations called out.

Figure 11-65 Distributed Unified CM Deployment



The following guidelines apply to the deployment in Figure 11-65:

- Enable the Cisco IOS RSVP Agent feature on a Cisco IOS router at each site, including the central site where Unified CM resides. At smaller sites, this router may coincide with the IP WAN router and PSTN gateway, while at larger sites they may be different platforms.
- In each Unified CM cluster, define a location for each site, and leave all bandwidth values as unlimited.
- Assign all devices located at each site to the appropriate location. This includes endpoints, gateways, conferencing resources, and the Cisco RSVP Agents themselves.
- Ensure that each Cisco RSVP Agent belongs to a media resource group (MRG) contained in the media resource group list (MRGL) of all devices at that site.
- In the Unified CM service parameters, set the **Default inter-location RSVP Policy** to **Mandatory** or **Mandatory (video desired)**, and set the **Mandatory RSVP mid-call error handle option** to **Call fails following retry counter exceeded** on both clusters.

- Enable RSVP on every WAN router interface in the network where congestion might occur, and configure the RSVP bandwidth based on the provisioning of the priority queue. (See RSVP Design Best Practices, page 11-57.)
- If you need to provision bandwidth separately for voice and video calls, also configure an RSVP application ID on the same WAN router interfaces.
- If the Cisco RSVP Agent is not co-resident with the IP WAN router, enable RSVP on the LAN interfaces connecting the agent to the WAN router.
- Ensure that RSVP SIP Preconditions are enabled on the SIP intercluster trunks (see Migration from Enhanced Locations Call Admission Control to RSVP SIP Preconditions, page 11-78, for the steps).
- Ensure that an inter-location RSVP policy of **Mandatory** or **Mandatory (video desired)** is set between the SIP intercluster trunk location and all locations, including itself (see intra-location RSVP policy below).
- Ensure that an intra-location RSVP policy of **Mandatory** or **Mandatory (Video Desired)** is set on the SIP intercluster trunk. An intra-location RSVP policy is set by selecting the location and configuring a policy for itself. This effectively ensures that calls within this location will engage RSVP call admission control. This is important for calls hairpinned on the trunk from call transfers or forwards back to the originating cluster.

Distributed Mixed Call Processing Deployments

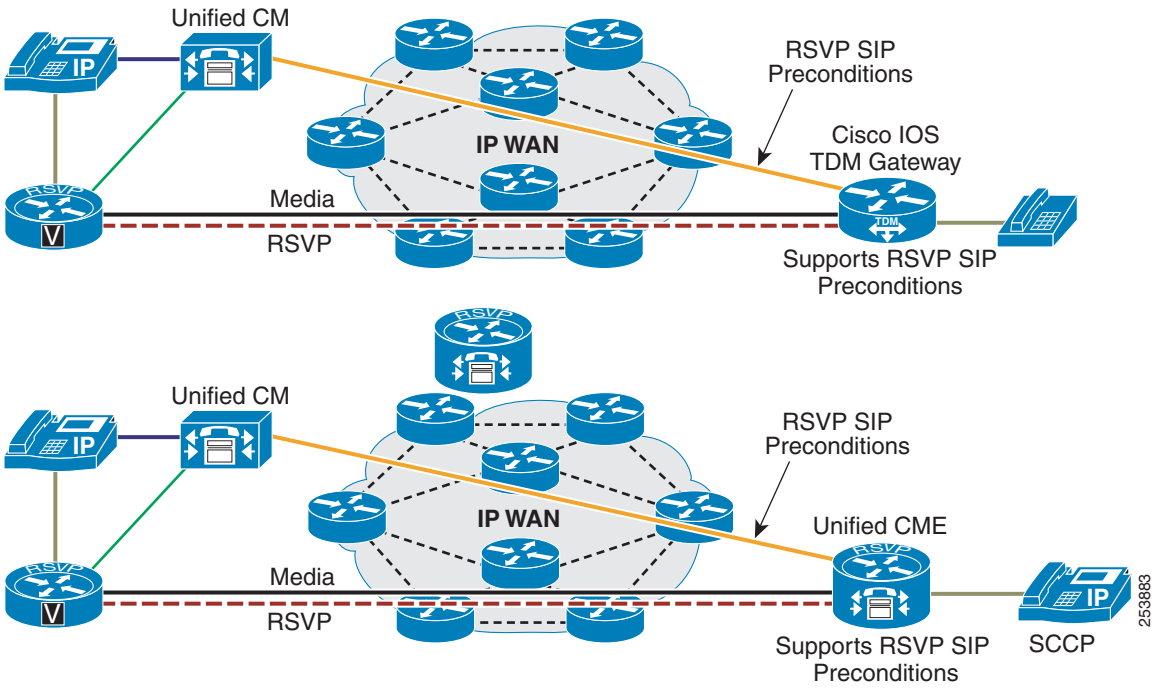
RSVP SIP Preconditions provides call admission control for distributed deployments of Unified Communications call control applications in a generic network topology.

This section contains a list of the supported RSVP SIP Preconditions deployment models. Hybrids and variations of these models are expected, and these are only high-level examples of the design possibilities, with best practices and design considerations called out. (Consult the specific product documentation for more information on configuration of these features.)

The following guidelines apply to the deployments in Figure 11-66:

- The deployment supports Unified CME SCCP integration for audio-only calls.
- Enable RSVP on every WAN router interface in the network where congestion might occur, and configure the RSVP bandwidth based on the provisioning of the priority queue. (See RSVP Design Best Practices, page 11-57.)
- If you need to provision bandwidth separately for voice and video calls, also configure an RSVP application ID on the same WAN router interfaces.
- Follow the recommendations listed in the Design Considerations for Unified CM Interoperability with SIP Cisco IOS TDM Gateway and Unified CME, page 11-88.

Figure 11-66 Unified CM to Cisco IOS Gateway (TDM) and to Unified CME (SCCP)



The following guidelines apply to the deployments in Figure 11-67:

- Follow the guidelines, best practices, limitations, and restrictions for SIP Cisco IOS TDM gateway and Unified CME interoperability listed in the *Cisco IOS SIP Configuration Guide*, available at

  http://www.cisco.com/en/US/docs/ios/voice/sip/configuration/guide/12_4t/sip_12_4t_book.html

- Ensure that the RSVP policy configured on each Unified CME or SIP Cisco IOS TDM gateway is consistent in order to avoid failed or unprotected calls. Use the following options under the **dial-peer** configuration when enabling RSVP reservations for the SIP Cisco IOS TDM gateway or Unified CME:

  ```
  req-qos guaranteed-delay audio
  acc-qos guaranteed-delay audio
  ```

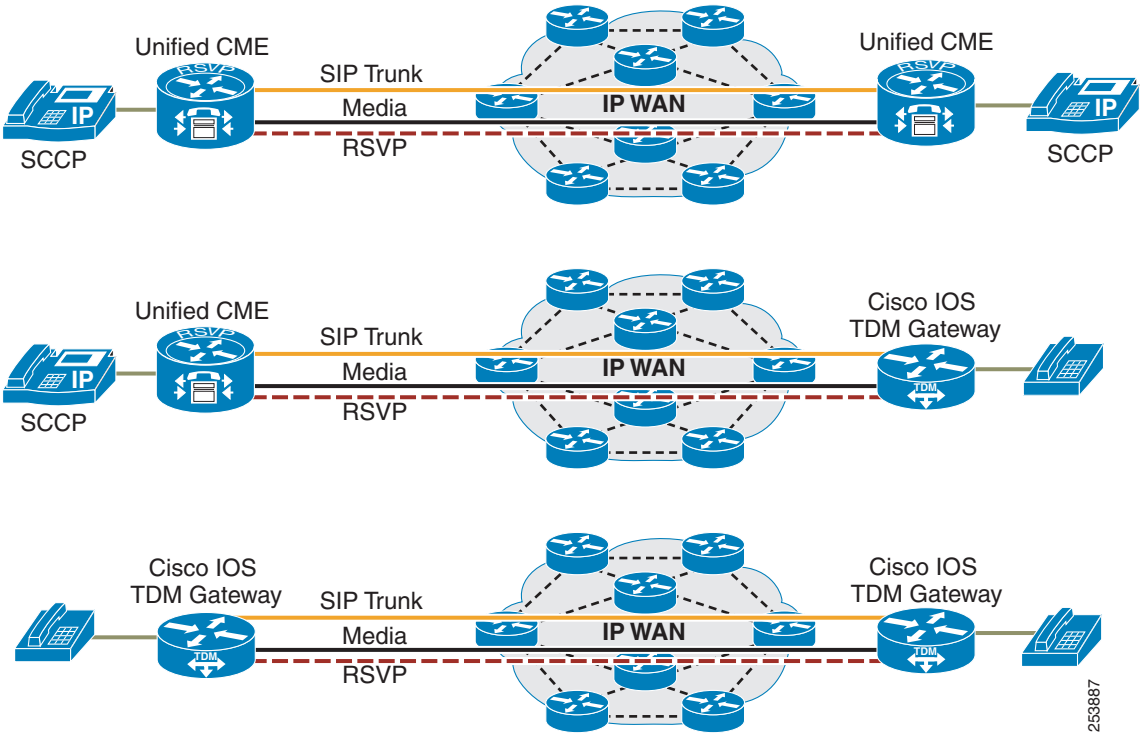
  This configuration ensures that for each voice call, the SIP Cisco IOS TDM gateway will request an RSVP reservation using the guaranteed delay service. The fact that both the requested QoS and the acceptable QoS specify this RSVP service means that the RSVP reservation is mandatory for the call to succeed. (That is, if the reservation cannot be established, the call will fail.)

- If Application ID is used, ensure that it is consistent across all of products in the solution (SIP Cisco IOS TDM gateway and Unified CME).

- Ensure that inbound and outbound dial peers are correctly matched to ensure that the appropriate dial peers configured with SIP Preconditions are utilized. For further information, refer to the *Cisco IOS SIP Configuration Guide*, available at

  http://www.cisco.com/en/US/docs/ios/voice/sip/configuration/guide/12_4t/sip_12_4t_book.html

Figure 11-67 Unified CME to Unified CME, Unified CME to Cisco IOS Gateway, and Cisco IOS Gateway to Cisco IOS Gateway



The following guidelines apply to the deployments in Figure 11-68:

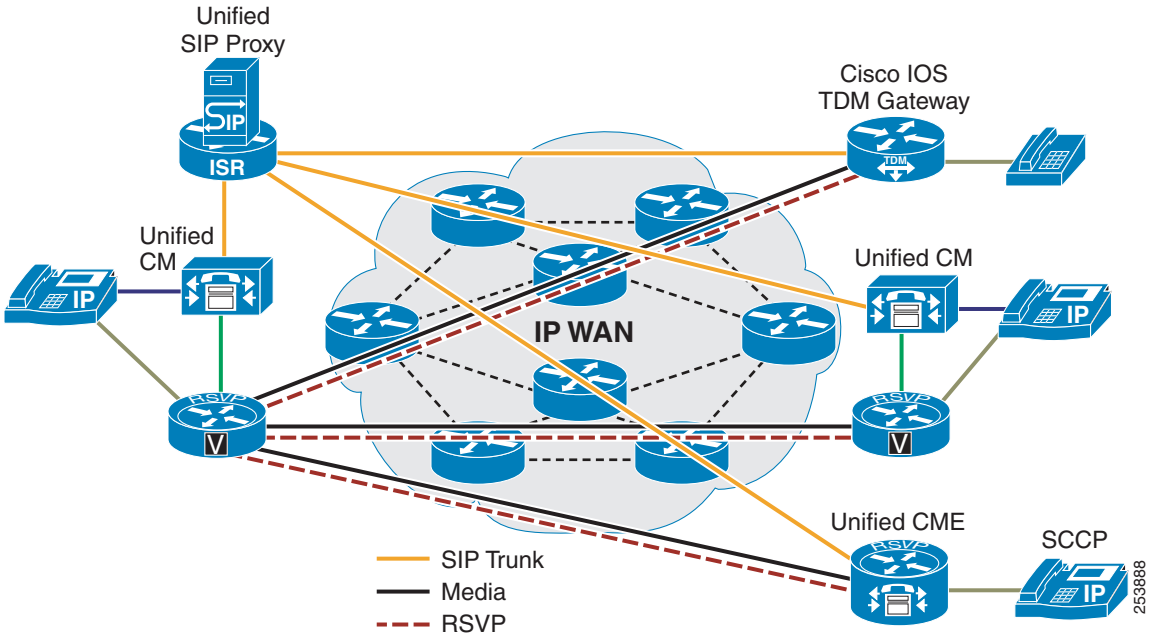
- Follow the guidelines stipulated for Figure 11-66 for Unified CM in distributed call processing deployments, and follow the guidelines stipulated for Figure 11-67 for Unified CME and SIP Cisco IOS TDM Gateways.
- Each Unified CM cluster will typically have a single trunk directed to the Cisco Unified SIP Proxy. Ensure that RSVP SIP Preconditions (end-to-end RSVP) is enabled on that trunk.
- Ensure that RSVP SIP Preconditions are enabled on the SIP trunk to the Cisco Unified SIP Proxy (see Migration from Enhanced Locations Call Admission Control to RSVP SIP Preconditions, page 11-78, for the steps).
- Ensure that an inter-location RSVP policy is configured on each Unified CM cluster between the IP phone locations and the SIP trunk location. This ensures that the SIP preconditions will be enabled for all calls engaged on the SIP trunk to the Cisco Unified SIP Proxy.
- If there are potential SIP destinations that do not support RSVP SIP Preconditions, then ensure that RSVP SIP Preconditions fallback to local RSVP is configured on the SIP trunk to allocate an RSVP Agent for those call flows. And if RSVP SIP Preconditions fallback is enabled, ensure that the RSVP Agent associated to the SIP trunk is in a physical site that will ensure RSVP path protection for those call flows.
- Unified CME and SIP Cisco IOS TDM gateways will also typically have a single SIP dial peer directed to the Cisco Unified SIP Proxy. Ensure that RSVP SIP Preconditions (SIP Preconditions support) is configured on those dial peers, as stipulated in the *Cisco IOS SIP Configuration Guide*, available at

  http://www.cisco.com/en/US/docs/ios/voice/sip/configuration/guide/12_4t/sip_12_4t_book.html

- For information on configuration, guidelines, best practices, limitations, and restrictions for Cisco Unified SIP Proxy, refer to the documentation at

  http://www.cisco.com/en/US/prod/collateral/modules/ps2797/data_sheet_c78-521390_ps2797_Products_Data_Sheet.html

Figure 11-68 All Components Capable of RSVP SIP Preconditions via Cisco Unified SIP Proxy



The following guidelines apply to the deployments in Figure 11-69:

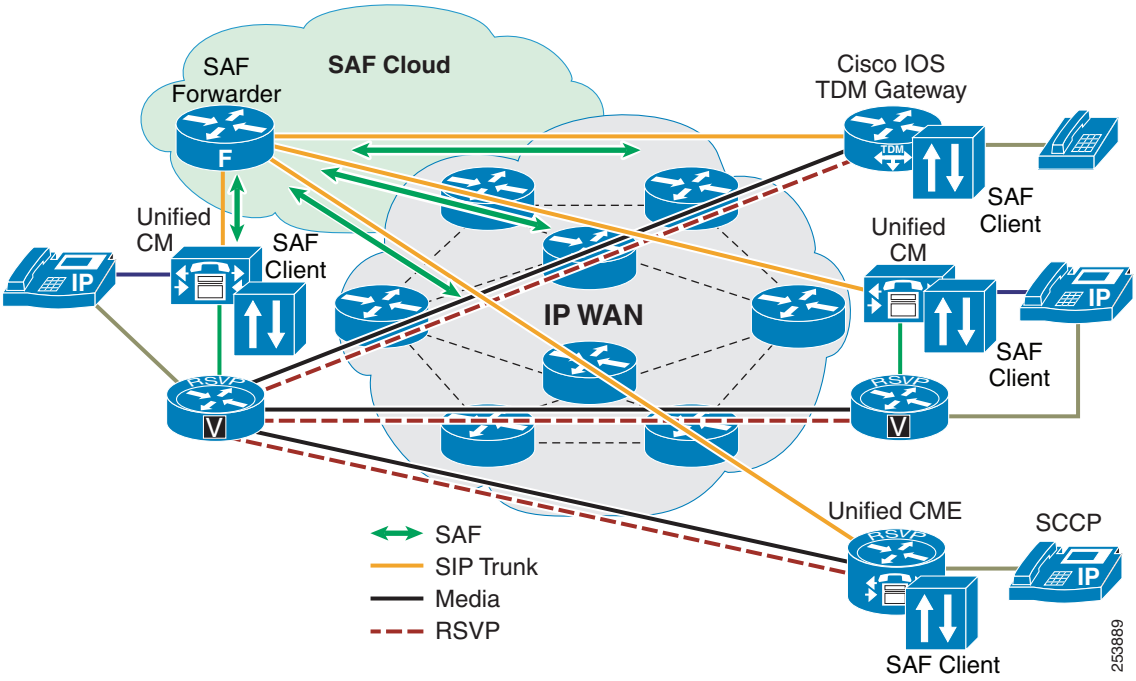
- Follow the guidelines stipulated for Figure 11-66 for Unified CM in distributed call processing deployments, and follow the guidelines stipulated for Figure 11-67 for Unified CME and SIP Cisco IOS TDM Gateways.
- Ensure that Services Advertisement Framework (SAF) and Call Control Discovery (CCD) are enabled in the network and functioning across the deployed products: For details on SAF configuration, refer to the following documents:
  - *Cisco IOS Service Advertisement Framework Configuration Guide*

    http://www.cisco.com/en/US/docs/ios/saf/configuration/guide/15_0/saf_15_0_book.html
  - *Cisco Unified Communications Manager Features and Services Guide*

    http://www.cisco.com/en/US/products/sw/voicesw/ps556/prod_maintenance_guides_list.html
- Enable RSVP SIP Preconditions on the SAF-enabled SIP trunk (see Migration from Enhanced Locations Call Admission Control to RSVP SIP Preconditions, page 11-78, for the steps).
- Enable RSVP SIP Preconditions on the SAF-enabled SIP trunk. Ensure that only SAF-enabled SIP trunks are used with Call Control Discovery in an RSVP SIP Preconditions deployment. SAF-enabled H.323 trunks will not function in an RSVP SIP Preconditions deployment.
- Ensure that an inter-location RSVP policy of **Mandatory** or **Mandatory (video desired)** is set between the SAF-enabled SIP trunk location and all locations, including itself (see the intra-location RSVP policy below). This ensures that the SIP preconditions will be enabled for all calls engaged on the SIP trunk to and from the SAF network.

- Ensure that an intra-location RSVP policy of **Mandatory** or **Mandatory (Video Desired)** is set on the SAF-enabled SIP trunk. An intra-location RSVP policy is set by selecting the location and configuring a policy for itself. This effectively ensures that calls within this location will engage RSVP call admission control. This is important for calls hairpinned on the trunk from call transfers or forwards back to the originating cluster.
- When using SAF-enabled SIP trunks in RSVP SIP Preconditions environments, Cisco recommends ensuring that CCD Automatic Failover is enabled and that calls that fail call admission control are routed to a local route group. For more information see the section on Call Control Discovery Automatic PSTN Failover as well as the DP chapter Local Route Group.

Figure 11-69 All Components Capable of RSVP SIP Preconditions via Service Advertisement Framework (SAF) and Call Control Discovery (CCD)



Call Admission Control Design Recommendations for TelePresence Video Interoperability Architectures

Video interoperability refers to the support for point-to-point video calls between Cisco TelePresence endpoints, Cisco Unified Communications video endpoints, and third-party video endpoints without requiring a Multipoint Control Unit (MCU). This section discusses the features in Enhanced Locations CAC and the design considerations and recommendations applicable to Quality of Service (QoS) for interoperable video calls.

This section explains how to supplement the design to allow for video interoperation between Unified Communications and TelePresence endpoints. For information on Unified Communications endpoints, refer to the chapter on Unified Communications Endpoints, page 18-1; and for information on TelePresence endpoints, refer to the Cisco TelePresence Endpoints product documentation at http://www.cisco.com/en/US/products/ps7060/index.html.

Note Third-party video endpoints should follow the same guidelines and recommendations as Cisco Unified Communications endpoints. Throughout this section, the term *UC endpoints* is used to refer to both third-party endpoints and Cisco Unified Communications Endpoints.

Starting with Cisco Unified CM 9.0, the Cisco TelePresence solution provides the ability to reserve network bandwidth and perform admission control for TelePresence calls. It is important to be familiar with Enhanced Locations call admission control (CAC) and RSVP CAC prior to designing TelePresence video interoperability. This section addresses both Enhanced Locations CAC and RSVP with regard to TelePresence video interoperability, and each CAC mechanism has its own benefits, design considerations, and requirements.

Additionally, TelePresence video interoperability in Unified CM enables Cisco Telepresence System (CTS) endpoints to communicate with non-CTS endpoints, provided that the installed CTS software supports such interoperability. For further information, refer to the document on *Interoperability Between CTS Endpoints and Other Cisco Endpoints or Devices*, available at

http://www.cisco.com/en/US/docs/telepresence/interop/endpoint_interop.html

Supported CAC Deployment Scenarios and Design Considerations

The design considerations for TelePresence video interoperability CAC are based on the following main deployment scenarios:

- Mixed Single Cluster — Mixed UC video endpoints and TelePresence endpoints registered to a single cluster

  This is a single-cluster design where TelePresence and UC video endpoints are registered to the same cluster. The call processing deployment model can be any of the single-cluster designs such as clustering over the WAN, multi-site centralized call processing, or single campus designs.

- Dedicated Multi-Cluster — UC video endpoints and TelePresence endpoints on separate dedicated clusters

  This is a multi-cluster design where the TelePresence endpoints are registered to a different cluster than the UC video endpoints. The call processing deployment model can be multi-site centralized or multi-site distributed cluster designs. All releases of Cisco Unified CM 8.*x* support this model for

deploying UC and TelePresence in the same enterprise. However, only Unified CM 8.6 and later releases support the new capability for UC video and TelePresence call interoperability in point-to-point video calls without the use of a video MCU.

- Mixed Multi-Cluster — UC video endpoints and TelePresence endpoints mixed in multi-cluster distributed deployments

  This is a multi-cluster design where TelePresence endpoints are spread across multiple clusters serviced by a single Cisco TelePresence System Manager (CTS-Manager). There are also deployment scenarios where some TelePresence endpoints can be co-located with the same Unified CM cluster as the UC endpoints (mixed single cluster model) while other TelePresence endpoints are registered to a dedicated cluster, and both clusters are serviced by a single CTS-Manager for TelePresence. The call processing deployment model can be multi-site, multi-cluster centralized, or multi-site/multi-cluster distributed designs. This is a hybrid model that combines aspects of the dedicated multi-cluster model with mixed single-cluster model where all endpoints are registered to the same cluster.

Cisco Unified Communications Manager Session Management Edition (SME) is also supported in the multi-cluster models. However, because Session Management Edition is a variation of the multi-site distributed call processing deployment model and is typically employed to interconnect large numbers of Unified Communications systems through a single front-end system, there are no specific guidelines for it with regard to TelePresence video interoperability.

Enhanced Locations CAC Design Considerations and Recommendations

When designing Enhanced Locations (E-L) CAC for TelePresence Video Interoperability, follow the design recommendations and considerations listed in this section.

Design Recommendations

The following design recommendations apply to TelePresence video interoperability solutions that employ Enhanced Locations (E-L) CAC:

- E-L CAC for TelePresence video interoperability is supported in all three deployment models: mixed single cluster, dedicated multi-cluster, and mixed multi-cluster.
- When deploying Unified Communications video and TelePresence video interoperability, 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 E-L CAC TelePresence endpoints can be managed in the same location as Unified Communications video endpoints. If TelePresence calls are not to be tracked through E-L 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 E-L CAC, then set immersive location and links bandwidth pool to a value according to the bit rate and number of calls to be allowed over the locations and link paths.
- E-L CAC performs call admission control end-to-end on location pairs; therefore, cross-cluster call transfers and forwards do not require QSIG tunneling with path replacement in order to perform end-to-end E-L CAC. However, Cisco recommends using QSIG path replacement when possible to optimize the call signaling path because this diminishes the number of call signaling legs in complex call forwarding or transfer scenarios.
- Intercluster SIP trunks should be associated with the shadow location.

- Only point-to-point video calls are supported between UC and TelePresence endpoints. No ad-hoc conferencing is supported unless a video MCU is available.
- 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.
- 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.

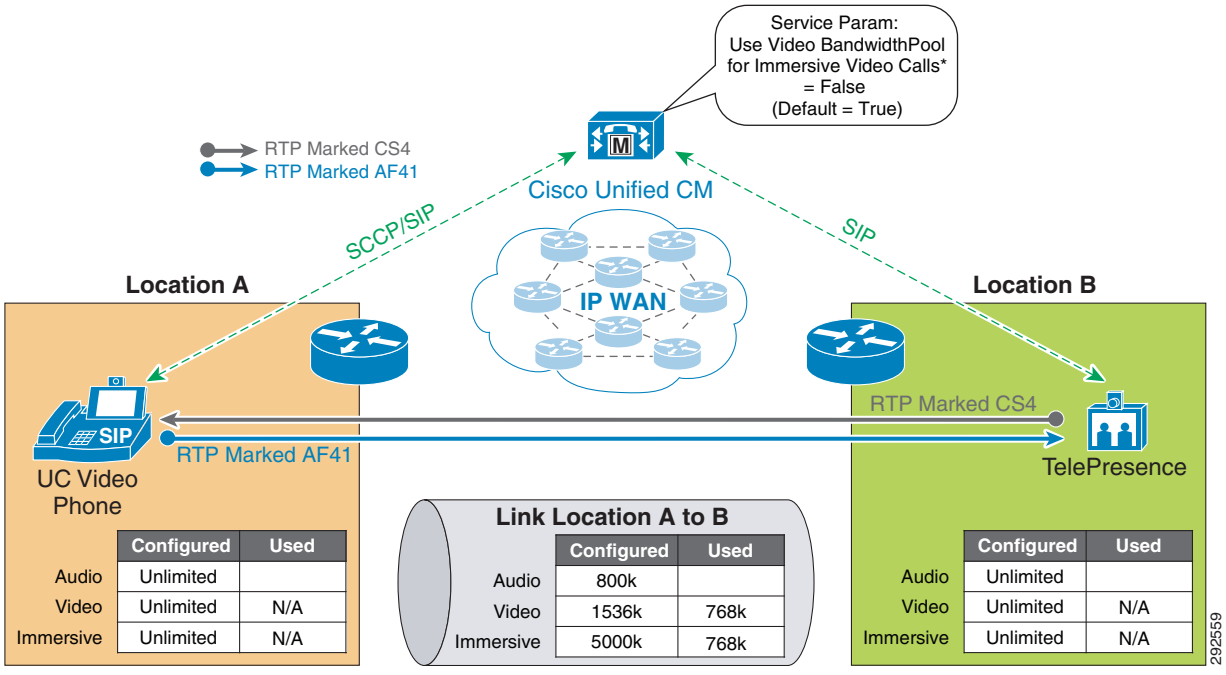
Design Considerations

When deploying Enhanced Locations CAC for TelePresence video interoperable calls, consider the affects of DSCP marking for both QoS classes.

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 11-70 illustrates the media marking and bandwidth accounting.

Figure 11-70 Bandwidth Deductions and Media Marking in a Multi-Site Deployment with Enhanced Locations CAC



Bandwidth Accounting for TelePresence Video Interoperability Calls

Enhanced Locations 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 11-70. 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 Locations 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.

For more information on the call flows for Enhanced Locations CAC and TelePresence interoperable calls, see the section on Enhanced Locations CAC for TelePresence Immersive Video, page 11-32.

RSVP CAC Design Considerations and Recommendations

In an RSVP solution for TelePresence video interoperability, the goal is to ensure RSVP CAC for both end-to-end UC endpoint calls and TelePresence video interoperability calls, while also ensuring that end-to-end TelePresence calls never invoke RSVP (RSVP Agent) because it is not currently supported and can cause specific TelePresence features to fail.

TelePresence video interoperability in RSVP deployments is very easy to achieve and provides better CAC support than Enhanced Locations CAC solutions, provided that a few design rules and deployment models are adhered to. The supported designs are the mixed single cluster and dedicated multi-cluster designs. The mixed multi-cluster designs are not recommended due to the configuration complexity of ensuring that end-to-end TelePresence calls across clusters do not invoke RSVP and that all other interoperable point-to-point call scenarios do invoke RSVP.

For details about RSVP architecture and general design and deployment considerations, see the section on Unified Communications Architectures Using Resource Reservation Protocol (RSVP), page 11-41. It is important to understand the RSVP principles and solution prior to reading this section independently as this section will uniquely cover recommendations and considerations specific to Telepresence video interoperability with regards to RSVP.

Design Recommendations

The following design recommendations apply to TelePresence video interoperability solutions that employ RSVP for call admission control:

- Locations-based RSVP is supported in mixed single cluster and dedicated multi-cluster designs. As mentioned previously, a mixed multi-cluster design can be used but at the cost of complex configuration of trunks, dial plan, and locations-based RSVP policy. Therefore, mixed multi-cluster designs are not recommended and will not be treated in these design recommendations.
- Both Local RSVP and RSVP SIP Preconditions are supported for intercluster calls, provided they are designed and configured according to the guidelines in the relevant sections of this chapter.

- When designing solutions for TelePresence video interoperability with RSVP, ensure that TelePresence-to-TelePresence calls never invoke RSVP because that functionality is not currently supported and can cause specific TelePresence features to fail. However, for TelePresence calls to/from UC video endpoints, RSVP should be invoked. To achieve this, ensure the following:
  - TelePresence endpoints are in CAC locations separate from UC endpoints.
  - TelePresence locations set their RSVP policy to **no reservation** for the location pairing with other TelePresence locations as well as their own location. For more information on RSVP Policy for location pairs, see the section on RSVP Policy, page 11-69.
  - TelePresence locations set their RSVP policy to **mandatory** or **mandatory video desired** for the location pairings with UC video endpoint locations.
  - For dedicated multi-cluster deployments, intercluster trunks require an RSVP policy pairing between UC endpoint locations as well as TelePresence endpoint locations.

    For dedicated TelePresence clusters, intercluster trunks pointing to dedicated UC endpoint clusters should have an RSVP policy location pairing with TelePresence endpoints set to **mandatory** or **mandatory video desired**. Intercluster trunks pointing to TelePresence clusters should have an RSVP policy location pairing with TelePresence endpoints set to **no reservation**.

    For dedicated UC endpoint clusters, intercluster trunks pointing to both dedicated UC clusters and TelePresence clusters should have an RSVP policy location pairing with UC endpoints set to **mandatory** or **mandatory video desired**.

Design Considerations

When deploying RSVP CAC for TelePresence video interoperable solutions, consider the following major factors:

RSVP Performs CAC for Both UC and TelePresence Endpoints

Unlike Enhanced Locations CAC, RSVP performs CAC for both the TelePresence and UC endpoints for TelePresence video interoperating calls. It also overwrites the endpoint QoS marking, thus ensuring the correct QoS marking from agent to agent. Figure 11-71 illustrates this point

Figure 11-71 Bandwidth Deductions and Media Marking in a Multi-Site Deployment with RSVP CAC

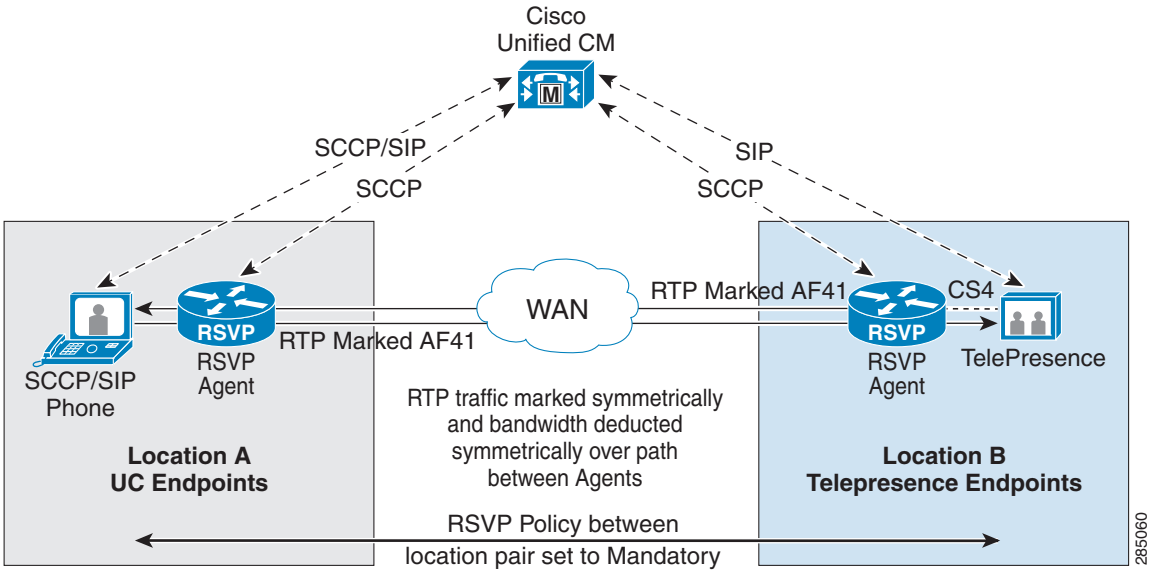


Figure 11-71 illustrates a call between a TelePresence endpoint and a Cisco Unified IP Phone 9900 Series video phone, where RSVP Agents are invoked for RSVP CAC. Two salient points are illustrated here:

- RSVP Agents re-mark the RTP media traffic so that the media is marked symmetrically between RSVP Agents with a DSCP value of AF41, which is consistent with the **DSCP for Video calls** setting. This provides symmetrically marked RTP traffic between endpoints, which is something that the locations CAC solution cannot achieve.
- RSVP deducts bandwidth over the media path between the RSVP Agents for the UC endpoint and the TelePresence endpoint. This provides CAC for audio and video streams in both directions for TelePresence video interoperability calls. This is also something that locations CAC inherently cannot achieve.

RSVP Should Not Be Used on Calls Between Endpoints Located in the Same Physical Site in a Mixed Single-Cluster Deployment

When deploying mixed single-cluster designs, where TelePresence endpoints and UC video endpoints are registered to the same Unified CM cluster and located in the same physical site or campus, it is important not to engage RSVP for calls between these devices. The RSVP policy location pair between the UC endpoint's location and the TelePresence endpoint's location should be set in these cases to **no reservation**. This is illustrated in Figure 11-72.

Figure 11-72 RSVP Policy Setting for Calls Between UC and TelePresence Endpoints in the Same Site

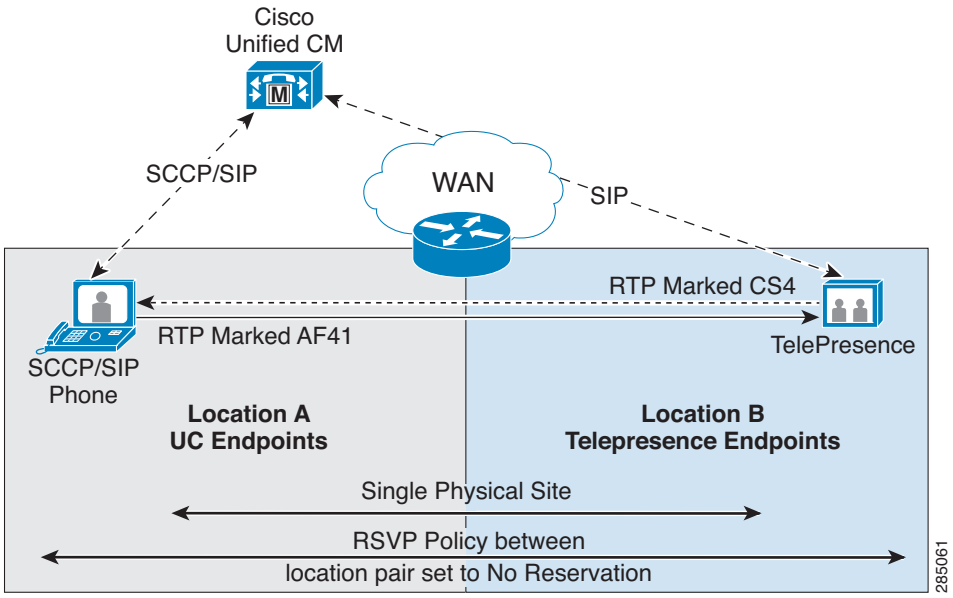
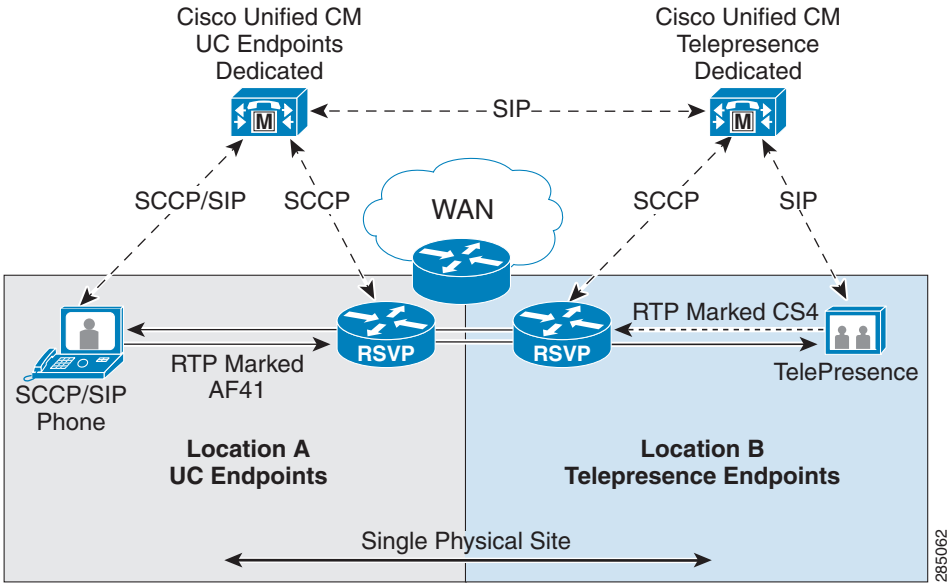


Figure 11-72 illustrates a TelePresence endpoint and a Cisco Unified IP Phone 9900 Series video phone in the same physical site but in separate CAC locations. The location pair policy is set to **no reservation** so that calls between TelePresence and UC endpoints in the same physical site or campus do not invoke RSVP. Note that the RTP streaming is asymmetrical. This will be the case but is usually inconsequential over the LAN.

RSVP Should Be Used on Calls Between Endpoints in Different Clusters

For dedicated multi-cluster deployments it is necessary to invoke RSVP even when the UC endpoint and the TelePresence endpoint are located in the same physical site. This will utilize RSVP Agent resources but the media will not be routed over the WAN uplink. Figure 11-73 illustrates this point.

Figure 11-73 Bandwidth Deductions and Media Marking in a Multi-Cluster, Single-Site Deployment with RSVP CAC



Note that the RSVP Agents as well as the WAN edge router in Figure 11-73 could all be either co-located or separated as depicted. For information on co-locating multiple RSVP Agents registered to separate clusters, see the section on Multiple Clusters Sharing a Single Platform for RSVP Agent, page 11-124.

Guidelines for Session Management Edition Deployments

For Session Management Edition deployments, follow the same guidelines that are stipulated in the section on Unified CM Session Management Edition with RSVP Deployments, page 11-120. Keep in mind, however, that RSVP is not recommended in mixed multi-cluster deployments.

Call Admission Control Design Recommendations for Unified CM Session Management Edition Deployments

Cisco Unified Communications Manager Session Management Edition (SME) has two main forms of call admission control available to it for admitting trunk-to-trunk audio and video calls: one is Enhanced Locations call admission control (CAC) and the other is Resource Reservation Protocol (RSVP) using RSVP-enabled Locations. This section covers design guidelines and best practices specific to Session Management Edition deployments. It does not cover the basic functions of Enhanced Locations CAC, Locations-based RSVP, or RSVP SIP Preconditions. Cisco highly recommends that you become familiar with the applicable sections of this chapter as a prerequisite to understanding the following design guidelines.

Session Management Edition with Enhanced Locations 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 Locations (E-L) CAC. For more information on Unified CM SME, refer to the *Cisco Unified Communications Manager Session Management Edition Deployment Guide*, available at

http://www.cisco.com/en/US/products/ps10661/products_implementation_design_guides_list.html

Recommendations and Design Considerations

- All leaf clusters that support E-L CAC should be enabled for intercluster E-L CAC with SME.
- SME can be used as a centralized bootstrap hub for the E-L CAC intercluster hub replication network. See LBM Hub Replication Network, page 11-23, for more information.
- All trunks to leaf clusters supporting E-L CAC should be SIP trunks placed in the shadow location to enable E-L CAC on the trunk between SME and the leaf clusters supporting E-L CAC.
- For TelePresence video interoperability, see the section on Call Admission Control Design Recommendations for TelePresence Video Interoperability Architectures, page 11-109.
- Connectivity from SME to any trunk or device other than a Unified CM that supports E-L CAC (some examples are third-party PBXs, gateways, Unified CM clusters prior to release 9.0 that do not support E-L 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 prior to 9.0, but they are not supported with Unified CM 9.*x* Enhanced Locations CAC. Shadow locations are special locations that enable trunks between Unified CM clusters that support E-L CAC to accomplish it end-to-end.
- SME can be used as a locations and link management cluster. See Figure 11-74 as an example of this.
- SME can support a maximum of 2,000 locations configured locally.

Figure 11-74 Unified CM SME as a Location and Link Management Cluster

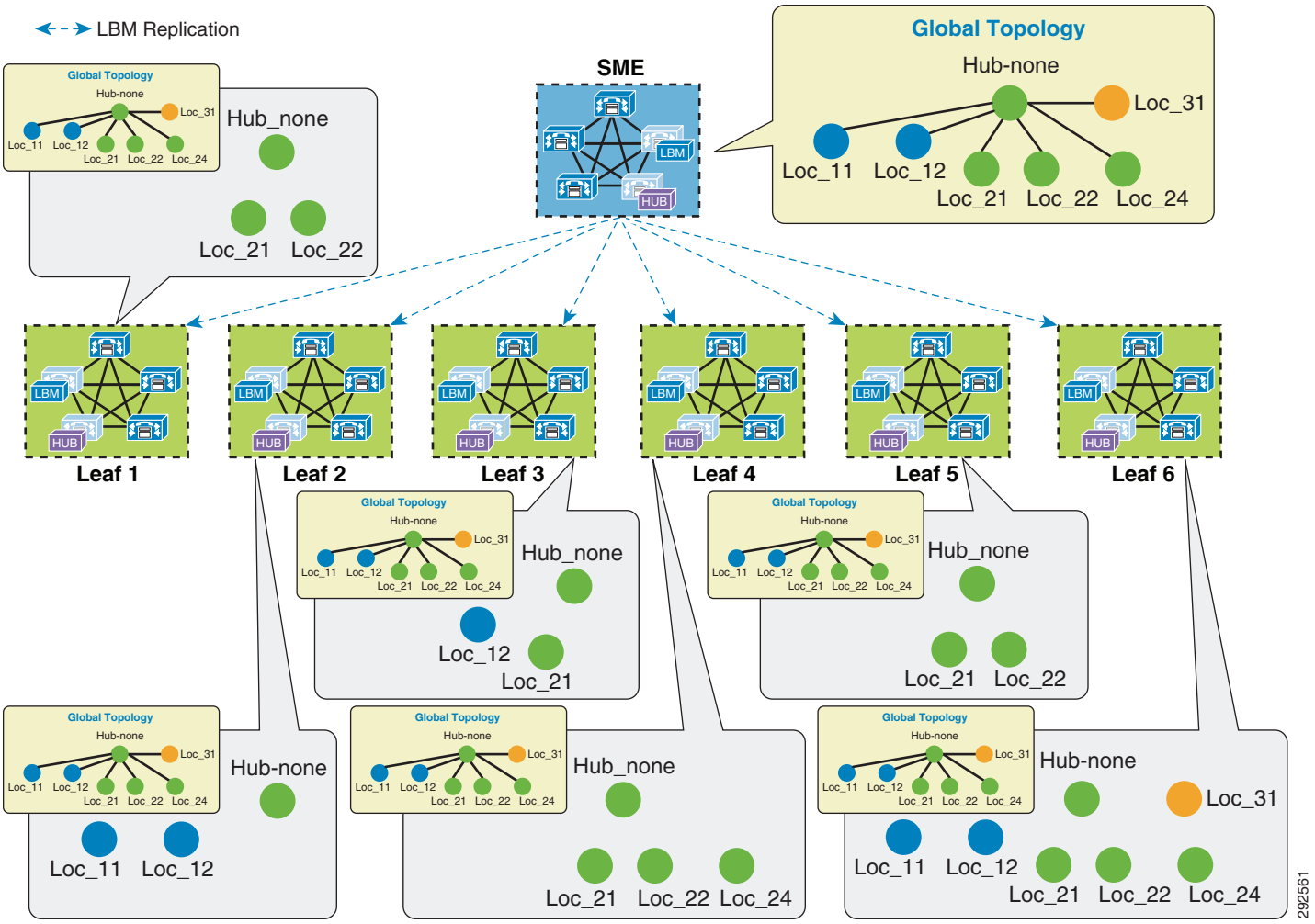
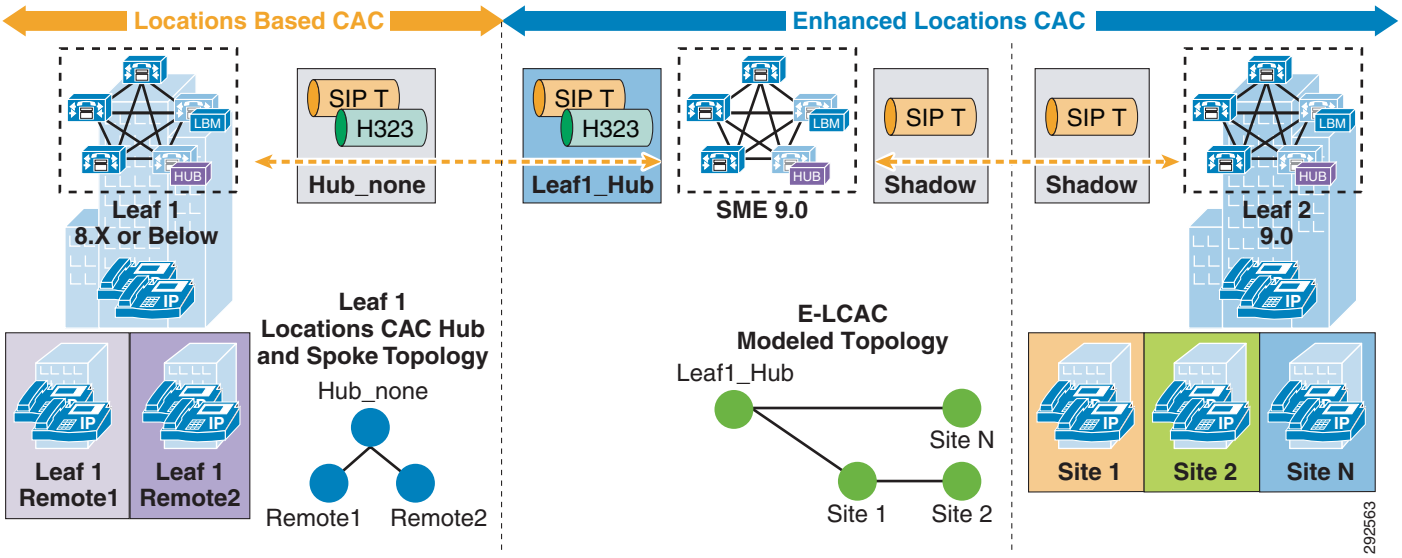


Figure 11-74 illustrates SME as a location and link management cluster. The entire location and link global topology is configured and managed in SME, and the leaf clusters configure locally only the locations that they require to associate with the end devices. When intercluster E-L CAC is enabled and 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 11-28.

Figure 11-75 illustrates an SME design where intercluster E-L 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 prior to 9.0 and are running traditional locations CAC (left). In this type of a deployment the locations managed by traditional locations CAC cannot be common or shared locations between E-L CAC-enabled clusters. 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 E-L 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 E-L CAC locations and links while Leaf 1 manages its remote locations with traditional locations CAC.

Figure 11-75 SME Design with Enhanced Locations CAC and Traditional Locations CAC in Leaf Clusters



QSIG Path Replacement Over Intercluster Trunks

In addition to providing features such as Call Back (on Busy/No Reply) between phones in Unified CM leaf clusters, QSIG also provides path replacement, which optimizes call signaling between clusters when, for example, a call is transferred or forwarded from one cluster to another. With E-L CAC, end-to-end QSIG path replacement is not required because E-L CAC deducts bandwidth over the correct locations and links path end-to-end between clusters. Nonetheless, QSIG path replacement is beneficial in optimizing the signaling path between clusters that do not support E-L CAC, as well as for third-party PBXs and gateways where supported, in order to avoid hair-pinning the trunk signaling with every forward or transfer off the cluster, PBX, or gateway. For more information on QSIG over intercluster trunks, refer to the *Cisco Unified Communications Manager Session Management Edition Deployment Guide*, available at

http://www.cisco.com/en/US/products/ps10661/products_implementation_design_guides_list.html

Unified CM Session Management Edition with RSVP Deployments

Unified CM Session Management Edition can be deployed in a number of RSVP deployments either with SIP Preconditions support between clusters or without it. For an overview of RSVP in Unified Communications designs and as a prerequisite to understanding the following content, see the section on Unified Communications Architectures Using Resource Reservation Protocol (RSVP), page 11-41, which includes detailed sections on Unified CM RSVP-enabled Locations and RSVP SIP Preconditions.

Session Management Edition Design with Leaf Clusters without RSVP SIP Precondition Support

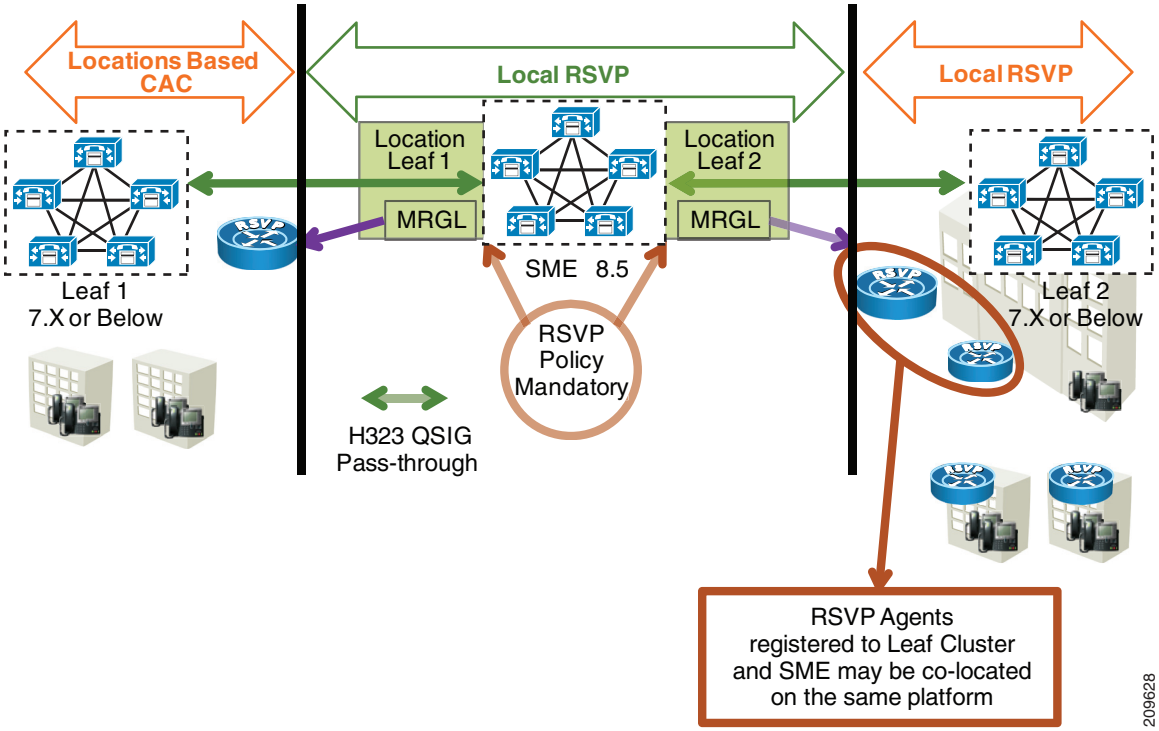
Unified CM Session Management Edition can be deployed with leaf clusters that maintain their own CAC support locally within the leaf cluster. In such deployments where the leaf clusters manage their own CAC mechanism for their own devices in remote sites, or where those leaf clusters are in a campus deployment and do not require CAC for intra-cluster calls (calls that remain within the leaf cluster), Session Management Edition can be leveraged to manage CAC for audio and video calls between those leaf clusters (intercluster calls) with RSVP. In these cases Session Management Edition is leveraging the Cisco RSVP Agents to handle CAC across the links between the leaf clusters but is not managing CAC for the links that the leaf clusters manage for their intra-cluster calls.

Requirements

- Intra-cluster calls are managed by the leaf cluster CAC.
- Intercluster calls are managed by Session Management Edition RSVP-enabled Locations CAC.
- Cisco highly recommends that the WAN bandwidth managed by Session Management Edition for intercluster calls should not be shared with the WAN bandwidth for intra-cluster calls managed by the leaf clusters. If the bandwidth from the same WAN links is managed by two separate CAC mechanisms, then there is the potential for double bandwidth counting because the two separate CAC mechanisms are not aware of each other.

Figure 11-76 depicts two leaf clusters deployed in different regional sites connecting to one another through Session Management Edition for intercluster connectivity. Leaf cluster 1 is using Locations CAC to manage any remote sites in its CAC domain, while Leaf cluster 2 is using RSVP to manage any remote sites under its CAC domain. Session Management Edition in this case leverages RSVP-enabled Locations to manage the CAC domain between leaf clusters using remote RSVP Agents. The remote RSVP Agents are simply standard RSVP Agents registered with Session Management Edition and associated to the leaf cluster trunk(s) through media resource group (MRG) and media resource group list (MRGL) functions, but they are physically located at a campus site and/or co-located with the leaf cluster and thus could be "remote" from the Session Management Edition cluster. These RSVP Agents should be at a head-end network WAN that interconnects the leaf clusters and is enabled to support RSVP.

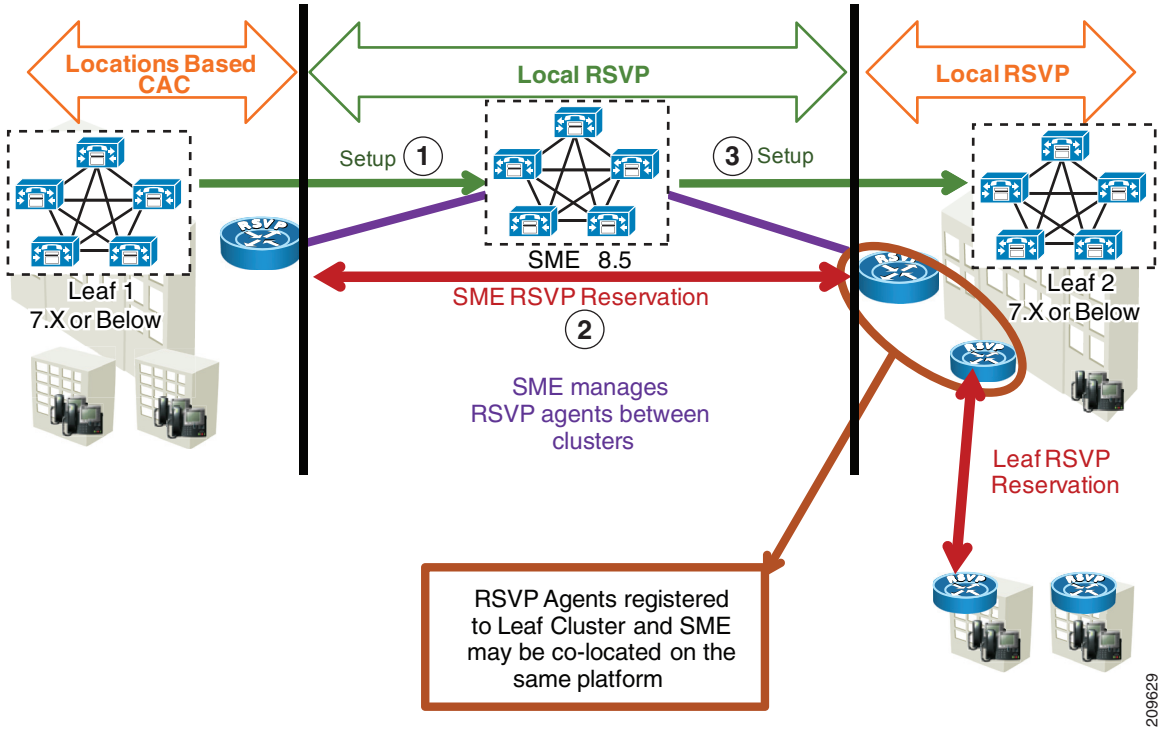
Figure 11-76 Session Management Edition Deployment with Leaf Clusters without SIP Precondition Support



In a call setup between Leaf 1 and Leaf 2, the call flow works as follows (see Figure 11-77):

1. Leaf 1 sets up a call to Session Management Edition. This is done over an H.323 or SIP intercluster trunk. QSIG Path Replacement support is preferred for H.323 prior to Unified CM 8.5 and SIP with Unified CM 8.5 and later releases.
2. Session Management Edition receives the call from Leaf 1 and allocates two RSVP Agents. The RSVP Agents are associated to the trunks through the media resource group and list functions (see Figure 11-76). Once allocated, the RSVP Agents reserve the bandwidth over the path between them.
3. If the bandwidth request is successful, the call is extended from Session Management Edition to Leaf 2 over the SIP or H.323 intercluster trunk. If the reservation fails, then the call is not extended to Leaf 2 and, depending on the configuration of call processing in Session Management Edition, the call could be extended elsewhere or torn down from Leaf 1.

Figure 11-77 Call Flow for Session Management Edition Deployment with Leaf Clusters without SIP Precondition Support



The following notes apply to Figure 11-77:

- If the call is extended to Leaf 2 in step 3, then Leaf 2 can use Locations CAC or RSVP as the admission control for that call leg. If using RSVP as in Figure 11-77, then Leaf 2 will associate an RSVP Agent to the SIP trunk pointing to Session Management Edition and will associate another RSVP Agent with the called party endpoint.
- In cases where the leaf cluster is doing RSVP locally and not with SIP Preconditions, the Session Management Edition Remote RSVP Agent and the leaf cluster RSVP Agent associated with the trunk to Session Management Edition can be co-located on the same routing platform. See Multiple Clusters Sharing a Single Platform for RSVP Agent, page 11-124, for further information.

Session Management Edition Design with Leaf Clusters with SIP Precondition Support

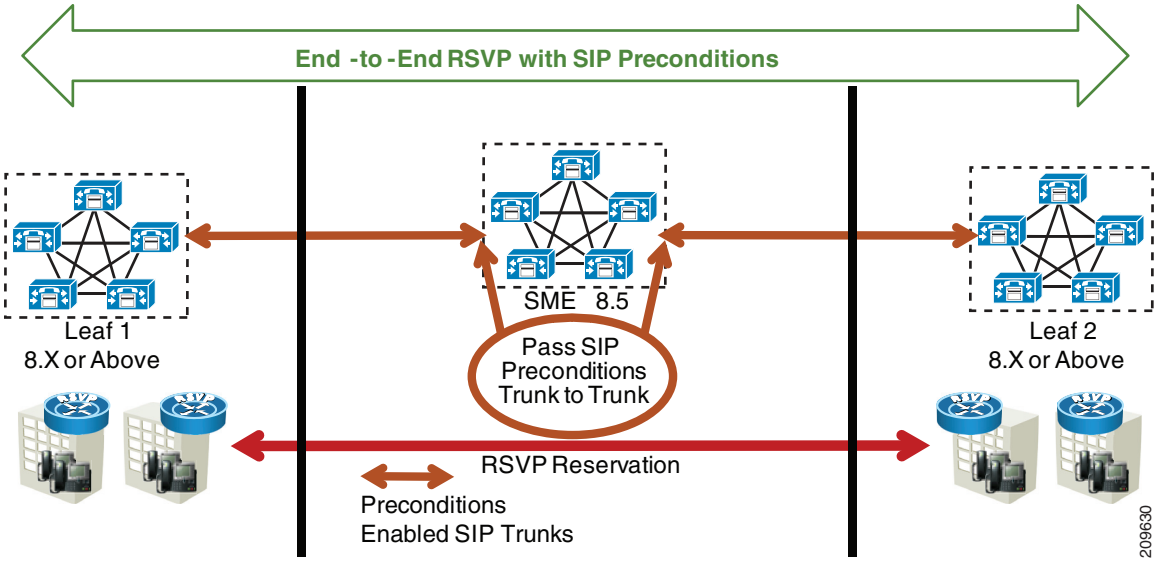
Unified CM Session Management Edition can also be deployed with leaf clusters that manage their CAC domain with RSVP and support RSVP SIP Preconditions. Session Management Edition in this case enables its intercluster trunks with RSVP SIP Preconditions and passes the SIP Preconditions from leaf cluster to leaf cluster, thus providing an end-to-end RSVP Agent deployment.

Requirements

- Leaf clusters with Unified CM 8.0 and later releases (preferably Unified CM 8.6.1 or later)
- Intra-cluster calls managed by the leaf cluster using Unified CM RSVP-enabled Locations
- Intercluster calls managed by Session Management Edition with SIP intercluster trunks (ICTs) configured to pass RSVP SIP Preconditions from leaf cluster to leaf cluster
- QSIG with path-replacement is not required, but Cisco recommends QSIG with path-replacement enabled on all SIP intercluster trunks (ICTs) to optimize call signaling for transferred and forwarded calls (requires Unified CM 8.5 or later release).

Figure 11-78 depicts the described solution where leaf clusters have Unified CM RSVP-enabled Locations for intra-cluster calls, and Session Management Edition as well as the SIP ICTs configured on both of the leaf clusters are enabled with RSVP SIP Preconditions. Session Management Edition in turn passes the SIP Preconditions along from leaf Unified CM to leaf Unified CM, which in turn pass those preconditions down to the branch RSVP Agents, thus providing an end-to-end RSVP reservation directly from branch to branch across multiple cluster boundaries.

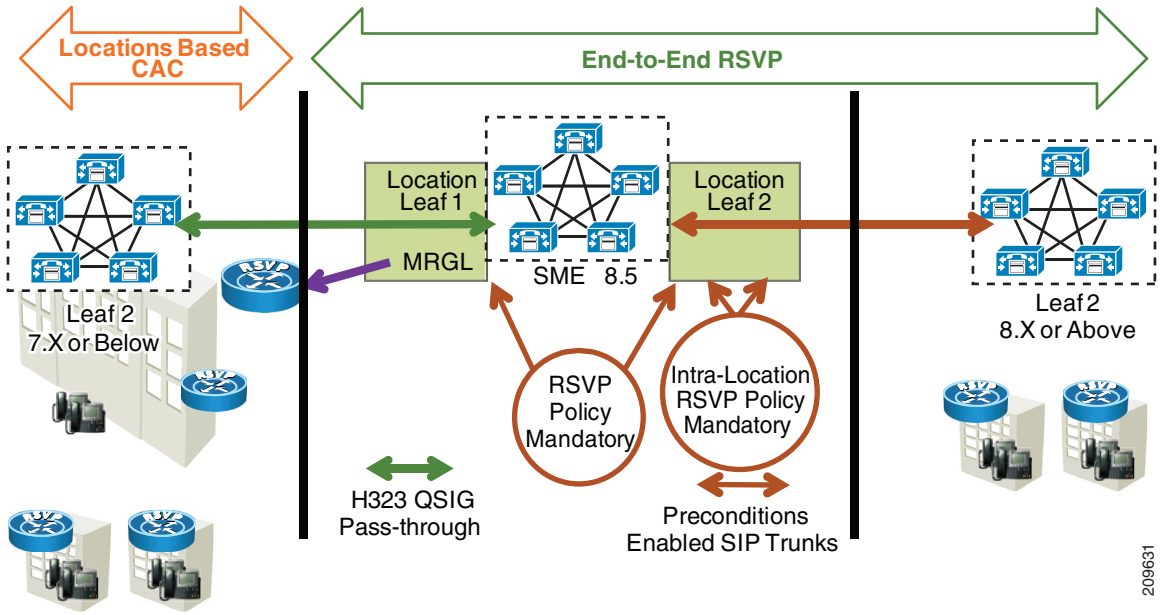
Figure 11-78 Session Management Edition Deployment with Leaf Clusters with SIP Precondition Support



Session Management Edition Design with Mixed Leaf Clusters (with and without SIP Precondition Support)

Unified CM Session Management Edition can also support a mixed environment where on one trunk it is supporting RSVP SIP Preconditions on the trunk to a leaf cluster that supports RSVP locally and RSVP SIP Preconditions (see example Leaf 2 in Figure 11-79) and on another trunk Session Management Edition associates an RSVP Agent and is doing RSVP locally to the trunk to a cluster that does not support RSVP SIP Preconditions (see Leaf 1 in Figure 11-79).

Figure 11-79 Session Management Edition Design with Mixed Leaf Clusters (with and without SIP Precondition Support)



Multiple Clusters Sharing a Single Platform for RSVP Agent

In some cases there might be the need to share a single router platform supporting RSVP Agent across multiple Unified CM clusters. In this case a single router platform supporting RSVP Agent, such as a Cisco Integrated Services Router (ISR), can be configured with multiple RSVP Agents, with each agent registered to a separate Unified CM cluster, each with dedicated software sessions. For information on the number of supported RSVP Agent sessions per platform, refer to the *Cisco RSVP Agent Data Sheet*, available at

http://www.cisco.com/en/US/products/ps6832/products_data_sheets_list.html

Figure 11-80 illustrates this with a Session Management Edition deployment co-located with two leaf clusters at the headquarters site. In this example one leaf cluster is Unified CM 7.*x* while the other leaf cluster and the Session Management Edition cluster is version 8.5. Each of the three clusters is configured with an RSVP Agent that supports 300 software sessions at the headquarters, and both leaf clusters share a Cisco ISR platform for RSVP Agent supporting 100 sessions each at two remote locations.

Figure 11-80 Session Management Edition Co-Located with Leaf Clusters (Multiple Clusters Sharing a Single Platform)