C H A P T E R **9**

Call Admission Control

Last revised on: June 4, 2010

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 9-1.

Figure 9-1   *Why Call Admission Control is Needed*
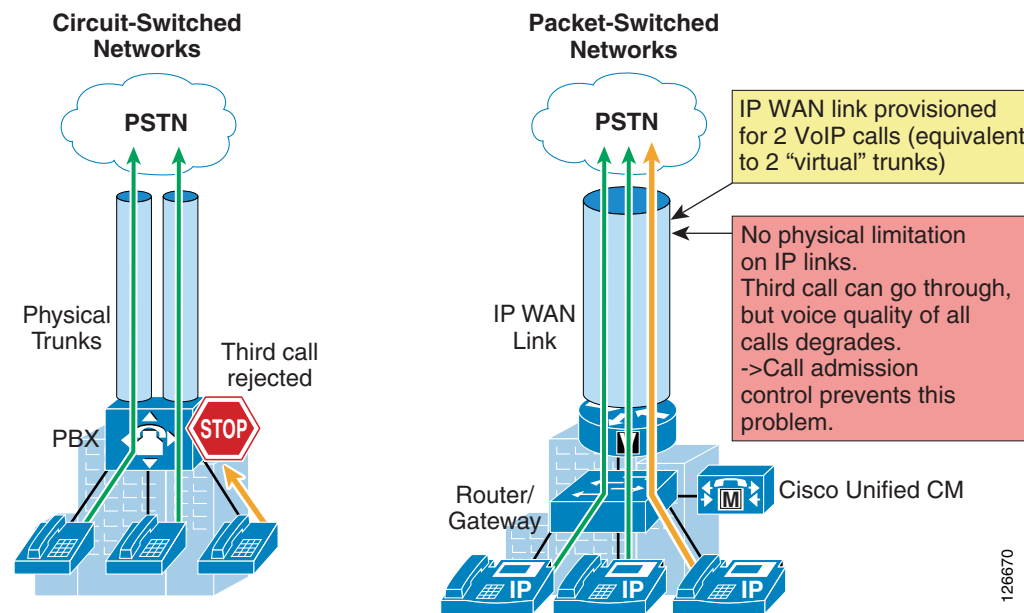
As shown on the left side of Figure 9-1, traditional TDM-based PBXs operate within circuit-switched networks, where a circuit is established each time a call is set up. As a consequence, when a legacy PBX is connected to the PSTN or to another PBX, a certain number of physical trunks must be provisioned. 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 9-1. 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:

- Design Recommendations for Call Admission Control, page 9-62

    This section summarizes the key best practices, recommendations, and notes about call admission control for the readers who are already familiar with the principles and mechanisms described in the remainder of this chapter.

- Call Admission Control Principles, page 9-3

    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 Elements, page 9-12

    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 locations, Cisco IOS gatekeeper, RSVP, and the Cisco Unified Border Element.

- Call Admission Control Design, page 9-37

    This section shows how to apply and combine the mechanisms described in the previous sections, based on the IP WAN topology (simple hub-and-spoke, two-tier hub-and-spoke, MPLS, or other topologies) and also based on the Cisco Unified Communications Manager deployment model adopted.

What's New in This Chapter

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

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

New or Revised Topic	Described in:
Cisco RSVP Agent	Unified CM RSVP-Enabled Locations, page 9-18
The name of the Cisco Multiservice IP-to-IP Gateway (IP-IP Gateway) has been changed to the Cisco Unified Border Element.	Cisco IOS Gatekeeper and Cisco Unified Border Element with RSVP, page 9-29 Cisco Unified Border Element Approach, page 9-60
Unified CM locations and regions	Locations and Regions Settings, page 9-14 Unified CM Support for Locations and Regions, page 9-15

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, then it presents the principles of topology-aware call admission control.
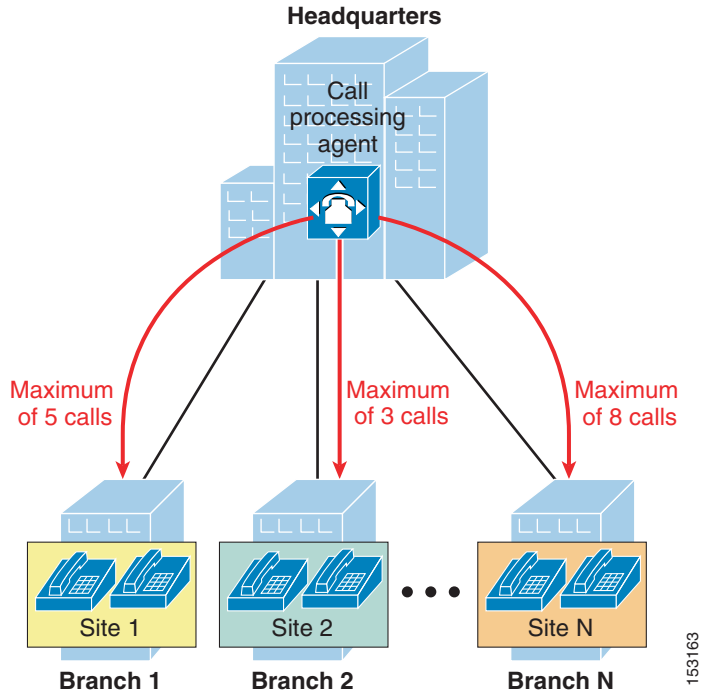
Topology-Unaware Call Admission Control

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 9-2, 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 9-2   *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 9-3.

Figure 9-3   *Domain of Applicability of Topology-Unaware Call Admission Control*
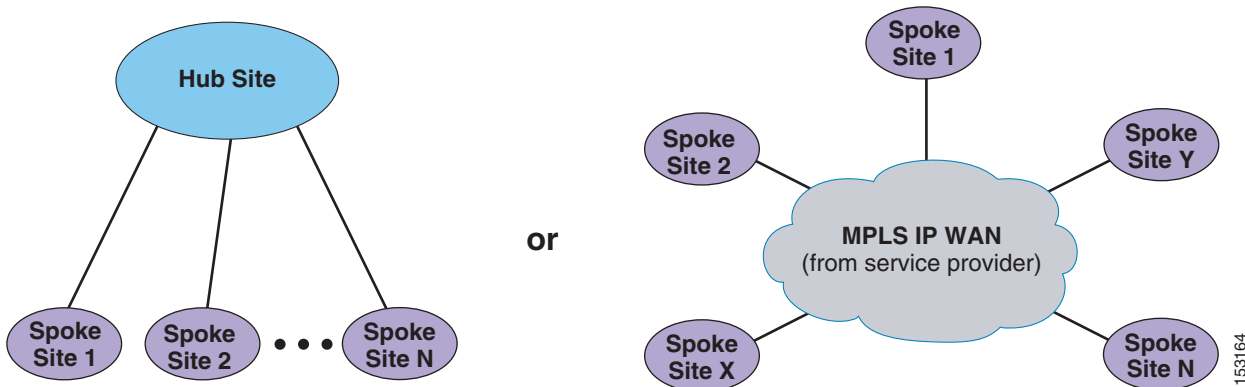
In a hub-and-spoke network or MPLS-based network such as those shown in Figure 9-3, each spoke site is assigned to a "site" within the call processing agent, and the number of calls or amount of bandwidth for that "site" is configured to match the bandwidth available for voice (and/or video) on the IP 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 9-4 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 9-4*   ***Topology Characteristics of Typical Enterprise Networks***
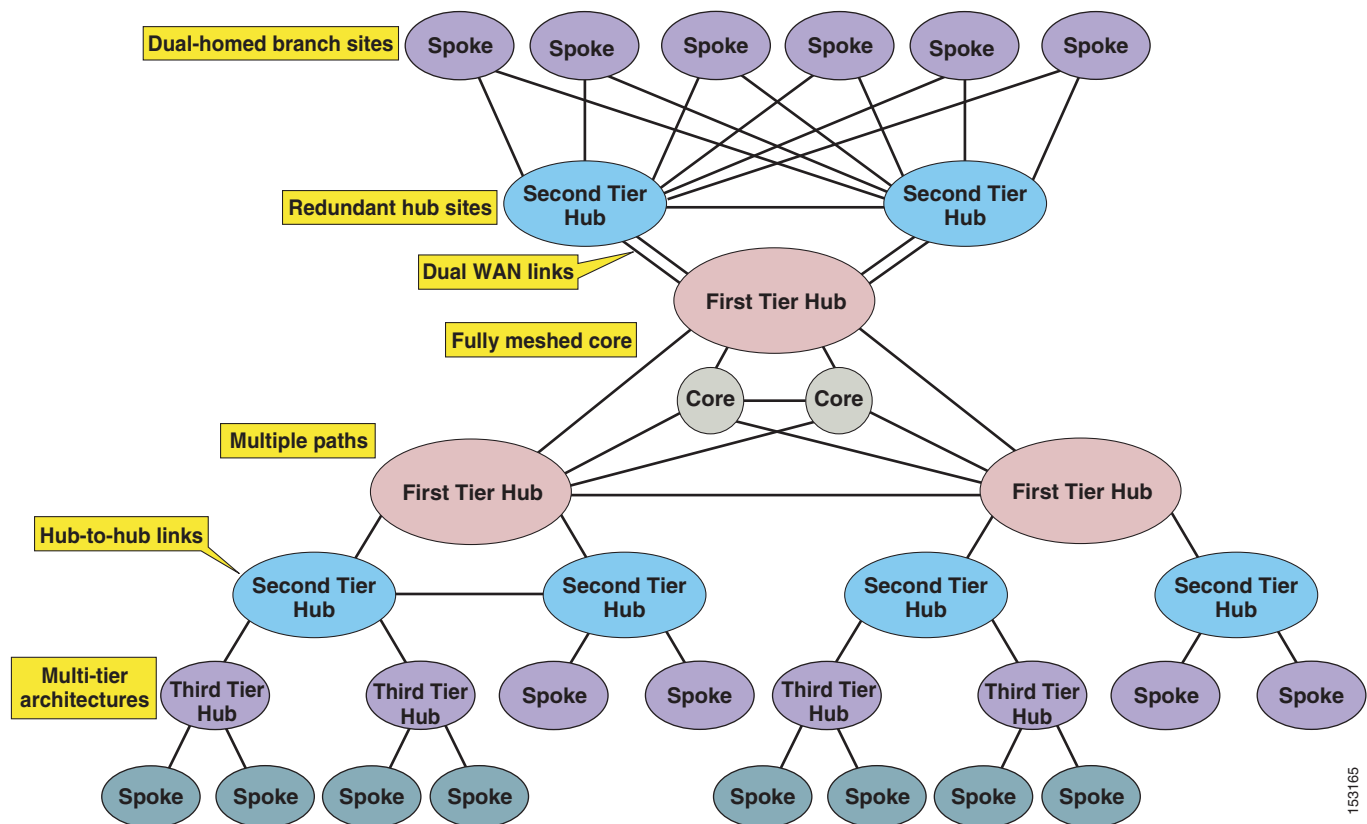
As explained in the section on Call Admission Control Design, page 9-37, it is sometimes possible to adapt a topology-unaware call admission control mechanism to a complex network topology, but there are limitations in terms of when this approach can be used and what behavior can be achieved. For example, 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 9-5 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 9-5   Topology-Unaware Call Admission Control in Presence of Dual Links
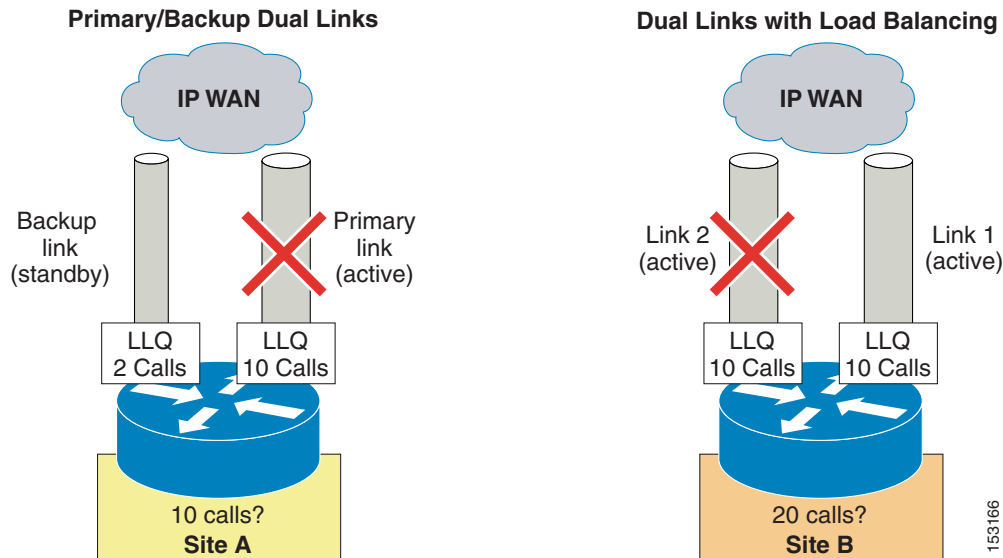
For the first example in Figure 9-5, branch office A is normally connected to the IP WAN via a primary link, whose Low Latency Queuing (LLQ) bandwidth is provisioned to allow a maximum of 10 simultaneous calls. When this primary link fails, a smaller backup link becomes active and preserves the connectivity 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 9-6. 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 3-46.
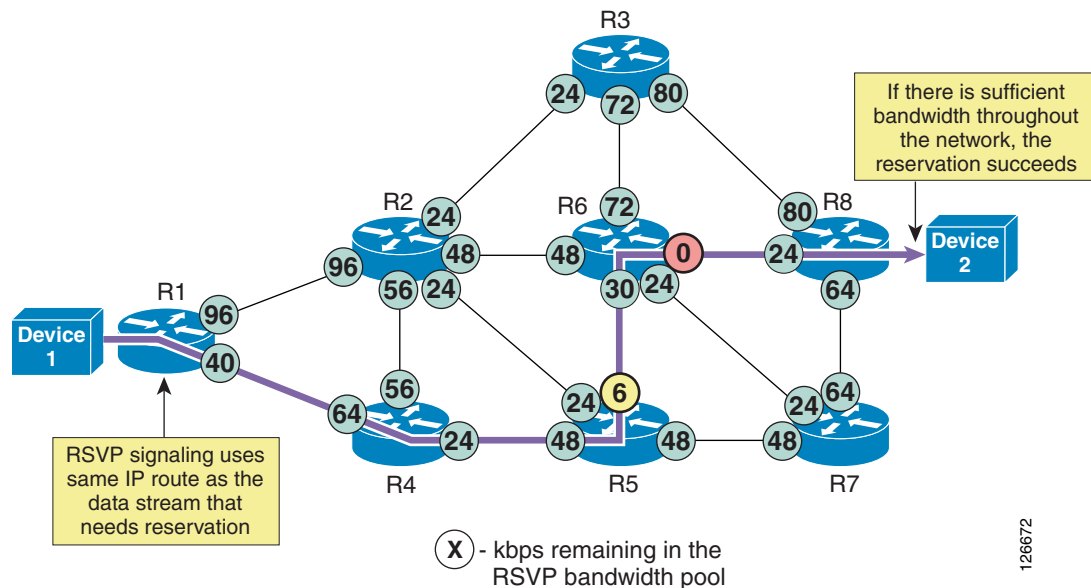
Assume that RSVP is enabled on each router interface in the network shown in Figure 9-6 and that the numbers shown in the circles represent the amount of available RSVP bandwidth remaining on each interface.

Figure 9-6   Sample Network to Show RSVP Principles

XX - kbps remaining in the
RSVP bandwidth pool

126671
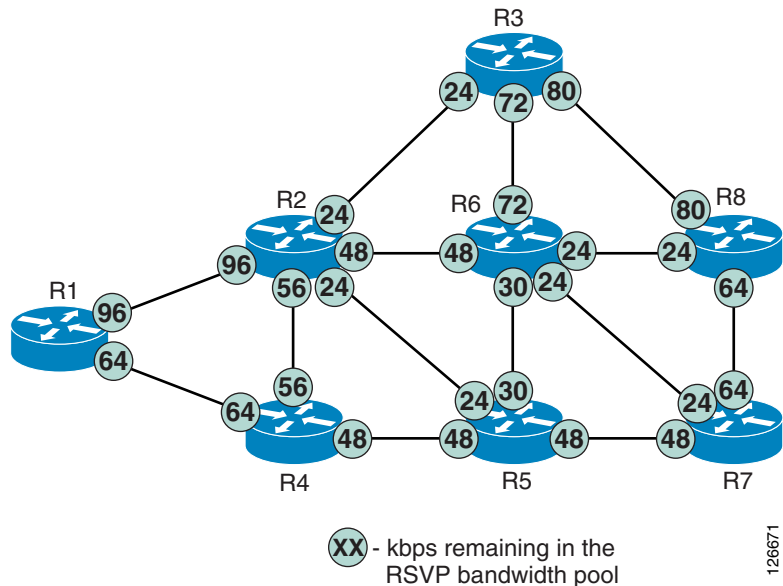
Now consider an RSVP-enabled application that wants to reserve a certain amount of bandwidth for a data stream between two devices. This scenario is depicted in Figure 9-7, which shows a particular data stream that requires 24 kbps of bandwidth from Device 1 to Device 2.

Figure 9-7   RSVP Signaling for a Successful Reservation


If there is sufficient bandwidth throughout the network, the reservation succeeds

RSVP signaling uses same IP route as the data stream that needs reservation

X - kbps remaining in the
RSVP bandwidth pool

126672

The following considerations apply to Figure 9-7:

- 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 9-7.

- 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 3-46, 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 9-8.

Figure 9-8   RSVP Signaling for an Unsuccessful Reservation
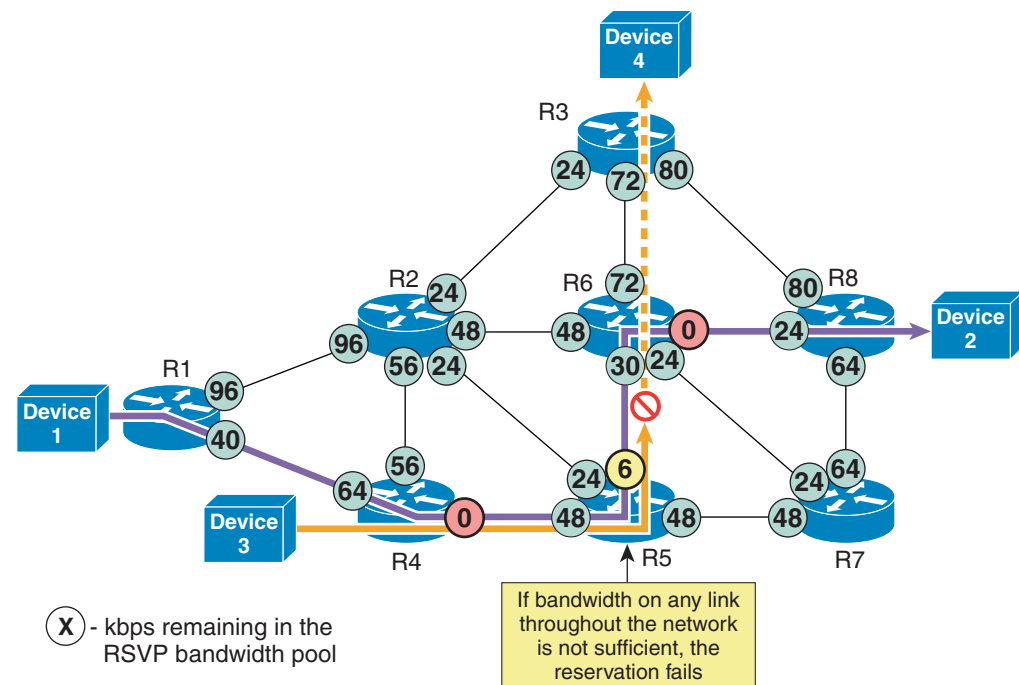
The following considerations apply to Figure 9-8:

- 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 9-8.

- 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 9-9, 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 9-9   Topology-Aware Call Admission Control for Dual Links
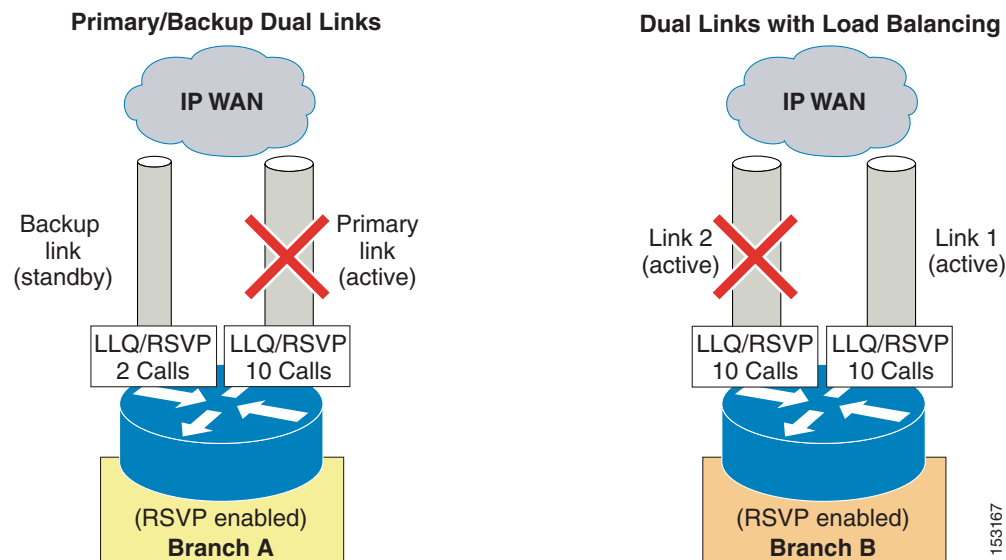
Similar considerations apply to branch B, connected to the IP WAN via two load-balanced links, as shown on the right side of Figure 9-9. RSVP is enabled on each of the two router interfaces, with a bandwidth value that matches the LLQ configuration (in this case, enough bandwidth for 10 calls). Branch 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 Simple MPLS Topologies, page 9-45.) 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 Elements

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 Static Locations, page 9-12

  - Cisco IOS Gatekeeper Zones, page 9-16

- Topology-aware mechanisms

  - Unified CM RSVP-Enabled Locations, page 9-18

  - Cisco IOS Gatekeeper and Cisco Unified Border Element with RSVP, page 9-29

Note Cisco Unified CM 5.0 introduces topology-aware call admission control by extending the concept of *locations*, which already existed in previous releases. Therefore, this document refers to the former, topology-unaware mechanism as *static locations* and to the new, topology-aware mechanism as *RSVP-enabled locations*.

Unified CM Static Locations

Unified CM provides a simple mechanism known as *static locations* for implementing call admission control in the centralized call processing deployment. When you configure a device in Unified CM, the device can be assigned to a location. A certain amount of bandwidth will be allocated for calls to or from each location. The locations configured in Unified CM are virtual locations and not real, physical locations. Unified CM has no knowledge of the physical locations of devices. Therefore, if a device is moved from one physical location to another, the system administrator must perform a manual update on its location configuration so that Unified CM can correctly calculate bandwidth allocation for that device. Each device is in location Hub_None by default. Location Hub_None is a special location that is configured by default with unlimited audio and video bandwidth, and location Hub_None cannot be deleted. If the devices at a branch location are configured in the Hub_None location, none of the phone calls to or from that branch device will be subject to any call admission control.

Unified CM allows you to define a voice and video bandwidth pool for each location. If the location's audio and video bandwidth are configured as **Unlimited**, there will be unlimited bandwidth available for that location, and every audio or video call to or from that location will be permitted by Unified CM. On the other hand, if the bandwidth values are set to a finite number of kilobits per second (kbps), Unified CM will allow calls in and out of that location as long as the aggregate bandwidth used by all active calls is less than or equal to the configured values. If the video bandwidth for the location is configured as **None**, every video call to or from that location is denied but the video calls within the same location are not affected.

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

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 static locations call admission control mechanism also takes into account the mid-call changes of 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. If this video call changes to an audio-only call via 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. Calls that change from audio to video will cause the opposite change of bandwidth allocation.

Table 9-2 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 allocated from the location's audio bandwidth pool. 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 speed of 384 kbps, Unified CM will allocate 384 kbps from the video bandwidth pool.

Table 9-2   Amount of Bandwidth Requested by the Static Locations Algorithm

Call Speed	Static Location 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 static location 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 configuration for the location Branch 1 allocates 256 kbps of available audio bandwidth and 384 kbps of available video bandwidth. In this case, the Branch 1 location 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 location 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).

Note   Call admission control does not apply to calls between devices within the same location.

When a call is placed from one location to the other, Unified CM deducts the appropriate amount of bandwidth from both locations. For example, a G.729 call between two locations causes Unified CM to deduct 24 kbps from the available bandwidth at both locations. When the call has completed, Unified CM returns the bandwidth to the affected locations. If there is not enough bandwidth at either branch location, 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-site call is denied by call admission control, Unified CM can automatically reroute the call to the destination via the PSTN connection by means of the Automated Alternate Routing (AAR) feature. For detailed information on the AAR feature, see Automated Alternate Routing, page 10-86.

Note   AAR is invoked only when the locations-based call admission control denies the call due to a lack of network bandwidth. 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 and Regions Settings

Locations work in conjunction with regions to define the characteristics of a network link. 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 on the link, and locations define the amount of available bandwidth for the link. 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 locations (for example, a branch office)
- The amount of bandwidth that is available in the WAN for voice and fax calls coming into and going out from a location. This bandwidth value is used by Unified CM for locations-based call admission control.
- The amount of video bandwidth that is available in the WAN for calls coming into and going out from a location. This bandwidth value is used by Unified CM for locations-based 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 voice codecs used for intraregion calls
- The voice codecs used for interregion calls
- The per-call video bandwidth used for intraregion calls and interregion calls
- 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 7.1(2) and later releases support 2000 locations and 2000 regions with Cisco MCS-7845 servers. To deploy up to 2000 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:

- Intraregion Audio Codec Default
- Interregion Audio Codec Default
- Intraregion Video Call Bandwidth Default
- Interregion Video Call Bandwidth Default
- Default inter-location RSVP Policy

When adding regions, you should then select **Use System Default** for the Audio Codec and Video Call Bandwidth values. If you are using RSVP call admission control, you should also select **Use System Default** for the RSVP Setting 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 2000 regions and 2000 locations, you can modify up to 200 of them to use non-default values. With a total of 1000 or fewer regions and locations, you can modify up to 500 of them to use non-default values. Table 9-3 summarizes these limits.

Table 9-3   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	2000	2000
200 to 500	1000	1000

Note   The audio codec value is used by both voice calls and fax calls. If you plan to use G.729 as the interregion codec value, use T.38 Fax Relay for fax calls. If you plan to use fax pass-through over the WAN, change the default Interregion Audio Codec value to G.711, or else add a region for fax machines to each location with a non-default codec value of G.711 (subject to the limits in Table 9-3).

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.

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. Refer to the section on Call Admission Control Design, page 9-37, for detailed topology examples.

The Cisco 2600, 3600, 3700, 2800, 3800, and 7200 Series routers all support the gatekeeper feature. You can configure Cisco IOS gatekeepers in a number of different ways for redundancy, load balancing, and hierarchical call routing. This section focuses on the call admission control aspect of the gatekeeper feature. For redundancy and scalability considerations, refer to the section on Gatekeeper Design Considerations, page 8-26. For call routing considerations, refer to Call Routing in Cisco IOS with a Gatekeeper, page 10-109.

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).

Use the **bandwidth** command to manage the number of calls that the gatekeeper will allow, thus providing call admission control functionality. This command has several options, but the most relevant are the following:

- The **interzone** option controls the amount of bandwidth for all calls into or out of a given local zone.
- The **total** option controls the amount of bandwidth for all calls into, out of, or within a given local zone.
- The **session** option controls the amount of bandwidth per call for a given local zone.
- The **remote** option controls the total amount of bandwidth to or from all remote zones.

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 9-4 shows the bandwidth values used by the gatekeeper feature for some of the most popular call speeds.

Table 9-4   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

Table 9-4   Gatekeeper Bandwidth Settings for Various Call Speeds (continued)

Call Speed	Gatekeeper Bandwidth Value
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 3-58, 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 9-10.

Figure 9-10   Example of Cisco IOS Gatekeeper bandwidth Commands
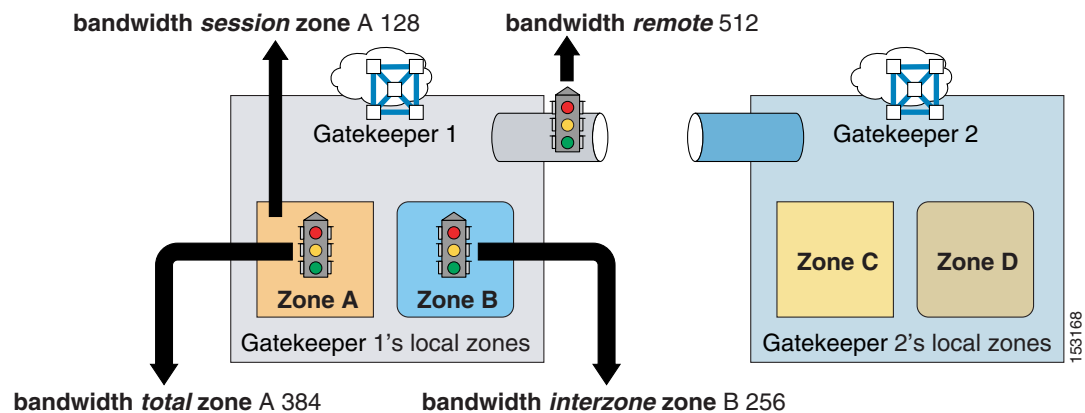
Assuming that all calls are voice-only calls using the G.711 codec, and given the configuration commands shown in Figure 9-10, the following statements hold true:

- 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 CM RSVP-Enabled Locations

Cisco Unified CM Release 5.0 introduces 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. The Cisco RSVP Agent feature has been introduced into Cisco IOS Release 12.4(6)T, and it is available on the Cisco 2800 Series and 3800 Series Integrated Services Routers platforms. There is also support for Cisco RSVP Agent on the following router platforms using the Cisco IOS Advanced IP Services Image Release 12.4(15)T5 or higher:

- Cisco 7200 Series Routers (with NPE-G1 or NPE-G2)
- Cisco 7201 Series Routers
- Cisco 7301 Series Routers

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 9-11 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 9-11 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.

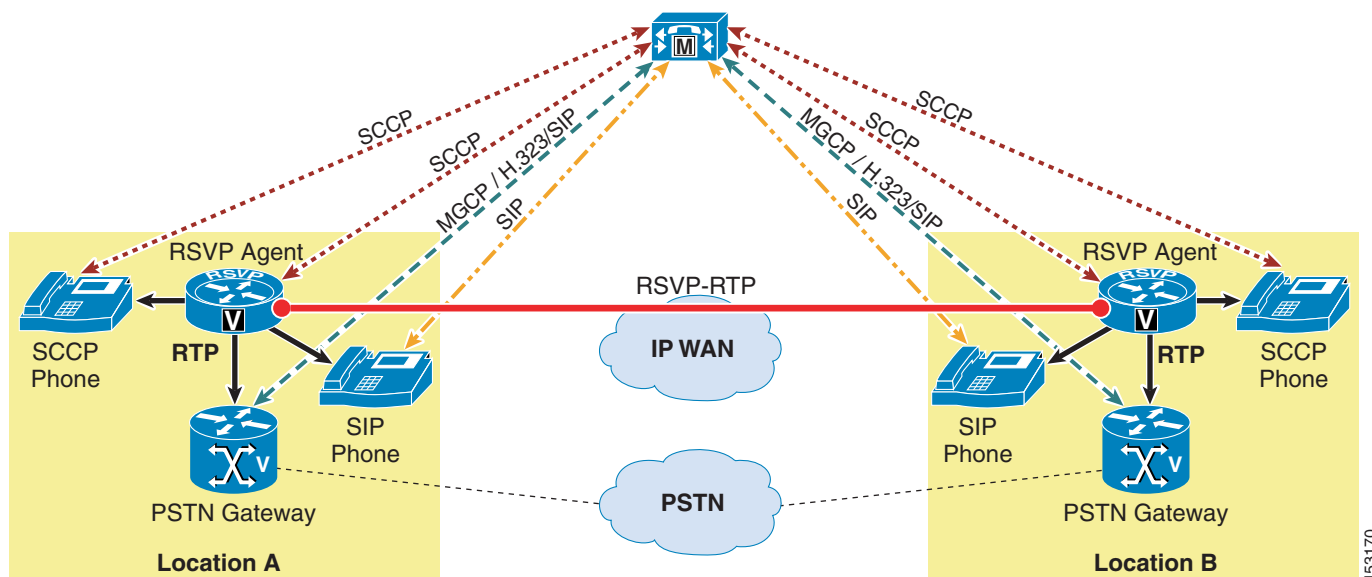*Figure 9-11*   *Protocol Flows for Locations with RSVP*

Figure 9-12 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 9-12) 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 using call admission control based on static locations.

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.

Figure 9-12   Cisco RSVP Agent Concept
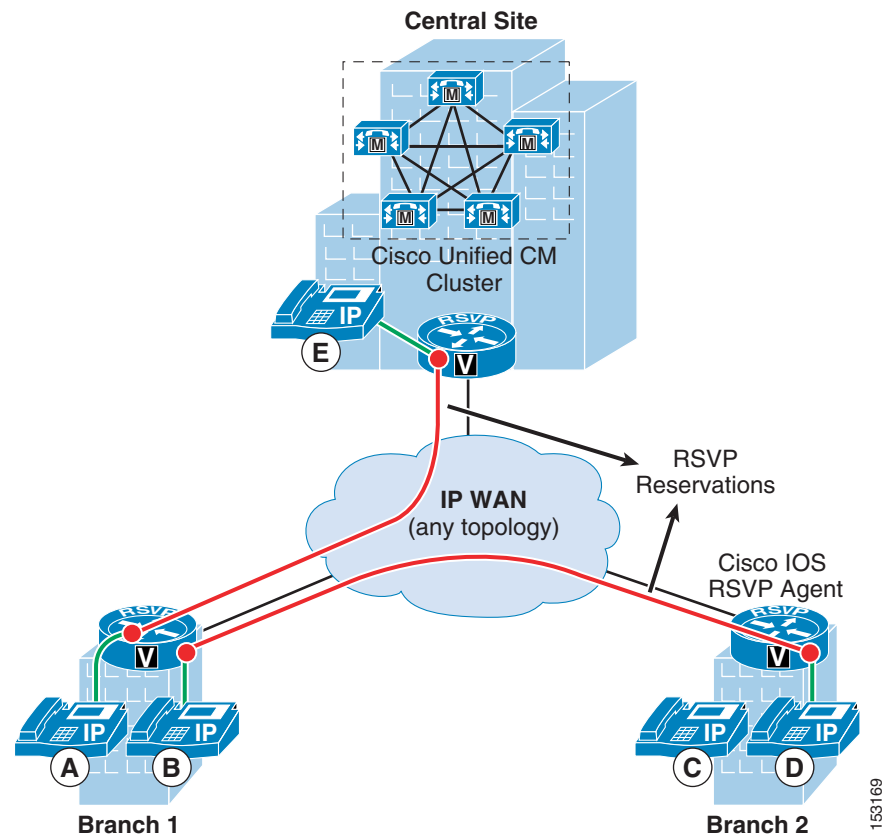
Cisco RSVP Agent Provisioning

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 6-1, 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.

The Cisco RSVP Agent can be associated to the endpoint device by a combination of the device pool, media resource group (MRG), and media resource group list (MRGL) configurations. The Cisco RSVP Agent can be included in an MRG, and the MRG can be part of an MRGL. The MRGL can be assigned to the endpoint device either directly or via the device pool. As illustrated in Figure 9-13, MRGL-Branch 1 can be associated to the IP phone either directly or via Device Pool-Branch 1. As a general rule, assign the MRGL directly to the endpoint device if the endpoint device requires an unique set of media resources; otherwise, assign the MRGL to the device pool in which the endpoint device is located.

Figure 9-13   Assigning a MRGL to an IP Phone
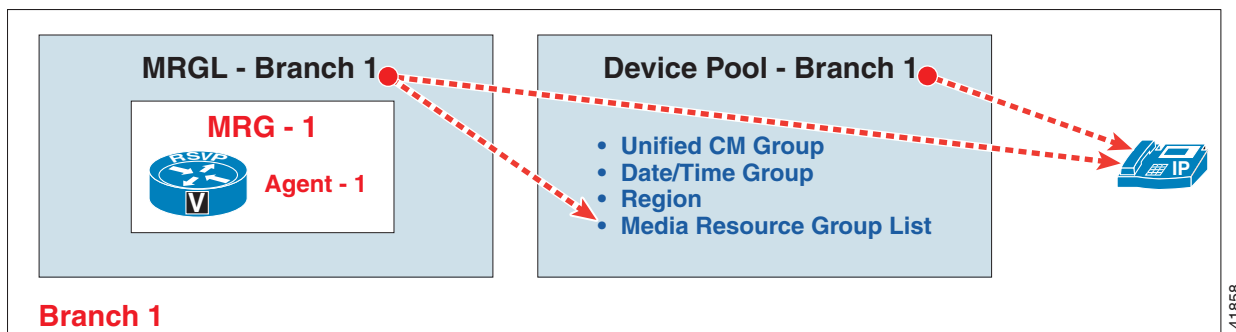
Unified CM allocates the Cisco RSVP Agent in the same way it allocates other conventional media resources including MTP, transcoder, conferencing resource, and annunciator.

Cisco recommends that you do not configure the Cisco RSVP Agent in the same MRG with other conventional media resources. Doing so could cause the Cisco RSVP Agent to be allocated for any call in need of an MTP device even though the call is not RSVP related.

Figure 9-14 shows how Cisco RSVP Agent load balancing is implemented via the MRG and MRGL configurations. For all the Cisco RSVP Agents within the same MRG, Unified CM load-balances and allocates the Cisco RSVP Agents in a round-robin fashion.

Figure 9-14   Cisco RSVP Agent Load Balancing
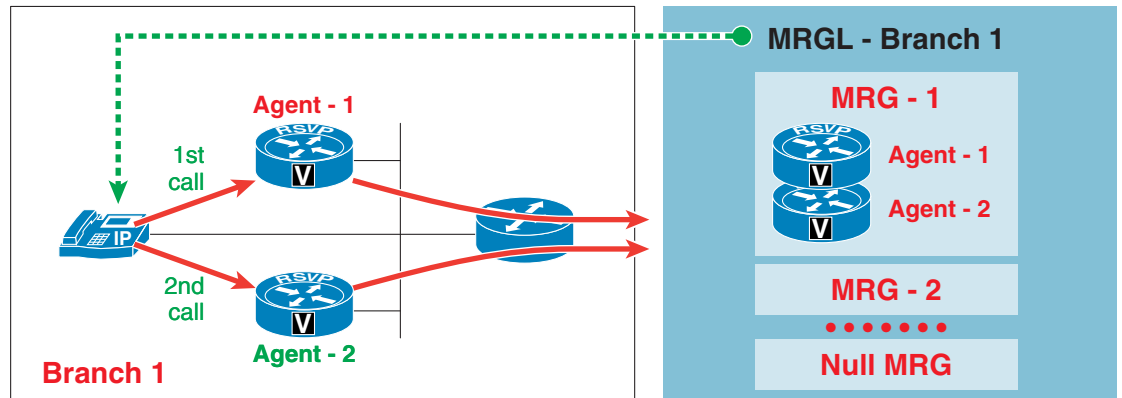
As Figure 9-14 illustrates, if both Cisco RSVP Agents in MRG-1 are available, Agent-1 will be selected for the first call and Agent-2 will be selected for the second call. If neither of the Cisco RSVP Agents in MRG-1 is available, Unified CM will try to search through MRG-2, MRG-3, and the rest of the MRGs until it finds a suitable Cisco RSVP Agent for the call. Any Cisco RSVP Agent that is not explicitly included in an MRG will be included in the Null MRG by default. Note that the Null MRG is always implicitly included as the last MRG in any MRGL configurations, but it is not displayed in Unified CM Administration. The Cisco RSVP Agent in the Null MRG is accessible by any endpoint devices in the Unified CM cluster. Therefore, Cisco recommends that you always configure a Cisco RSVP Agent in an MRG. For details on Unified CM's media resource allocation process and the associated best practices, see Media Resources, page 6-1.

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 9-15 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 9-15   Configuring Maximum Sessions on the Cisco RSVP Agent
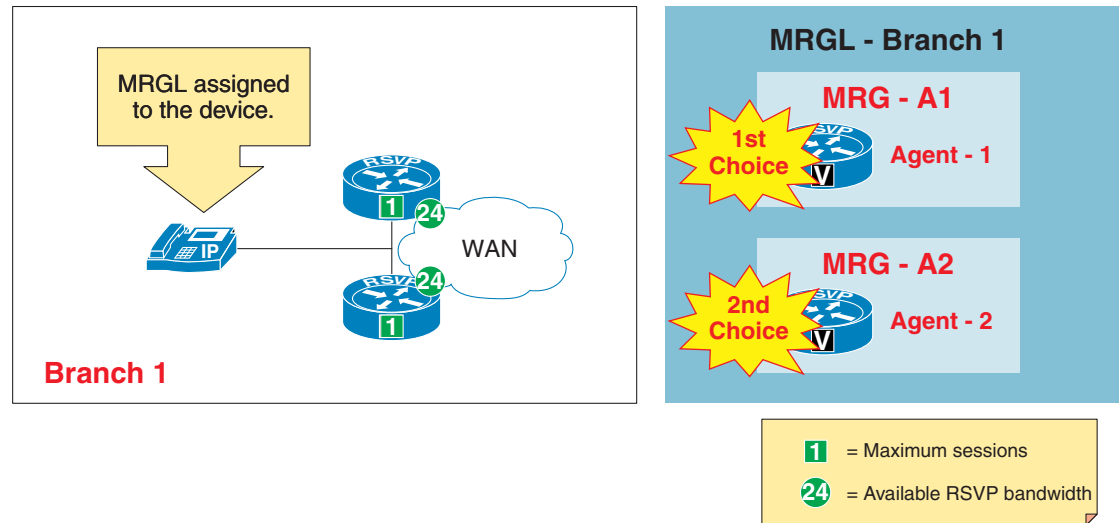
Pass-Through Codec

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 codec configured in Unified CM Administration. For example, if the G.729 codec is the inter-region codec 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 2800 IOS 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 5.0.1
sccp ccm 20.11.1.51 identifier 2 priority 2 version 5.0.1
```

```
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 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 only static-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.

Figure 9-16 shows both the default and recommended configurations of the RSVP clusterwide parameters. 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**.

*Figure 9-16   Setting RSVP Clusterwide Parameters*

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 from Static Locations to RSVP Call Admission Control

The example in this section illustrates the best practices for migrating from the traditional static-locations call admission control to the RSVP-based call admission control mechanism.

Figure 9-17 shows a centralized call processing deployment with the static-locations call admission control mechanism. There are four locations in the Unified CM cluster, including the Hub_None location and three branches. For simplicity of illustration, the bandwidth used in this example refers only to the voice stream bandwidth. Table 9-5 and Table 9-6 show that every branch location is statically provisioned with 256 kbps of bandwidth, and the Hub_None location is provisioned with **Unlimited** bandwidth. The RSVP setting between any pair of locations is configured with **Use System Default**, and the clusterwide RSVP setting is configured with the default value **No Reservation**.

Figure 9-17 **Configuration with Static Locations for Call Admission Control**
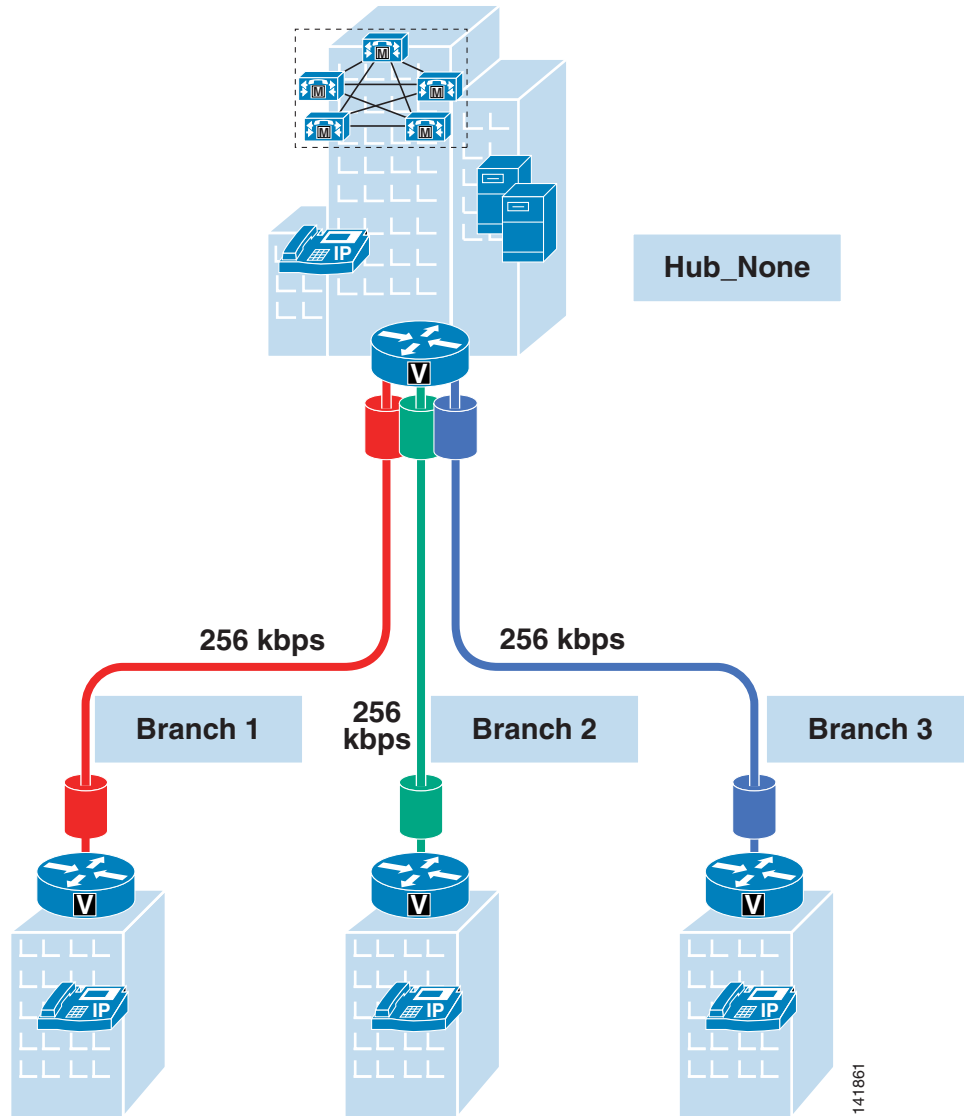
Table 9-5 **Locations and Bandwidth Settings for the Example in Figure 9-17**

Location Name	Bandwidth
Hub_None	Unlimited
Branch 1	256 kbps
Branch 2	256 kbps
Branch 3	256 kbps

Table 9-6   RSVP Policy for the Example in Figure 9-17

Locations Pair		Policy
Any	Any	No Reservation

To migrate to RSVP-based call admission control, Cisco recommends migrating one location at a time. For example, if Branch 1 is the first location to be migrated, observe the following procedures:

- Set up a Cisco RSVP Agent in the Branch 1 location and assign it to the Branch 1 MRG and MRGL to associate it with the Branch 1 IP phones.

- Set up another Cisco RSVP Agent in the Hub_None location and include this Cisco RSVP Agent in the MRG and MRGL that is associated to all the IP phones in the remaining three locations, including the Hub_None location. Note that this Cisco RSVP Agent should not be included in the Null MRG or Branch 1 MRG, otherwise it is possible that a Branch 1 phone will use the Cisco RSVP Agent at the Hub_None location to make RSVP reservations.

- Configure the Branch 1 bandwidth as **Unlimited**.

- Configure the RSVP setting between Branch 1 and any other location as **Mandatory**. For example, for a call between Branch 1 and Branch 2 phones, the voice stream will still be hair-pinned through the Hub_None location. For the first call leg between the Branch 1 and Hub_None locations, a RSVP reservation will be made between the Branch 1 and Hub_None Cisco RSVP Agents. For the second call leg between the Hub_None and Branch 2 locations, Unified CM will perform call admission control based on static locations by checking the bandwidth availability for the Branch 2 location.

Table 9-7 and Table 9-8 show the location bandwidth and RSVP policy configuration after the migration at Branch 1.

Table 9-7   Locations and Bandwidth Settings After Migration of Branch 1

Location Name	Bandwidth
Hub_None	Unlimited
Branch 1	Unlimited
Branch 2	256 kbps
Branch 3	256 kbps

Table 9-8   RSVP Policy After Migration of Branch 1

Locations Pair		Policy
Branch 1	Any	Mandatory
All other locations	All other locations	No Reservation

Table 9-9 and Table 9-10 show the location bandwidth and RSVP policy configurations after the clusterwide migration. With the clusterwide migration completed, any inter-site call will require a RSVP reservation to be made directly between two Cisco RSVP Agents, and the voice stream will be transported over the bandwidth reservation path.

The following procedures can be used to migrate Branch 2 and Branch 3 to RSVP call admission control:

- Set up a Cisco RSVP Agent in the Branch 2 location and assign it in the Branch 2 MRG and MRGL that is associated with the Branch 2 IP phones. Be sure to remove the Cisco RSVP Agent of the Hub_None location from the Branch 2 MRG so that the Cisco RSVP Agent in the Hub_None location will no longer be accessed by the IP phones in Branch 2.

- Configure the Branch 2 bandwidth as **Unlimited**.

- Configure the RSVP setting between Branch 2 and any other location as **Mandatory**.

- Set up a Cisco RSVP Agent in the Branch 3 location and assign it in the Branch 3 MRG and MRGL that is associated with the Branch 3 IP phones. Be sure to remove the Cisco RSVP Agent of the Hub_None location from the Branch 3 MRG so that the Cisco RSVP Agent of the Hub_None location will no longer be accessed by the IP phones in Branch 3.

- Configure the Branch 3 bandwidth as **Unlimited**.

- Configure the RSVP setting between Branch 3 and any other location as **Mandatory**.

Table 9-9   *Locations and Bandwidth Settings After Migration Is Complete*

Location Name	Bandwidth
Hub_None	Unlimited
Branch 1	Unlimited
Branch 2	Unlimited
Branch 3	Unlimited

Table 9-10   *RSVP Policy After Migration Is Complete*

Locations Pair		Policy
Any	Any	Mandatory

RSVP Application ID

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 Cisco IOS Release 12.4(6)T or later on the Cisco RSVP Agent router and Cisco Unified CM Release 5.0. 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.

The RSVP Application ID uses a model to separate the bandwidth of voice and video traffic that is different than the static locations model. In static locations, the voice and video streams of a video call are both deducted from the video bandwidth counter. When using the RSVP Application ID, the voice stream is deducted from the audio bandwidth pool while the video stream is deducted from the video bandwidth pool. With this change in the call admission control model, you can now reserve a certain amount of bandwidth for voice calls and allow them to use the entire available bandwidth in the priority queue, thus ensuring that all the available bandwidth can be used for voice calls if no video calls are in progress. If there is enough available bandwidth in the priority queue, calls can optionally be enabled

for video. You can set limits on how much bandwidth the video-enabled calls can consume, but if voice calls are consuming all the available bandwidth, it might not be possible to place a video call at all. For detailed information on how to configure the RSVP Application ID, RSVP policy, and LLQ, see RSVP Application ID, page 3-56.

Cisco IOS Gatekeeper and Cisco Unified Border Element with RSVP

The Cisco Unified Border Element (formerly the Cisco Multiservice IP-to-IP Gateway) can be used to ease the restriction of hub-and-spoke topologies for the IP WAN connections between Unified CM clusters and/or H.323 gateways.

This Cisco IOS feature provides a mechanism to enable H.323 Voice over IP (VoIP) and videoconferencing calls from one IP network to another. The main purpose of the Cisco Unified Border Element gateway is to provide a control point and a demarcation for VoIP and video calls traversing administrative domains. This gateway performs most of the same functions of a PSTN-to-IP gateway, but typically joins two IP call legs rather than a PSTN and an IP call leg.

The most interesting feature of the Cisco Unified Border Element in an enterprise IP Communications environment is that it can generate an RSVP reservation for each call that traverses it. As described in the section on Topology-Aware Call Admission Control, page 9-7, RSVP is a network-based signaling protocol that provides a topology-aware call admission control mechanism that does not require a hub-and-spoke topology but works with any network topology.

Note   For detailed information on Cisco support for deploying the Cisco Unified Border Element with RSVP, refer to the *Cisco Unified Border Element Application Guide* available at http://www.cisco.com.

As a consequence, you can perform call admission control over any IP WAN topology by inserting two Cisco Unified Border Elements in the call flow and enabling RSVP between them. Figure 9-18 shows a basic example where two sites, A and B, each have a Unified CM cluster and are connected via an IP WAN that has an arbitrary topology. A Cisco Unified Border Element is also located at each site, and the two Unified CM clusters are configured so that all inter-site calls are routed via a trunk that points to the local Cisco Unified Border Element. When a call is set up between Site A and Site B, the following events occur:

- Unified CM at Site A sets up a call through an H.323 trunk to Site A's Cisco Unified Border Element. (This is call leg 1 in the figure.)

- Site A's Cisco Unified Border Element attempts to establish another call to Site B's Cisco Unified Border Element, but first it uses RSVP to allocate bandwidth resources along the IP WAN path.

- If the RSVP reservation is successful, call leg 2 is established between the two Cisco Unified Border Elements.

- Site B's Cisco Unified Border Element generates another call to Site B's Unified CM cluster. (This is call leg 3 in the figure.)

Figure 9-18   Simple Example of the Cisco Unified Border Element for RSVP Call Admission Control
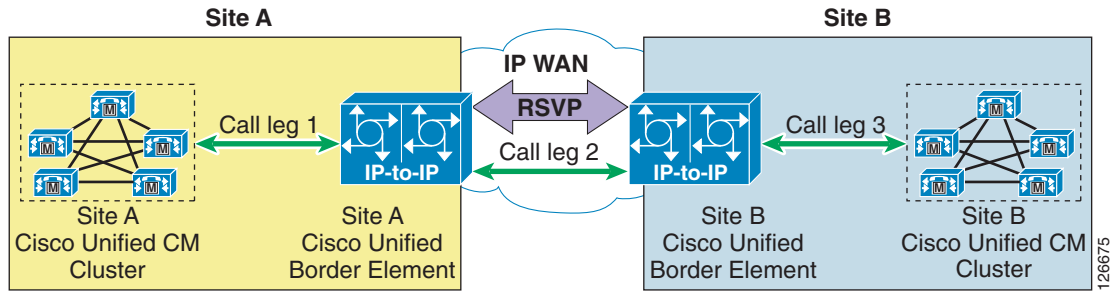
The example in Figure 9-18 is a simple scenario in which all calls between Unified CM clusters are routed via a pair of Cisco Unified Border Elements. However, in many real-world cases this approach might not prove scalable or flexible enough. For these cases, Cisco IOS gatekeepers can be used to provide a wider range of communication options between Unified CM clusters, H.323 gateways, H.323 videoconferencing endpoints, and Cisco Unified Border Elements.

Note   All scenarios involving Cisco Unified Border Elements described in this section apply to calls between different Unified CM clusters. Cisco does not recommend inserting Cisco Unified Border Elements for calls between endpoints registered to the same Unified CM cluster. For RSVP-based call admission control between endpoints registered to the same Unified CM cluster, see Unified CM RSVP-Enabled Locations, page 9-18.

Via-Zone Gatekeeper

Traditional Cisco IOS gatekeeper functionality has been extended to accommodate for Cisco Unified Border Elements through the concept of a *via-zone gatekeeper*. A via-zone gatekeeper differs from legacy gatekeepers in how it uses LRQ and ARQ messages for call routing. Using via-zone gatekeepers will maintain normal gatekeeper functionality and extend it with additional features. Legacy gatekeepers examine incoming LRQs based on the called number, and more specifically the dialedDigits field in the destinationInfo portion of the LRQ. Via-zone gatekeepers look at the origination point of the LRQ before looking at the called number. If an LRQ comes from a gatekeeper listed in the via-zone gatekeeper's remote zone configurations, the gatekeeper checks to see that the zone remote configuration contains an **invia** or **outvia** keyword. If the configuration contains these keywords, the gatekeeper uses the new via-zone behavior; if not, it uses legacy behavior.

For ARQ messages, the gatekeeper determines if an **outvia** keyword is configured on the destination zone. If the **outvia** keyword is configured and the zone named with the **outvia** keyword is local to the gatekeeper, the call is directed to a Cisco Unified Border Element in that zone by returning an ACF pointing to the Cisco Unified Border Element. If the zone named with the **outvia** keyword is remote, the gatekeeper sends a location request to the outvia gatekeeper rather than the remote zone gatekeeper. The **invia** keyword is not used in processing the ARQ.

Figure 9-19 shows an example of how Cisco Unified Border Elements and via-zone gatekeepers can be used in conjunction with Unified CM clusters and legacy gatekeepers to provide call routing and call admission control. The following considerations apply to this scenario:

- Site A's Unified CM clusters use the Site A gatekeeper to route calls directly between them.

- The Site A gatekeeper sends all calls directed to Site B's E.164 numbers to the Site A via-zone gatekeeper.

- The Site A via-zone gatekeeper inserts a Cisco Unified Border Element for all calls coming from or destined to the Site A gatekeeper.

- The Site A Cisco Unified Border Element attempts RSVP reservations before sending calls toward Site B's Cisco Unified Border Element.

- The Unified CM clusters, gatekeepers, and the Cisco Unified Border Element at Site B are configured in a similar fashion to their counterparts at Site A.

Figure 9-19   Cisco Unified Border Element for RSVP Using Via-Zone Gatekeepers
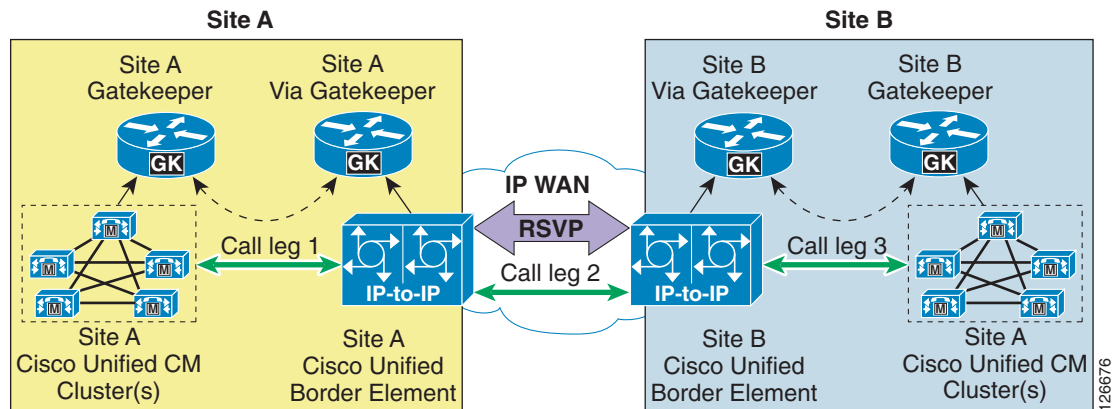
Design Best Practices

When deploying Cisco Unified Border Elements in conjunction with Unified CM in order to enable RSVP call admission control in the IP WAN, observe these design best practices:

- When configuring trunks in Unified CM for either voice or video communication with other Unified CM clusters through one or more Cisco Unified Border Elements, use gatekeeper-controlled H.225 trunks. When using Cisco IOS Release 12.4(6)T or later and Cisco Unified CM Release 4.1 or later, MTP resources are no longer required to invoke supplementary services such as hold/resume, transfer, and conference across Cisco Unified Border Elements. In order to ensure interoperability, you must configure the following items:

    - On Unified CM Administration trunk configuration page, leave the **Media Termination Point required** field unchecked (default configuration), and also uncheck the field **Wait for Far End H.245 Terminal Capability Set**.

    - In Unified CM Administration, under the Advanced Service Parameters page for Unified CM, set the **Send H225 User Info Message** field to **H225 Info For Call Progress Tone**.

    - On the Cisco Unified Border Elements, configure the following Cisco IOS commands to ensure interoperability with Unified CM when invoking supplementary services:

    ```
    voice service voip
       h323
          emptycapability
          h245 passthru tcsnonstd-passthru
    ```

- The MTP resources are preferred in some deployments to provide a proxy functionality and to terminate the signaling and media streams on behalf of endpoint devices. If the MTP resources are required, Cisco recommends that you place the MTP resources in the same site as the Cisco Unified

Border Element to avoid further IP WAN bandwidth usage when calling across the intercluster trunks. These MTP resources may be software-based (for example, on a Cisco MCS server or on a Cisco IOS router) or hardware-based (for example, on a Catalyst 6500 with a Cisco Communications Media Module or on a Cisco IOS router with an NM-HDV network module). Refer to the chapter on Media Resources, page 6-1, for a complete list of available MTP resources. However, when MTPs are used, media packets will transit through the initial MTP resources for the entire duration of the call, with the potential for hairpinning in the case of subsequent call transfers. Note that video calls will not be established across clusters through Cisco Unified Border Elements if the **Media Termination Point required** option is checked on the H.225 trunks (because MTPs do not support video calls).

- Configure the Cisco Unified Border Element as an H.323 gateway in Unified CM only when you want all intercluster calls to use the Cisco Unified Border Element. In this case, the Cisco Unified Border Element can still use a gatekeeper to resolve the remote destination.

- Configure gatekeeper-controlled intercluster trunks in Unified CM when you want to use a gatekeeper to resolve intercluster calls and decide whether they need to go through an Cisco Unified Border Element or be routed directly. This approach is more flexible and allows for greater scalability.

- Compatibility with Cisco Unified CM 3.3(2) or later on the Cisco Unified Border Element is available on Cisco IOS Release 12.3(1) or later. Cisco recommends that you use Cisco IOS Release 12.4(6)T or later.

- Keep the gatekeeper and via-zone gatekeeper functions separate by running them on different router platforms. Each Cisco Unified Border Element should have a dedicated via-zone gatekeeper.

- You can run the via-zone gatekeeper function and the Cisco Unified Border Element function on the same router platform (co-resident); however, be aware of the scalability requirements described in the section on Redundancy, page 9-32.

- Do not use Cisco Unified Border Elements for calls between endpoints controlled by the same Unified CM cluster.

- Use RSVP-enabled locations to provide topology-aware call admission control between endpoints controlled by the same Unified CM cluster.

- Use the following options under the dial-peer configuration when enabling RSVP reservations for the Cisco Unified Border Element:

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

This configuration ensures that for each voice or video call, the Cisco Unified Border Element 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). For more information on configuration details, see Configuration Guidelines, page 9-33.

Redundancy

Redundancy and scalability can be provided by registering multiple Cisco Unified Border Elements with the same via-zone gatekeeper and in the same via-zone. The via-zone gatekeeper will automatically distribute the incoming calls between all the Cisco Unified Border Elements registered in the same via-zone, using a round-robin algorithm.

When a Cisco Unified Border Element fails, it loses its registration to the via-zone gatekeeper, and the gatekeeper removes it from the list of available resources.

It is also possible to manually configure maximum-load thresholds within the Cisco Unified Border Element so that a certain Cisco Unified Border Element stops being selected for new calls when more than a certain percentage of its circuits are in use, and it becomes available again when the circuits in use drop below a certain percentage. The Cisco IOS commands used for this configuration are:

- On the Cisco Unified Border Element:

    ip circuit *max-calls max-call-number*

    The above command specifies an aggregated session capacity of the Cisco Unified Border Element, in terms of call legs. The default value is 1000 reserved call legs. Any call being handled by the Cisco Unified Border Element costs two sessions from the available IP circuits, one session for the inbound call leg and another session for the outbound call leg. Like a regular H.323 gateway, the Cisco Unified Border Element will automatically send its session capacity information to the via-zone gatekeeper using the H.323 version 4 protocol.

- On the gatekeeper:

    endpoint resource-threshold onset *onset-threshold* **abatement** *abatement-threshold*

    The above command makes the via-zone gatekeeper monitor the call volume in each of its gateways, including the Cisco Unified Border Element. The via-zone gatekeeper does the active call counting when the gateway reports its session capacity information in an admission request (ARQ) or disengaged request (DRQ) message. If the active call capacity usage in a particular gateway is above the high-water mark (Range = 1 to 99; Default = 90), the via-zone gatekeeper will stop sending calls to that gateway. If the gateway's active call volume falls below the low-water mark (Range = 1 to 99; Default = 70), the via-zone gatekeeper will resume sending calls to the gateway. These thresholds are global values and affect all gateways registered with a given gatekeeper.

With both of the above commands configured, the via-zone gatekeeper is able to calculate the current session capacity utilization on the Cisco Unified Border Element, which can prevent the via-zone gatekeeper from sending a call to the Cisco Unified Border Element without enough capacity resources. Otherwise, the gatekeeper call admission control would fail, with an admission reject (ARJ) or location reject (LRJ) message being returned to the originating device.

For more details on these commands, refer to the Cisco IOS command reference documentation available at

http://www.cisco.com

Configuration Guidelines

This section presents a simple configuration example based on the network diagram shown in Figure 9-20. This section is not intended as an exhaustive command reference guide, but rather as a collection of guidelines that can prove helpful in the most common deployment scenarios. Complete information on how to configure Cisco Unified Border Elements and via-zone gatekeepers is provided in the online documentation for the Cisco Unified Border Element, available at

http://www.cisco.com

Figure 9-20   *Configuration Example for Cisco Unified Border Element with Via-Zone Gatekeeper*
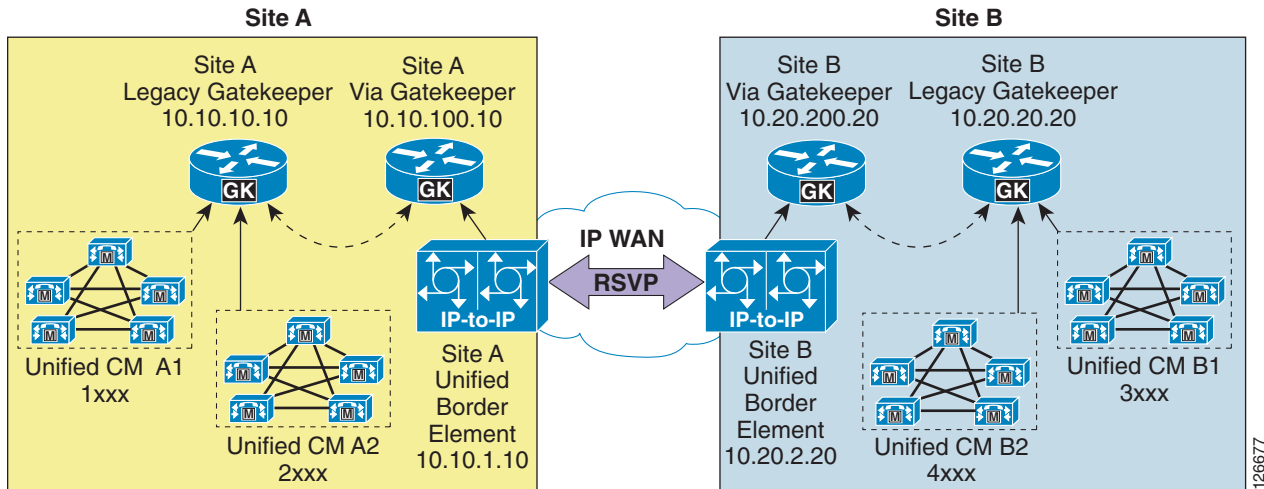
For the network shown in Figure 9-20, assume that there are two Unified CM clusters at Site A: cluster A1, with phone extensions 1*xxx*, and cluster A2, with phone extensions 2*xxx*. There are also two Unified CM clusters at Site B: cluster B1, with phone extensions 3*xxx*, and cluster B2, with phone extensions 4*xxx*.

The following subsections show the relevant configurations for devices located at Site A, so that calls within Site A are routed directly between Unified CM clusters (using Site A's legacy gatekeeper) while calls to Site B are routed through the two Cisco Unified Border Elements (using the respective legacy gatekeepers and via-zone gatekeepers).

Unified CM

Both cluster A1 and cluster A2 use a gatekeeper-controlled intercluster trunk, which is an intercluster trunk (ICT) without MTP required and which points to Site A's legacy gatekeeper.

A [34]XXX route pattern points to the ICT through a route list and route group construct, so as to reach Site B's clusters through the gatekeepers and the Cisco Unified Border Elements.

Another route pattern (2XXX for cluster A1 and 1XXX for cluster A2) allows cluster A1 and A2 to communicate with each other through the gatekeeper, by pointing to the ICT through a route list and route group construct.

To support supplementary services with a Cisco Unified Border Element and RSVP between clusters, the ICT must have the parameter "Calling Party Selection" set to "Last Redirect Number" or "First Redirect Number" for calls to complete.

Legacy Gatekeeper

The legacy gatekeeper at Site A routes calls between clusters A1 and A2 directly, while it sends all calls for Site B (extensions 3*xxx* and 4*xxx*) to Site A's via-zone gatekeeper. Example 9-1 shows the relevant configuration.

Example 9-1   *Legacy Gatekeeper Configuration for Site A*

```
gatekeeper
 zone local CCM-A1 customer.com 10.10.10.10
 zone local CCM-A2 customer.com
```

```
zone remote A-VIAGK customer.com 10.10.100.10
zone prefix CCM-A1 1...
zone prefix CCM-A2 2...
zone prefix A-VIAGK 3...
zone prefix A-VIAGK 4...
gw-type-prefix 1#* default-technology
arq reject-unknown-prefix
no shutdown
```

Via-zone Gatekeeper

The via-zone gatekeeper at Site A sends calls directed to Site B's Unified CM clusters (extensions 3*xxx* and 4*xxx*) to Site B's via-zone gatekeeper and invokes a Cisco Unified Border Element for calls to or from Site B. Calls directed to Site A's clusters are routed to Site A's legacy gatekeeper without invoking a Cisco Unified Border Element. Example 9-2 shows the relevant configuration.

Example 9-2   Via-Zone Gatekeeper Configuration for Site A

```
gatekeeper
 zone local A-VIAGK customer.com 10.10.100.10
 zone remote CCM-A1 customer.com 10.10.10.10
 zone remote CCM-A2 customer.com 10.10.10.10
 zone remote B-VIAGK customer.com 10.20.200.20 invia A-VIAGK outvia A-VIAGK
 zone prefix B-VIAGK 3...
 zone prefix B-VIAGK 4...
 zone prefix CCM-A1 1...
 zone prefix CCM-A2 2...
 gw-type-prefix 1#* default-technology
 arq reject-unknown-prefix
 no shutdown
```

The following considerations apply to the configuration shown in Example 9-2:

- The **invia** and **outvia** keywords in the command line related to the B-VIAGK remote zone trigger the via-zone gatekeeper behavior for that zone. This means that, for all calls destined to or coming from the B-VIAGK remote zone, the via-zone gatekeeper will invoke a Cisco Unified Border Element resource registered in the A-VIAGK local zone.

- The absence of **invia** and **outvia** keywords in the command line related to the CCM-A1 and CCM-A2 remote zones means that the standard gatekeeper behavior is applied, and no Cisco Unified Border Element is invoked for calls to or from these zones.

Cisco Unified Border Element

The Cisco Unified Border Element at Site A requests RSVP reservations for voice and video calls directed toward Site B's Unified CM clusters (extensions 3*xxx* and 4*xxx*) but not for calls directed toward Site A's Unified CM clusters (extensions 1*xxx* and 2*xxx*). Example 9-3 shows the relevant configuration.

Example 9-3   Cisco Unified Border Element Configuration for Site A

```
voice service voip
   allow-connections h323 to h323
   h323
      emptycapability
      h245 passthru tcsnonstd-passthru
!
gateway
!
```

```
interface FastEthernet0/1
   ip address 10.10.1.10 255.255.255.0
   ip rsvp bandwidth 200
   ip rsvp data-packet classification none
   ip rsvp resource-provider none
   h323-gateway voip interface
   h323-gateway voip id A-VIAGK ipaddr 10.10.100.10
   h323-gateway voip h323-id A-CUBE
   h323-gateway voip bind srcaddr 10.10.1.10
   h323-gateway voip tech-prefix 1#
!
dial-peer voice 5 voip
   session target ras
   incoming called-number [3-4]...
   codec g729r8
!
dial-peer voice 10 voip
   destination-pattern [3-4]...
   session target ras
   req-qos guaranteed-delay audio
   req-qos guaranteed-delay video
   acc-qos guaranteed-delay audio
   acc-qos guaranteed-delay video
   codec g729r8
!
dial-peer voice 15 voip
session target ras
   incoming called-number [1-2]...
   req-qos guaranteed-delay audio
   req-qos guaranteed-delay video
   acc-qos guaranteed-delay audio
   acc-qos guaranteed-delay video
   codec g729r8
!
dial-peer voice 20 voip
   destination-pattern [1-2]...
   session target ras
   codec g729r8
```

The following considerations apply to the configuration shown in Example 9-3:

- The **emptycapability** command enables the H.245 Empty Capabilities Set (ECS) between Unified CM and the Cisco Unified Border Element to invoke supplementary services for the established calls.

- The **req-qos guaranteed-delay** [**audio** | **video**] commands specify that the Cisco Unified Border Element should request a guaranteed-delay RSVP reservation for voice and video calls that use dial-peer 10 or 15.

- The **acc-qos guaranteed-delay** [**audio** | **video**] commands specify that the minimum acceptable QoS level for voice and video calls is also a guaranteed-delay RSVP reservation. This means that, if the RSVP request fails, the call will also fail, which equates to making the RSVP reservation mandatory. To configure the Cisco Unified Border Element so that the RSVP reservation is optional (so that the call succeeds even if the reservation fails), use the commands **acc-qos best-effort** [**audio** | **video**] instead.

Call Admission Control Design

This section describes how to apply the call admission control mechanisms to the various Unified CM deployment models and to the following IP WAN topologies:
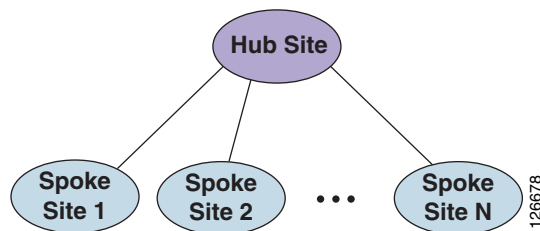
- Simple Hub-and-Spoke Topologies, page 9-37
- Two-Tier Hub-and-Spoke Topologies, page 9-41
- Simple MPLS Topologies, page 9-45
- Generic Topologies, page 9-52

For each topology, these sections present different sets of design considerations based on the Unified CM deployment model adopted.

Simple Hub-and-Spoke Topologies

Figure 9-21 depicts a simple hub-and-spoke topology, also known as a star topology. In this type of network topology, all sites (called *spoke sites*) are connected via a single IP WAN link to a central site (called the *hub site*). There are no direct links between the spoke sites, and every communication between them must transit through the hub site.

Figure 9-21   A Simple Hub-and-Spoke Topology

The design considerations in this section apply to simple hub-and-spoke topologies that use traditional Layer 2 IP WAN technologies such as:

- Frame Relay
- ATM
- Frame Relay/ATM Service Interworking
- Leased Lines

For IP WAN deployments based on the MPLS technology, refer to the section on Simple MPLS Topologies, page 9-45.

The remainder of this section contains design best practices for simple hub-and-spoke topologies according to the Unified CM deployment model adopted:

- Centralized Unified CM Deployments, page 9-38

  One or more Unified CM clusters are located at the hub site, but only phones and gateways are located at the spoke sites.

- Distributed Unified CM Deployments, page 9-39

  A Unified CM cluster or Cisco Unified Communications Manager Express (Unified CME) is located at each site.

Centralized Unified CM Deployments

In multisite WAN deployments with centralized call processing in a simple hub-and-spoke topology, use Unified CM static *locations* for implementing call admission control. Figure 9-22 shows an example of how to apply this mechanism to such a topology.

Figure 9-22   Call Admission Control for Simple Hub-and-Spoke Topologies Using Static Locations
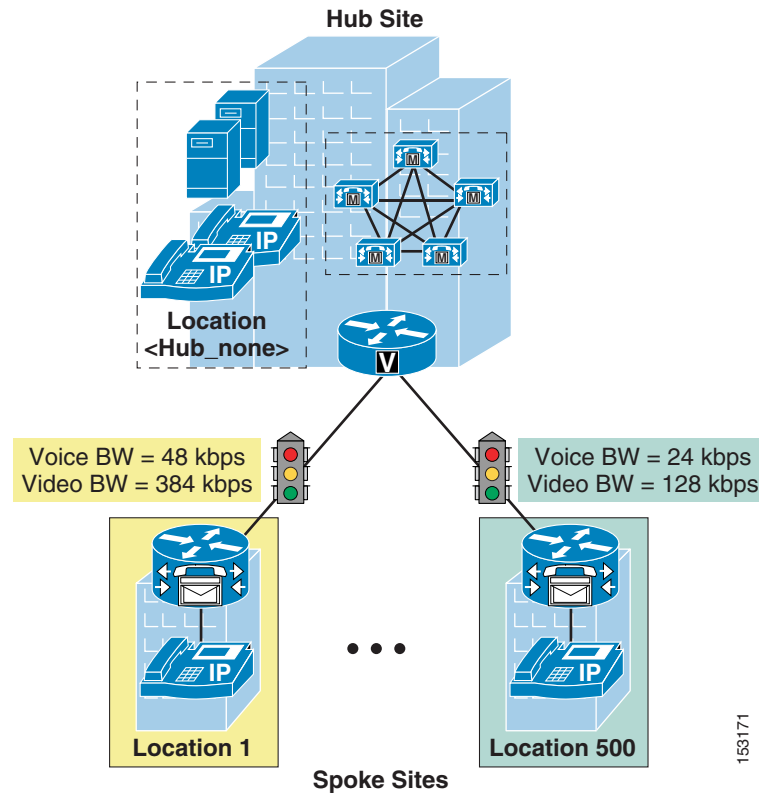
Follow these guidelines when using static locations for call admission control:

- Configure a separate location in Unified CM for each spoke site.
- Configure the appropriate bandwidth limits for voice and video calls for each site according to the types of codecs used at that site. (See Table 9-2 for recommended bandwidth settings.)
- Assign all devices at each spoke site to the appropriate location.
- Leave devices at the hub site in the Hub_None location.
- If you move a device to another location, change its location configuration as well.
- Unified CM supports up to 2000 locations.
- If you require automatic rerouting over the PSTN when the WAN bandwidth is not sufficient, configure the automated alternate routing (AAR) feature on Unified CM. (See Automated Alternate Routing, page 10-86.)
- If multiple Unified CM clusters are located at the same hub site, leave the intercluster trunk devices in the Hub_None location. You may use a gatekeeper for dial plan resolution. However, gatekeeper call admission control is not necessary in this case because all IP WAN links are controlled by the locations algorithm.

Note   If one or more sites have dual connectivity to the IP WAN and you want to take full advantage of the bandwidth available on both links, Cisco recommends that you deploy topology-aware call admission control, as described in the section on Generic Topologies, page 9-52. See also Limitations of Topology-Unaware Call Admission Control, page 9-5, for more information.

Distributed Unified CM Deployments

For distributed call processing deployments on a simple hub-and-spoke topology, you can implement call admission control with a Cisco IOS gatekeeper. In this design, the call processing agent (which could be a Unified CM cluster, Cisco Unified Communications Manager Express (Unified CME), or an H.323 gateway) registers with the Cisco IOS gatekeeper and queries it each time the agent wants to place an IP WAN call. The Cisco IOS gatekeeper associates each call processing agent with a zone that has specific bandwidth limitations. Thus, the Cisco IOS gatekeeper can limit the maximum amount of bandwidth consumed by IP WAN voice calls into or out of a zone.

Figure 9-23 illustrates call admission control with a gatekeeper. In brief, when the call processing agent wants to place an IP WAN call, it first requests permission from the gatekeeper. If the gatekeeper grants permission, the call processing agent places the call across the IP WAN. If the gatekeeper denies the request, the call processing agent can try a secondary path (the PSTN, for example) or can simply fail the call.

Figure 9-23   Call Admission Control for Hub-and-Spoke Topologies Using a Gatekeeper
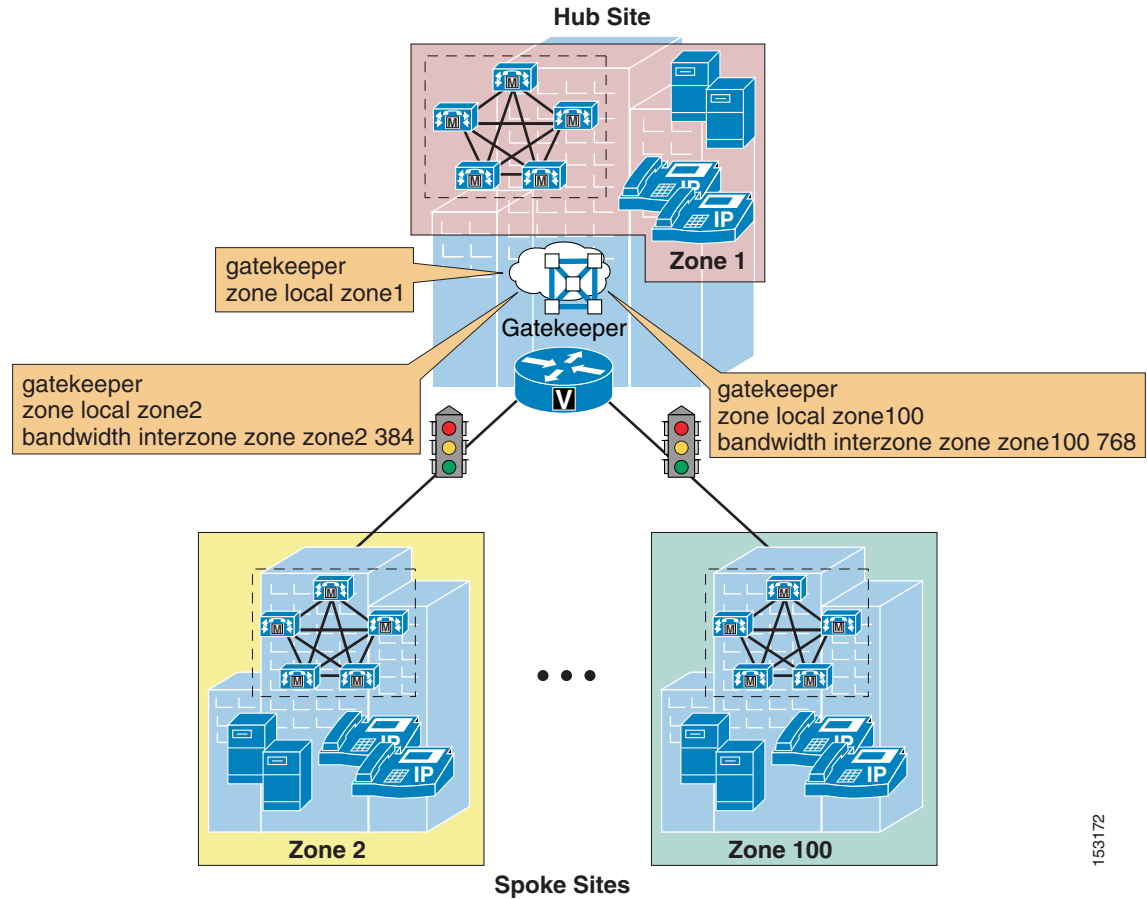
Follow these guidelines when deploying call admission control with a gatekeeper:

- In Unified CM, configure an H.225 gatekeeper-controlled trunk if you have a mixed environment with Cisco Unified Communications Manager Express (Unified CME) and H.323 gateways.

- In Unified CM, configure an intercluster gatekeeper-controlled trunk if you have an environment exclusively based on Unified CM clusters.

- Ensure that the zone configured in Unified CM matches the correct gatekeeper zone for the site.

- Each Unified CM subscriber listed in the device pool's Unified CM Redundancy Group registers a gatekeeper-controlled trunk with the gatekeeper. (Maximum of three.)

- Calls are load-balanced across the registered trunks in the Unified CM cluster.

- Unified CM supports multiple gatekeepers and trunks.

- You can place the trunk in a route group and route list construct to provide automatic PSTN failover. (See Dial Plan, page 10-1, for more details.)

- Configure a separate zone in the gatekeeper for each site supporting Unified CMs, Unified CME, or H.323 gateways.

- Use the **bandwidth interzone** command on the gatekeeper to control bandwidth between Unified CM clusters, Unified CME servers, and H.323 devices registered directly with the gatekeeper. (See Table 9-4 for bandwidth settings by codec type.)

- A single Cisco IOS gatekeeper can support up to 100 zones or sites.

- You can provide gatekeeper redundancy by using gatekeeper clustering (alternate gatekeeper) or Cisco Hot Standby Router Protocol (HSRP). Use HSRP only if gatekeeper clustering is not available in your software feature set.

Note   If one or more sites have dual connectivity to the IP WAN and you want to take full advantage of the bandwidth available on both links, Cisco recommends that you deploy topology-aware call admission control, as described in the section on Generic Topologies, page 9-52. See also Limitations of Topology-Unaware Call Admission Control, page 9-5, for more information.

Two-Tier Hub-and-Spoke Topologies

Figure 9-24 depicts a two-tier hub-and-spoke topology. This type of network topology consists of sites at three hierarchical levels: the first-tier hub site, the second-tier hub sites, and the spoke sites. A group of spoke sites are connected to a single second-tier hub site, and each second-tier hub site is in turn connected to the single first-tier hub site. As in the simple hub-and-spoke topology, there are no direct links between the spoke sites, and every communications between them must transit through the second-tier hub site. Similarly, there are no direct links between the second-tier hub sites, and all communications between them must transit through the first-tier hub site.

Figure 9-24   A Two-Tier Hub-and-Spoke Topology
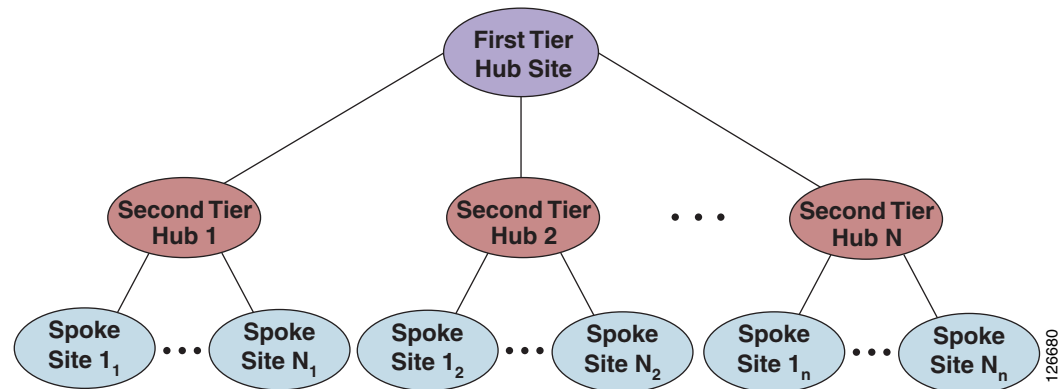
The design considerations in this section apply to two-tier hub-and-spoke topologies that use traditional Layer 2 IP WAN technologies such as:

- Frame Relay

- ATM

- Frame Relay/ATM Service Interworking

- Leased Lines

For IP WAN deployments based on the MPLS technology, refer to the section on Simple MPLS Topologies, page 9-45.

The remainder of this section contains design best practices for two-tier hub-and-spoke topologies according to the Unified CM deployment model adopted:

- Centralized Unified CM Deployments, page 9-42

   One or more Unified CM clusters are located at the first-tier hub site, but only phones and gateways are located at the second-tier hub sites and the spoke sites.

- Distributed Unified CM Deployments, page 9-44

   Unified CM clusters are located at the first-tier hub site and at the second-tier hub sites, while only endpoints and gateways are located at the spoke sites.

Centralized Unified CM Deployments

Figure 9-25 depicts a single centralized Unified CM cluster deployed in a two-tier hub-and-spoke IP WAN topology. In this scenario, the Unified CM cluster is located at the first-tier hub site, while all second-tier hub and spoke sites contain only endpoints and gateways.

Figure 9-25   Two-Tier Hub-and-Spoke Topology with Centralized Unified CM
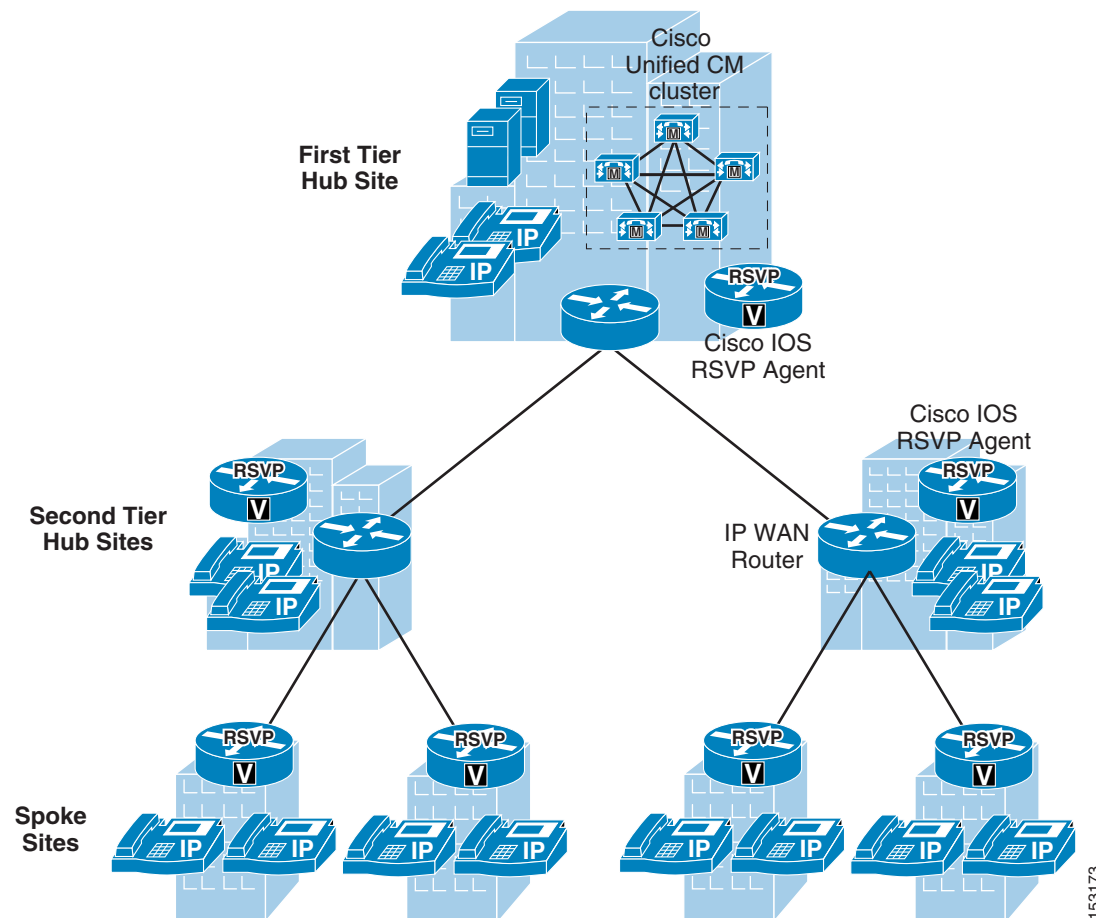
This scenario requires that you deploy topology-aware call admission control, which for a single Unified CM cluster means using RSVP-enabled locations. Figure 9-26 shows how this mechanism can be deployed.

Figure 9-26   *Call Admission Control for Two-Tier Hub-and-Spoke Topologies Using Locations with RSVP*
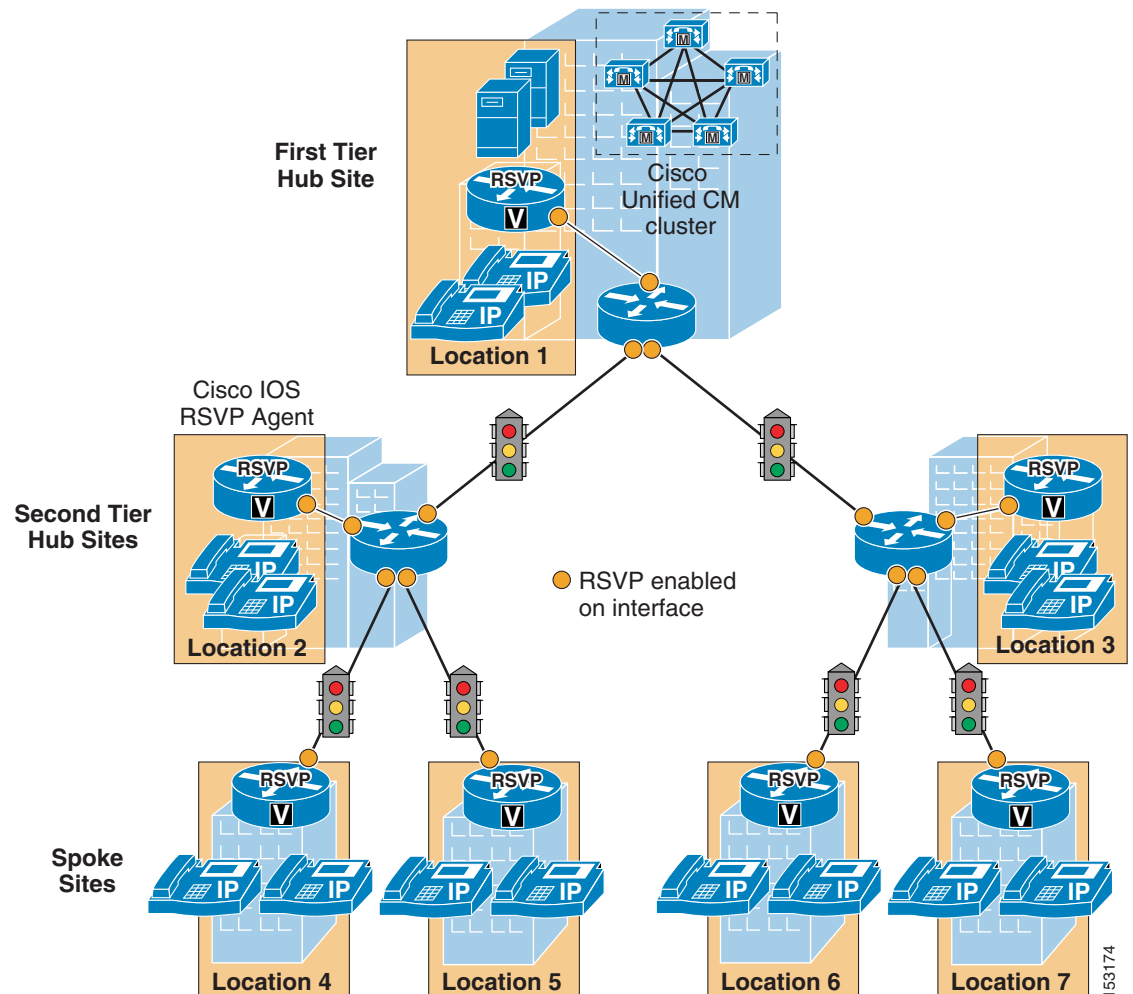
The following guidelines apply to these deployments:

- Enable the Cisco IOS RSVP Agent feature on a Cisco IOS router at each site. 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 Agent itself).

- 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.

- 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 (as illustrated in Figure 9-26).

Distributed Unified CM Deployments

To provide call admission control in deployments that use a two-tier hub-and-spoke topology, with Unified CMs at the first-tier and second-tier hub sites, you can combine the static locations and gatekeeper zone mechanisms as illustrated in Figure 9-27.

Figure 9-27   Combining the Locations and Gatekeeper Mechanisms for Call Admission Control
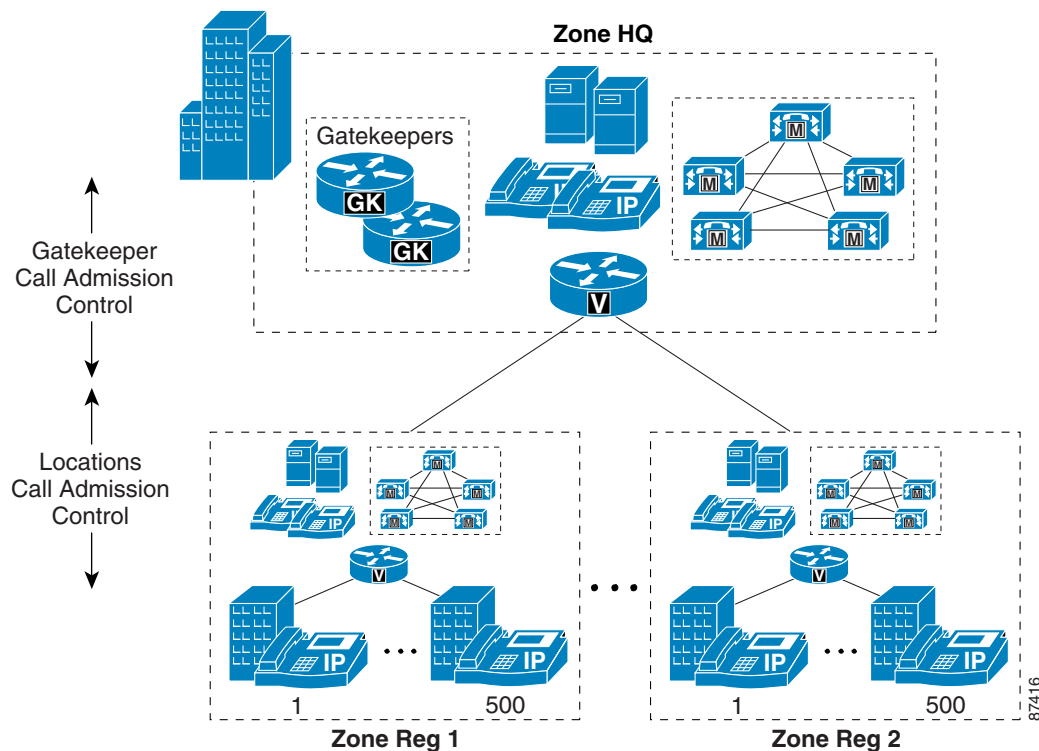
Follow these recommendations when combining gatekeeper zones with static locations for call admission control:

- Use call admission control based on static locations for sites with no local Unified CM (that is, the spoke sites).

- Use gatekeeper-based call admission control between Unified CM clusters (that is, between the first-tier hub site and the second-tier hub sites).

- For each site without a local Unified CM, configure a location for that site in the Unified CM cluster supporting the site.

- Configure the appropriate bandwidth limits for voice and video calls at each site according to the type of codec used at that site. (See Table 9-2 and Table 9-4 for bandwidth settings.)

- Assign each device configured in Unified CM to a location. If you move a device to another location, change its location configuration as well.

- Unified CM supports up to 2000 locations.

- Each Unified CM cluster registers a gatekeeper-controlled trunk with the gatekeeper.

- On the gatekeeper, configure a zone for each Unified CM cluster, and use the **bandwidth interzone** command to control the number of calls to and from each cluster.

> **Note**   If one or more sites have dual connectivity to the IP WAN and you want to take full advantage of the bandwidth available on both links, Cisco recommends that you deploy topology-aware call admission control, as described in the section on Generic Topologies, page 9-52. See also Limitations of Topology-Unaware Call Admission Control, page 9-5, for more information.

Simple MPLS Topologies

Figure 9-28 shows an IP WAN (from a service provider) based on the Multiprotocol Label Switching (MPLS) technology. The main design difference between traditional Layer 2 WAN services offered by service providers and services based on MPLS is that, with MPLS, the IP WAN topology does not conform to a hub-and-spoke but instead provides "full-mesh" connectivity between all sites.

This topology difference means that, from an IP routing perspective on the enterprise side of the network, each site is one IP hop away from all other sites. Thus, there is no need to transit through a hub site to reach any other site. In fact, there is no concept of a "hub site." All sites are considered equal, and the only difference between them is the amount of bandwidth that they are allowed to use across the IP WAN.

*Figure 9-28   MPLS IP WAN from a Service Provider, and Its Topology Equivalent*
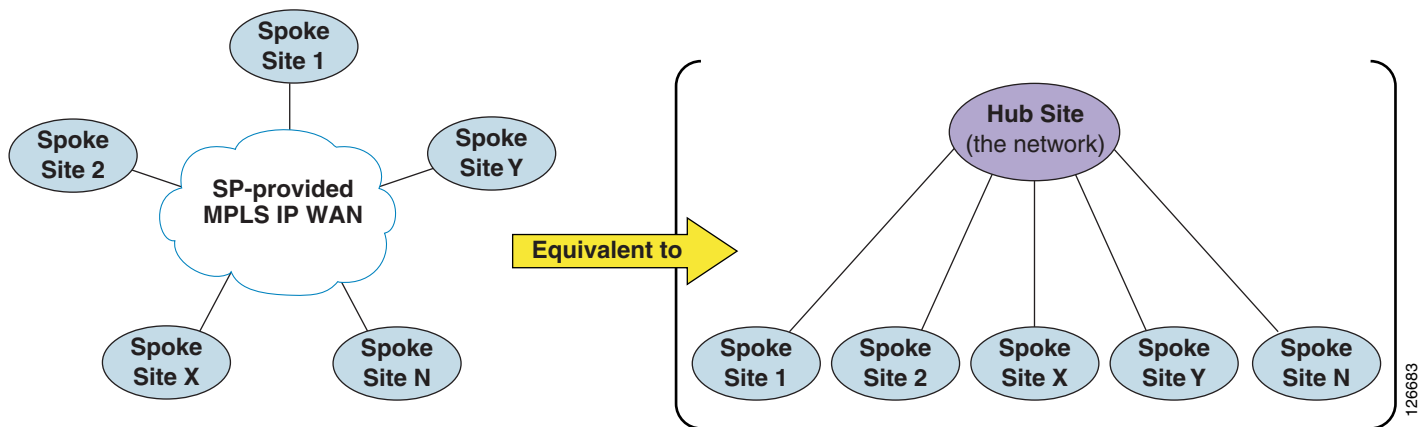
Based on these considerations, it is easy to see that, from a call admission control perspective, a service-provider IP WAN service based on MPLS is in reality equivalent to a hub-and-spoke topology without a hub site. (See Figure 9-28.) In fact, the network itself could be considered as the hub site, while all the enterprise sites (including the headquarters, or central site) are equivalent to spoke sites. These differences have implications on how to perform call admission control, which are described in the remainder of this section.

An exception to the above considerations that is worth mentioning here is represented by multisite deployments where an MPLS-based WAN co-exists with an IP WAN based on a traditional Layer 2 technology, such as Frame Relay or ATM. Such a scenario could occur, for example, in the case of network migration phases, company mergers, or various other situations.

As shown in Figure 9-29, integrating a hub-and-spoke IP WAN based on a traditional Layer 2 technology (such as Frame Relay) with an MPLS-based IP WAN results in a network topology that is neither a simple hub-and-spoke nor a full-mesh, but rather is equivalent to a two-tier hub-and-spoke.

In this case the first-tier hub site is represented by the MPLS network, the second-tier hub sites are represented by the MPLS-based sites as well as the MPLS-enabled Frame Relay hub site, and the spoke sites are represented by the Frame Relay spoke sites. Therefore, for design considerations on such deployments, refer to the section on Two-Tier Hub-and-Spoke Topologies, page 9-41.

Figure 9-29   Co-existence of MPLS Sites and Frame Relay Sites, and the Topology Equivalent
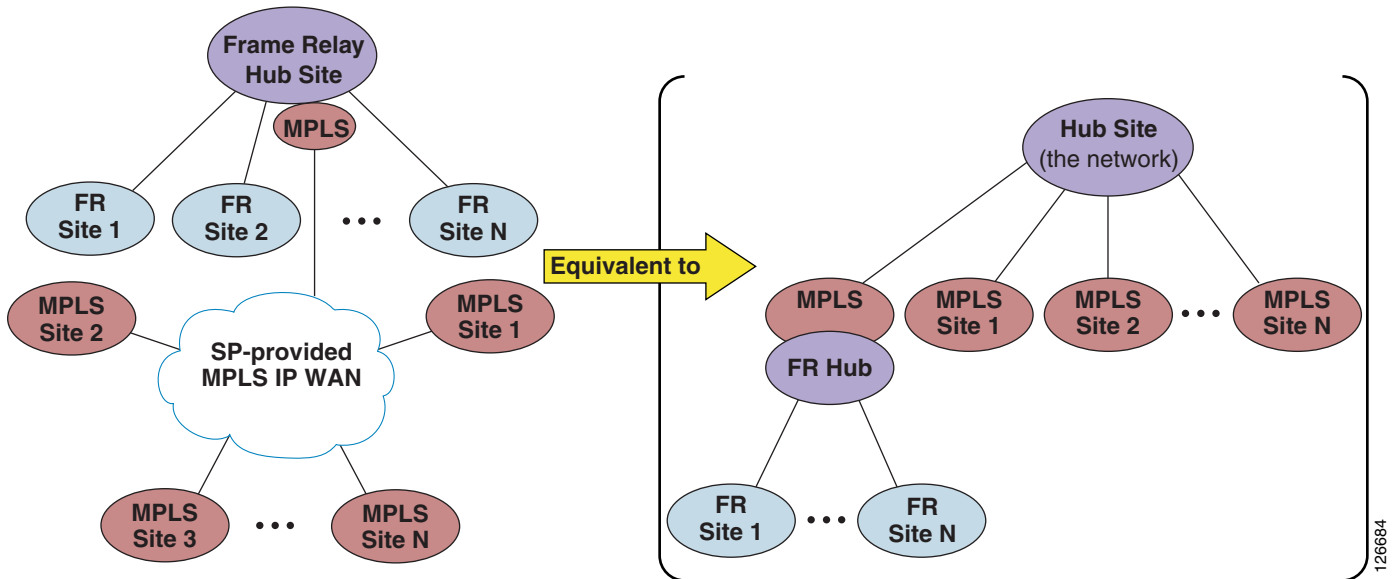
The remainder of this section contains design best practices for MPLS-based topologies according to the Unified CM deployment model adopted:

- Centralized Unified CM Deployments, page 9-47

    One or more Unified CM clusters are located at only one site, while only endpoints and gateways are located at all other sites.

- Distributed Unified CM Deployments, page 9-50

    Unified CM clusters are located at multiple sites, while endpoints and gateways are located at all other sites.

Note   This section focuses on enterprise deployments where the MPLS WAN service is provided by a service provider. In cases where the MPLS network is deployed by the enterprise itself, call admission control can be performed effectively if one of the following two conditions is satisfied: (1) routing in the MPLS network is configured so that it is equivalent to a hub-and-spoke, or (2) bandwidth in the core of the MPLS network is heavily over-provisioned so that congestion can occur only at the edge.

> **Note**   If one or more sites have dual connectivity to the IP WAN and you want to take full advantage of the bandwidth available on both links, Cisco recommends that you deploy topology-aware call admission control, as described in the section on Generic Topologies, page 9-52. Particular care must be taken to guarantee symmetric routing in the presence of load-balanced links. See also Limitations of Topology-Unaware Call Admission Control, page 9-5, and Special Considerations for MPLS Networks, page 9-11, for more information, and contact your local Cisco account team for further assistance.

Centralized Unified CM Deployments

In multisite WAN deployments with centralized call processing in an MPLS topology, use Unified CM static *locations* for implementing call admission control.

In a hub-and-spoke WAN topology (for example, Frame Relay or ATM), each link to and from a branch site terminates at the central site. For example, in a Frame Relay network, all permanent virtual circuits (PVCs) from the branch routers are aggregated at the central site's head-end router. In such a scenario, there is no need to apply call admission control to devices at the central site because the bandwidth accounting occurs at the branch ends of the WAN links. Therefore, within the Unified CM locations configuration, devices at the central site are left in the Hub_None location, while devices at each branch are placed in their appropriate location to ensure proper call admission control.

With an MPLS WAN network, all branches are deemed to be adjacent at Layer 3, thus they do not have to rely on the central site for connectivity. Figure 9-30 illustrates a spoke-to-spoke call between two branch sites in this type of deployment.

Figure 9-30   *Spoke-to-Spoke Calls in an MPLS Deployment*
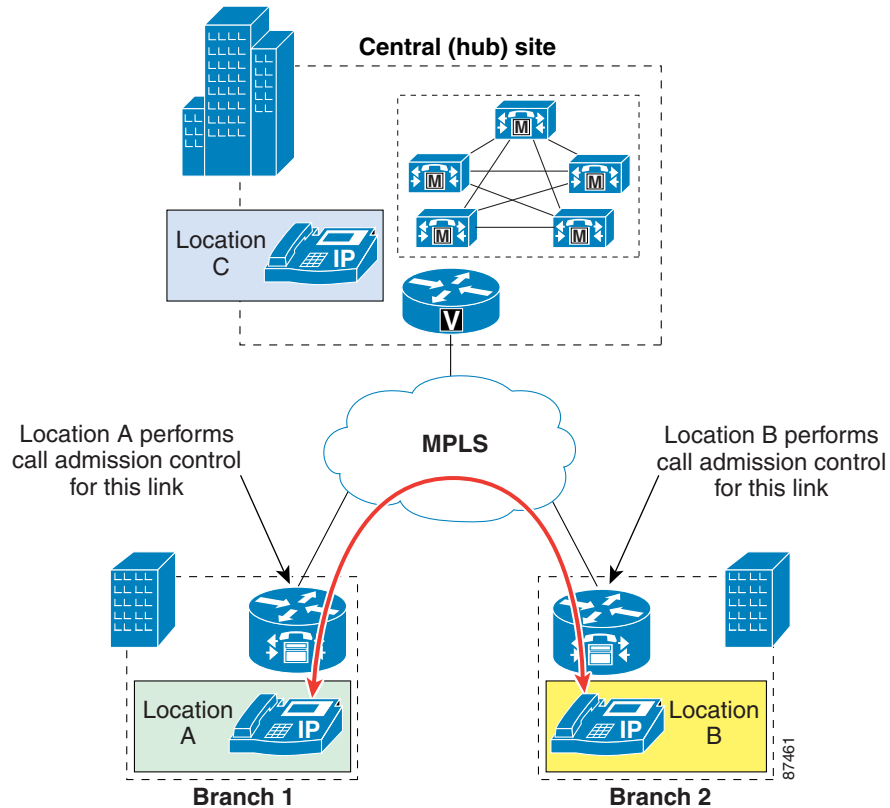
Also, in an MPLS WAN, the link connecting the central site to the WAN does not aggregate every branch's WAN link. By placing all the central site devices in their own call admission control location (that is, not in the Hub_None location), this configuration requires that call admission control be performed on the central site link independently of the branch links. (See Figure 9-31.)

Note   Some devices such as trunks do not terminate media and are normally left in the Hub_None location. However, to avoid errors in call admission control when an MTP is required on a trunk, the trunk must be assigned to a location other than Hub_None and any MTP in the trunk's MRGL must be physically located at the site associated with that location. This configuration is required because an MTP cannot be assigned a location directly, so it inherits the location of the device that selected it.

Figure 9-31   Calls to and from the Hub in an MPLS Deployment
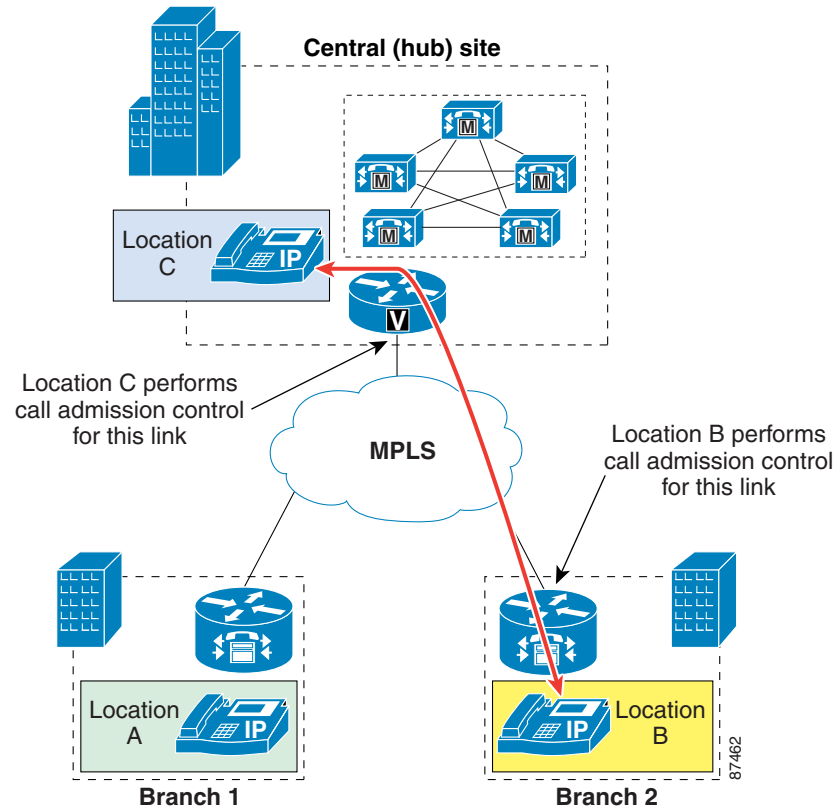
When all the available bandwidth for a particular site has been utilized, you can provide automatic failover to the PSTN by using the automated alternate routing (AAR) feature within Unified CM. (For more information on AAR, see Automated Alternate Routing, page 10-86.)
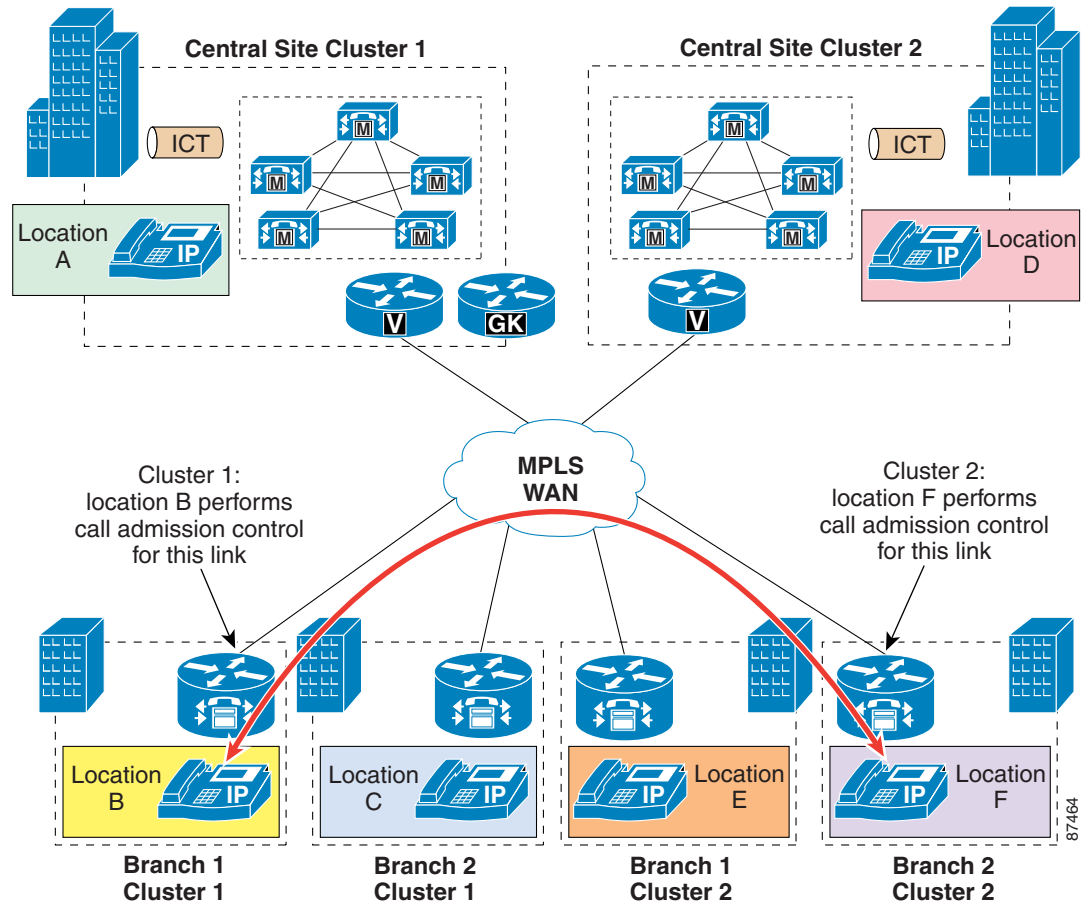
Distributed Unified CM Deployments

In multisite deployments where a Unified CM cluster is present at more than one site without any branch locations and the sites are linked through an MPLS WAN, a gatekeeper can provide dial-plan resolution as well as call admission control between the sites, with each site being placed in a different gatekeeper zone. This is the same mechanism adopted for hub-and-spoke topologies based on Layer 2 WAN technologies. (See Figure 9-32.)

Figure 9-32   Gatekeeper Call Admission Control in a Distributed Deployment with MPLS
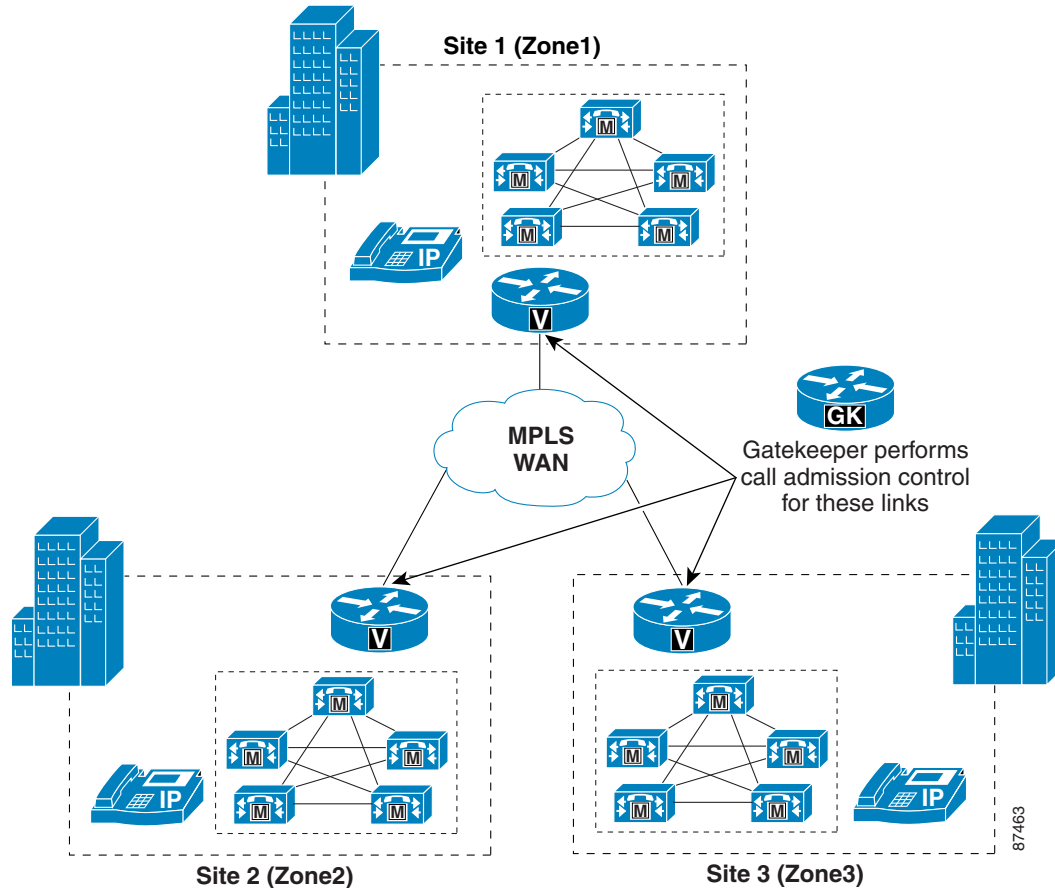
In deployments where branch sites are required, a gatekeeper can be used for dial-plan resolution between clusters, but a gatekeeper is not recommended for call admission control.

When calls occur between branches belonging to different clusters, the audio path is established between the two branches directly, with no media transiting through each cluster's central site. Therefore, call admission control is required only on the WAN links at the two branches. (See Figure 9-33.)

Figure 9-33   Multiple Clusters Connected by Intercluster Trunks (ICTs)

As in the centralized Unified CM deployments, devices that terminate media at each site (including the central sites for each cluster) must be placed in an appropriately configured location.

Note that the intercluster trunks are purely signaling devices, and there is no media transiting through them. Therefore, all intercluster trunks must be left in location Hub_None. The exception is when the trunk requires an MTP, in which case the trunk and MTP should both be in the location of the site in which they reside.

When all the available bandwidth for a particular site has been used, you can provide automatic failover to the PSTN by using a combination of the following two methods:

- The route list and route group construct for calls across multiple Unified CM clusters

- The automated alternate routing (AAR) feature for calls within a Unified CM cluster (For more information on AAR, see Automated Alternate Routing, page 10-86.)

Generic Topologies

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

As Figure 9-34 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 9-34   A Generic Topology
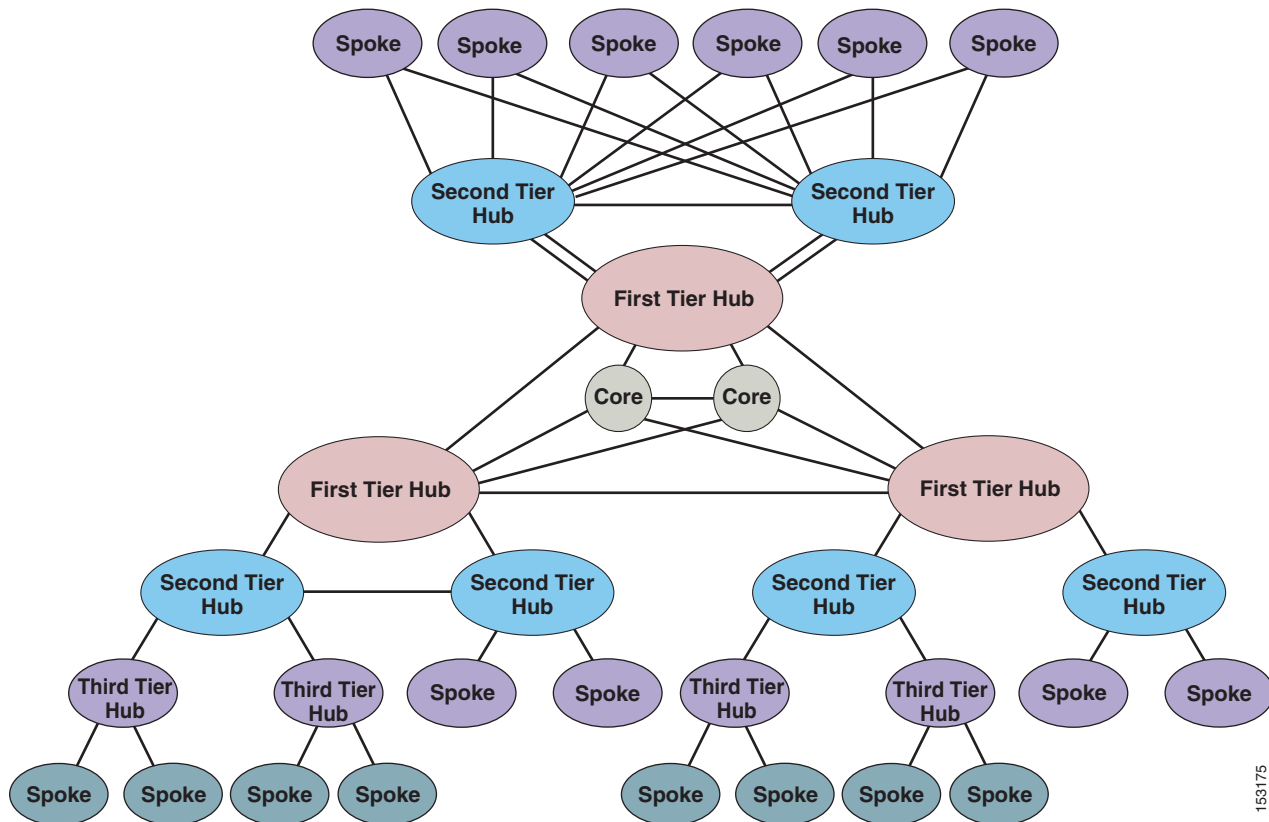
The complex nature of these networks requires the adoption of topology-aware call admission control mechanisms based on RSVP. In particular, these mechanisms can properly control bandwidth in presence of any of the following topology aspects:

- 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 9-53

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

- Distributed Unified CM Deployments, page 9-56

  Unified CM clusters are located at multiple sites, and endpoints and gateways are located at all other sites.

Centralized Unified CM Deployments

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

- Single Unified CM Cluster, page 9-53
- Co-Located Unified CM Clusters, page 9-54

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 9-35.

*Figure 9-35   A Single Unified CM Cluster in a Generic Topology*
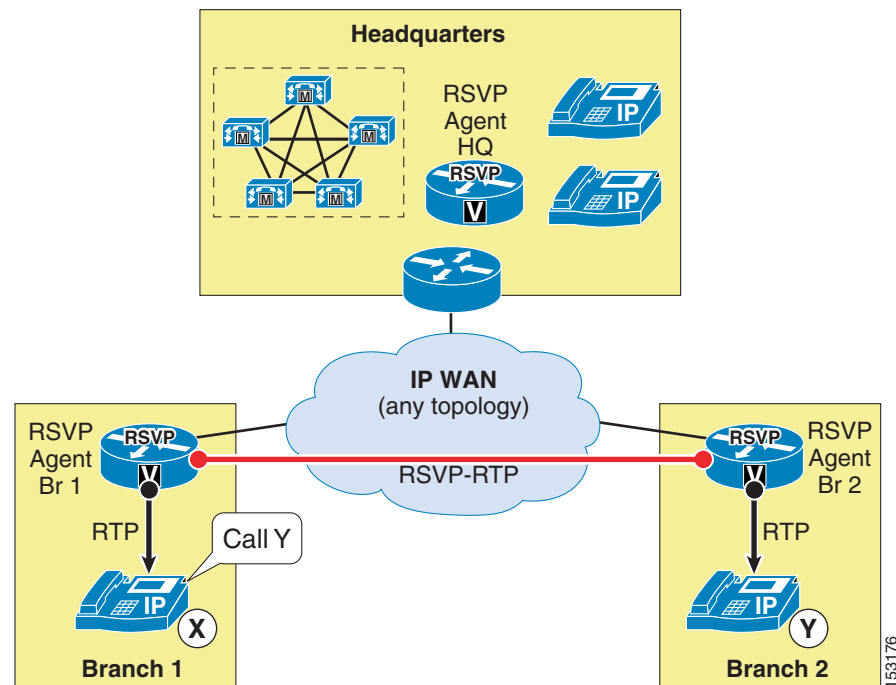
The following guidelines apply to this type of deployment:

- 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 3-58.)

- 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 high-speed link, provided that no congestion occurs in the priority queue of the link and that bandwidth for voice and video can therefore be considered unlimited.

Figure 9-36 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 9-36   *Co-located Unified CM Clusters in a Generic Topology*
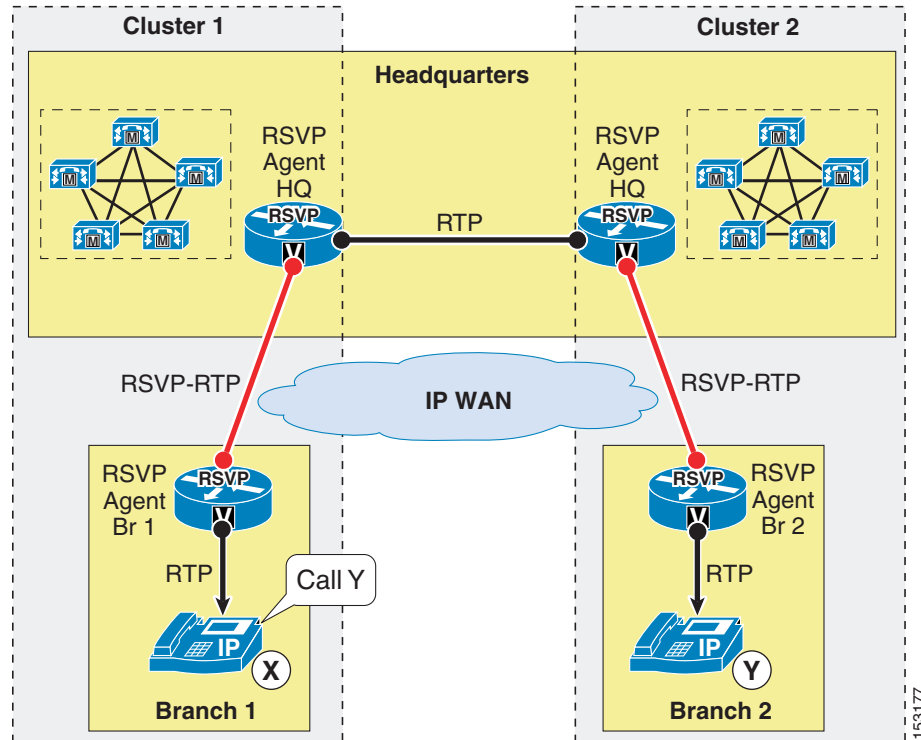
In addition to the guidelines listed in Single Unified CM Cluster, page 9-53, observe the following best practices for this type of deployment:

- For each cluster, define an intercluster trunk to enable communications to the other clusters. A gatekeeper may be used for dial plan resolution, but it is not needed for call admission control.

- Assign the intercluster trunk to the same location used for all devices located at the central site (HQ in the example of Figure 9-36).

- Ensure that the intercluster trunk is assigned to a device pool that specifies an MRGL that in turn points to an MRG containing the Cisco RSVP Agent located at the central site (Cisco RSVP Agent HQ1 in the case of Cluster 1 in Figure 9-36).

- Use the AAR feature to provide automatic PSTN failover in case of call admission control failure within a cluster.

- Use the route list and route group constructs to provide automatic PSTN failover in case of call admission control failure across clusters.

- Both media and signaling traffic are hair-pinned via the central site for calls between two branch sites belonging to different clusters (as shown in Figure 9-36, where the call between phone X in Branch 1 and phone Y in Branch 2 is hair-pinned via the HQ site).

Distributed Unified CM Deployments

In order to provide call admission control for distributed Unified CM deployments in a generic network topology, two approaches are possible, depending on the number of Unified CM clusters involved:

- Remote Cisco RSVP Agent Approach, page 9-57

  This solution applies to deployments where three or fewer Unified CM clusters are located at different sites connected via an IP WAN with limited bandwidth.

- Cisco Unified Border Element Approach, page 9-60

  This solution applies to deployments where any number of Unified CM clusters are located at different sites connected via an IP WAN with limited bandwidth.
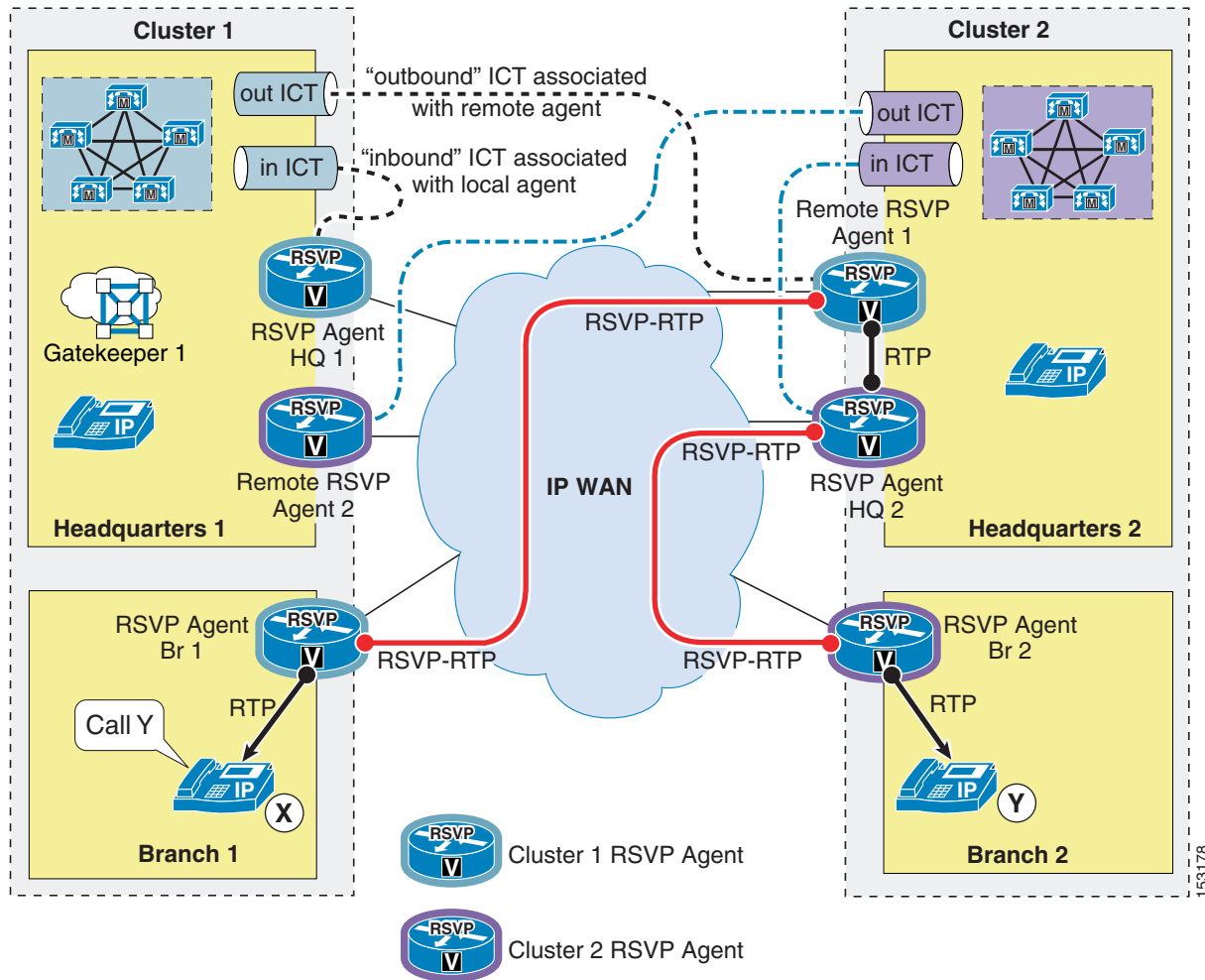
Note   In deployments where the Unified CM clusters are located at sites connected via a high-speed IP WAN link, it is possible to treat this scenario similar to that described in the section on Co-Located Unified CM Clusters, page 9-54, provided that congestion does not occur in the priority queue of the IP WAN link.

Remote Cisco RSVP Agent Approach

To provide call admission control in generic topologies where three or fewer Unified CM clusters are located at different sites, you can extend the concept of RSVP-enabled locations to cover intercluster calls by defining "remote" Cisco RSVP Agents, as shown in Figure 9-37.

Figure 9-37   Remote Cisco RSVP Agent Approach for Distributed Clusters in a Generic Topology

Note For simplicity, the descriptions in this section are based on an example with two Unified CM clusters, as shown in Figure 9-37. For deployments with three Unified CM clusters, some additional remarks are provided at the end of the section.

In addition to the guidelines listed in the section on , observe the following best practices for these deployments:

- For each cluster, define two intercluster trunks (ICTs) to enable communications to the other cluster: one "outbound" intercluster trunk and one "inbound" intercluster trunk.

- Configure a Cisco IOS gatekeeper for dial plan resolution (not for call admission control), defining one zone per Unified CM cluster. For example:

```
gatekeeper
  zone local cluster1 customer.com 10.10.10.10
  zone local cluster2 customer.com
```

- For each cluster, register the inbound trunk with the gatekeeper in the normal zone for that cluster. (For example, the inbound trunk of Cluster 1 registers in zone cluster1, while the inbound trunk of Cluster 2 registers with zone cluster2.)

- For each cluster, register the outbound trunk with the gatekeeper in specially created zones. For example:

```
gatekeeper
  zone local cluster1 customer.com 10.10.10.10
  zone local cluster2 customer.com
  zone local cluster1-to-cluster2 customer.com
  zone local cluster2-to-cluster1 customer.com
```

- Set up your Unified CM dial plan so that outbound calls to the other cluster use the outbound intercluster trunk. (For example, in Cluster 1, have a 2XXX route pattern that points to the outbound trunk via a route list and route group construct.)

- Set up your gatekeeper dial plan so that calls destined to a given cluster are routed to its inbound trunk. For example:

```
gatekeeper
  zone local cluster1 customer.com 10.10.10.10
  zone local cluster2 customer.com
  zone local cluster1-to-cluster2 customer.com
  zone local cluster2-to-cluster1 customer.com
  zone prefix cluster1 1...
  zone prefix cluster2 2...
```

- Assign the inbound trunk to the same location as all devices located at that site (for example, location HQ1 for Cluster 1's inbound trunk and location HQ2 for Cluster 2's inbound trunk).

- Assign the outbound trunk to a newly created location (for example, location remote-to-HQ2 for Cluster 1's outbound trunk toward Cluster 2, and location remote-to-HQ1 for Cluster 2's outbound trunk toward Cluster 1).

- At each of the two sites where Unified CM resides, have an instance of Cisco RSVP Agent registered with the local cluster (for example, Cisco RSVP Agent HQ1 at site HQ1 and Cisco RSVP Agent HQ2 at site HQ2).

- Assign these local Cisco RSVP Agents to the same location as all devices located at that site (for example, location HQ1 for Cisco RSVP Agent HQ1 and location HQ2 for Cisco RSVP Agent HQ2).

- For each cluster, assign the local Cisco RSVP Agent to an MRG contained in the MRGL of all devices located at the central site, including the MRGL used by the inbound trunk via its device pool.

- At each of the two sites where Unified CM clusters reside, add an instance of Cisco RSVP Agent registered with the other Unified CM cluster. (For example, Remote Cisco RSVP Agent 1 is registered with Cluster 1 and placed at site HQ2, where Cluster 2 resides, while Remote Cisco RSVP Agent 2 is registered with Cluster 2 and is placed at site HQ1, where Cluster 1 resides).

- Assign these remote Cisco RSVP Agents to the locations created for the outbound trunks (for example, location remote-to-HQ2 within Cluster 1 for Remote Cisco RSVP Agent 1 and location remote-to-HQ1 within Cluster 2 for Remote Cisco RSVP Agent 2).

- For each cluster, assign the remote Cisco RSVP Agent to an MRG contained in the MRGL used by the outbound trunk (via its device pool).

Note   While logically separate, the remote Cisco RSVP Agent instances may reside on the same router platform as the local Cisco RSVP Agent registered to the other cluster. For the example in Figure 9-37, this means that Remote Cisco RSVP Agent 1 and Cisco RSVP Agent HQ2 may actually be located on the same router platform, and the same applies to Remote Cisco RSVP Agent 2 and Cisco RSVP Agent HQ1.

In Figure 9-37, when phone X at Branch 1 places a call to phone Y at Branch 2, Unified CM Cluster 1 will route the call to the gatekeeper via the outbound trunk. Because phone X is assigned to location Branch 1 and the outbound trunk is associated to the location remote-to-HQ2, Unified CM Cluster 1 will initiate an RSVP reservation across the IP WAN between Cisco RSVP Agent Br 1 and Remote Cisco RSVP Agent 1. (The latter is located at HQ2, together with Cluster 2.)

The gatekeeper will then route the call to the inbound trunk of Cluster 2, based on its zone prefix configuration.

Unified CM Cluster 2 then receives a call from the inbound trunk (associated with location HQ1) and phone Y (associated with location Branch 2), so it will initiate another RSVP reservation across the IP WAN between Cisco RSVP Agent HQ2 and Cisco RSVP Agent Br 2.

The call is established across five call legs (four call legs if the Remote RSVP Agent 1 and the RSVP Agent HQ2 are co-resident), two of which traverse the IP WAN and are RSVP-enabled.

Note   In presence of deployments involving three Unified CM clusters, the same considerations apply, except that instead of having a single outbound trunk per cluster, you need two of them, one for each of the other two clusters that will be called. Similarly, each cluster needs two remote Cisco RSVP Agents, each located with one of the other two clusters.

Cisco Unified Border Element Approach

The complexity of the remote Cisco RSVP Agent approach described in the previous section quickly increases with the number of Unified CM clusters, and it is therefore limited to a maximum of three clusters.

To provide call admission control in generic topologies where more than three Unified CM clusters are located at different sites, you can combine RSVP-enabled locations for calls within a cluster with RSVP-enabled Cisco Unified Border Elements for calls between clusters, as shown in Figure 9-38.

Figure 9-38   Cisco Unified Border Element Approach for Distributed Clusters in a Generic Topology
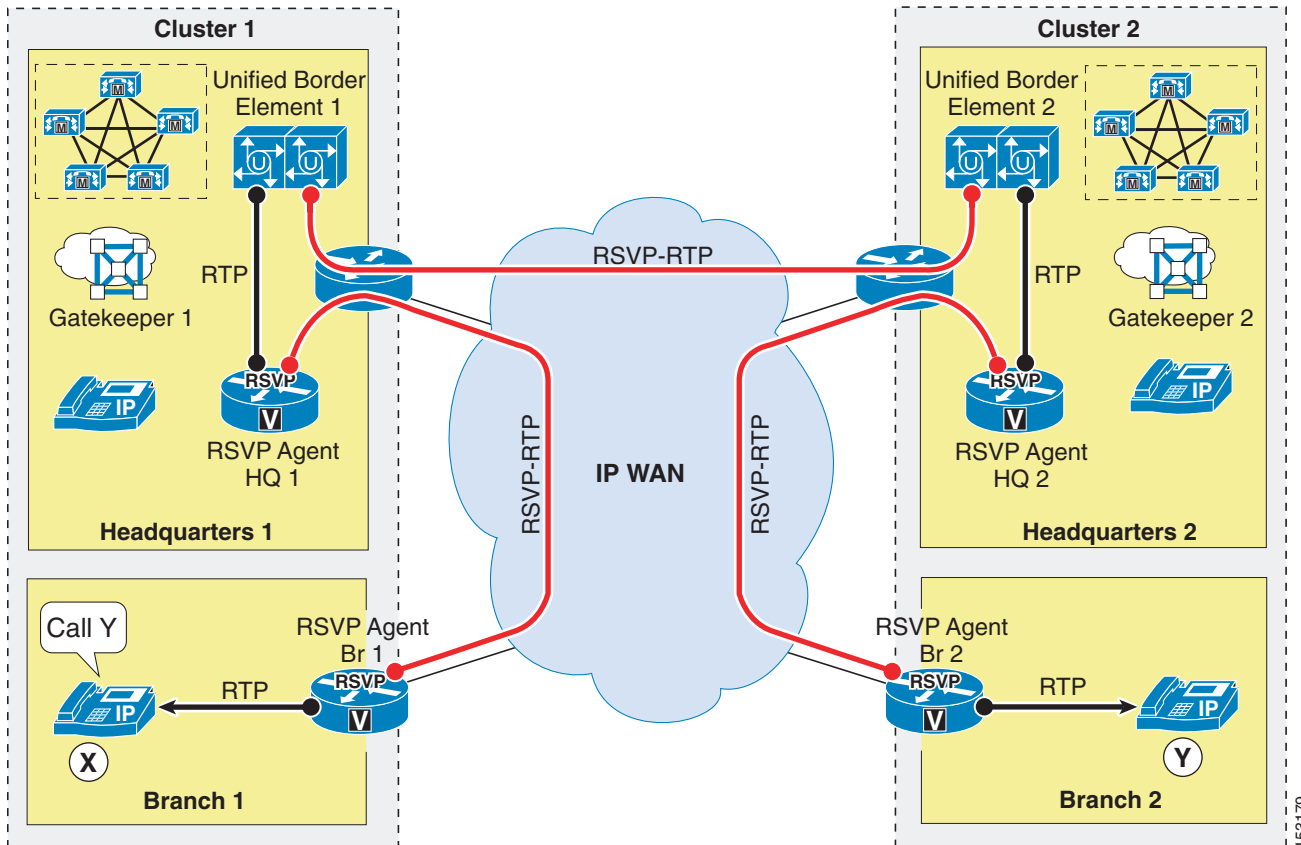
In addition to the guidelines listed in the section on , observe the following best practices for these deployments:

- For each cluster, define a gatekeeper-controlled intercluster trunk to enable communications to the other clusters. (Gatekeeper zones are used for dial plan resolution but are not needed for call admission control in this scenario.)

- Assign the intercluster trunk to the same location used for all devices located at the central site for that cluster.

- Ensure that the intercluster trunk is assigned to a device pool that specifies an MRGL that in turn points to an MRG containing the Cisco RSVP Agent located at the central site (for example, Cisco RSVP Agent HQ1 for Cluster 1 in Figure 9-38).

- For each cluster, place a Cisco Unified Border Element at the central site for that cluster and enable it to use RSVP for VoIP calls across the IP WAN.

- For each cluster, configure a gatekeeper as a via-zone gatekeeper so that it invokes the local Cisco Unified Border Element for all calls in or out of the respective zone. (Note that the gatekeeper can be co-resident with the Cisco Unified Border Element.)

- Use the AAR feature to provide automatic PSTN failover in case of call admission control failure within a cluster.

- Use the route list and route group constructs to provide automatic PSTN failover in case of call admission control failure across clusters.

- For calls between two branch sites belonging to different clusters, both media and signaling traffic are hair-pinned via the central sites of the respective clusters (as shown in Figure 9-38, where the call between phone X in Branch 1 and phone Y in Branch 2 is hair-pinned via the HQ1 and HQ2 sites).

Note   While logically separate, the Cisco RSVP Agents, gatekeeper, and Cisco Unified Border Element functions may reside on the same router platform. For example, in the scenario shown in Figure 9-38, Cisco Unified Border Element 1, Gatekeeper 1, and Cisco RSVP Agent HQ1 may resident on the same router platform, as may Cisco Unified Border Element 2, Gatekeeper 2, and Cisco RSVP Agent HQ2.

In Figure 9-38, when phone X at Branch 1 places a call to phone Y at Branch 2, Unified CM Cluster 1 will route the call to Gatekeeper 1 via the intercluster trunk. Because phone X is assigned to location Branch 1 and the intercluster trunk is associated to the location HQ1, Unified CM Cluster 1 will initiate an RSVP reservation across the IP WAN between Cisco RSVP Agent Br 1 and Cisco RSVP Agent HQ1.

Gatekeeper 1 then routes the call to Cisco Unified Border Element 1 based on the via-zone configuration, and Cisco Unified Border Element 1 establishes an RSVP reservation with Cisco Unified Border Element 2 across the IP WAN. Cisco Unified Border Element 2 in turn contacts Unified CM Cluster 2 via Gatekeeper 2.

Unified CM Cluster 2 then receives a call from the intercluster trunk associated with location HQ2 and directed to phone Y (associated with location Branch 2), so it will initiate another RSVP reservation across the IP WAN between Cisco RSVP Agent HQ2 and Cisco RSVP Agent Br 2.

The call is established across seven call legs, three of which traverse the IP WAN and are RSVP-enabled.

Design Recommendations for Call Admission Control

This section briefly summarizes the best practices for providing call admission control in various Cisco Unified Communications Manager (Unified CM) deployments.

The following recommendations apply to deployments with a single Unified CM cluster:

- For simple hub-and-spoke topologies with no dual links, use Unified CM static locations. Leave the hub site devices in the Hub_None location.

- For Multiprotocol Label Switching (MPLS) topologies with no dual links, use Unified CM static locations, with devices at every site (including the central site) assigned to a location.

- For any other topology, use Unified CM RSVP-enabled locations. Cisco recommends the **Mandatory** or **Mandatory (video desired)** policy as the default RSVP policy between sites. The Cisco RSVP Agent feature may reside on the IP WAN router in smaller sites or run on standalone platforms in larger sites.

The following recommendations apply to deployments with multiple Unified CM clusters:

- For simple hub-and-spoke topologies with no dual links, use Cisco IOS gatekeeper zones between sites where Unified CM clusters reside.

- For two-tier hub-and-spoke topologies with no dual links where Unified CM clusters are located at the first and second level hub sites, use Cisco IOS gatekeeper zones for the links between first- and second-level hub sites and use Unified CM static locations for the links between second-level hub sites and spoke sites.

- For MPLS topologies with no dual links, use Unified CM static locations, with every site in a location and with no gatekeeper zones. Leave intercluster trunks in the Hub_None location unless an MTP is required. You may use a gatekeeper for intercluster call routing, but it is not needed for call admission control.

- For any other topology and three or fewer clusters, use RSVP-enabled locations and the "remote agent" approach.

- For any other topology and more than three clusters, use RSVP-enabled locations within each cluster and gatekeepers with RSVP-enabled Cisco Unified Border Elements across clusters.