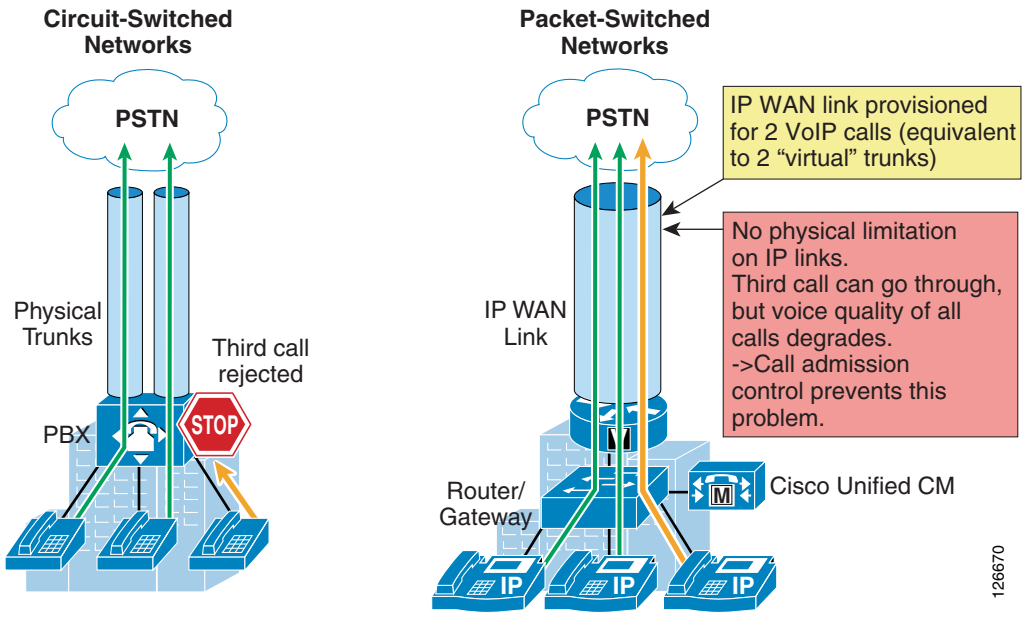
CHAPTER 9

Call Admission Control

Last revised on: February 13, 2008

The call admission control function is an essential component of any IP telephony system that involves 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:

- Best Practices Summary, page 9-2

  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 9-11

  This section describes the call admission control mechanisms available through the various components of a Cisco IP Communications system, such as Cisco Unified CallManager locations, Cisco IOS gatekeeper, RSVP, and the IP-to-IP gateway.

- Call Admission Control Design, page 9-25

  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 CallManager deployment model adopted.

Best Practices Summary

This section briefly summarizes the best practices for providing call admission control in various Cisco Unified CallManager deployments. The remainder of this chapter explains these best practices in more detail.

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

- For simple hub-and-spoke topologies, use Cisco Unified CallManager static locations. Leave the hub site devices in the <None> location.
- For Multiprotocol Label Switching (MPLS) topologies, use Cisco Unified CallManager static locations, with devices at every site (including the central site) assigned to a location.

The following recommendations apply to deployments with multiple Cisco Unified CallManager clusters:

- For simple hub-and-spoke topologies, use Cisco IOS gatekeeper zones between sites where Cisco Unified CallManager clusters reside.
- For two-tier hub-and-spoke topologies where Cisco Unified CallManager 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 Cisco Unified CallManager static locations for the links between second-level hub sites and spoke sites.

- For MPLS topologies with no dual links, use Cisco Unified CallManager static locations, with every site in a location and with no gatekeeper zones. Leave intercluster trunks in the <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 generic topologies where Cisco Unified CallManager clusters are located at each site, use RSVP-enabled IP-to-IP gateways across clusters.

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

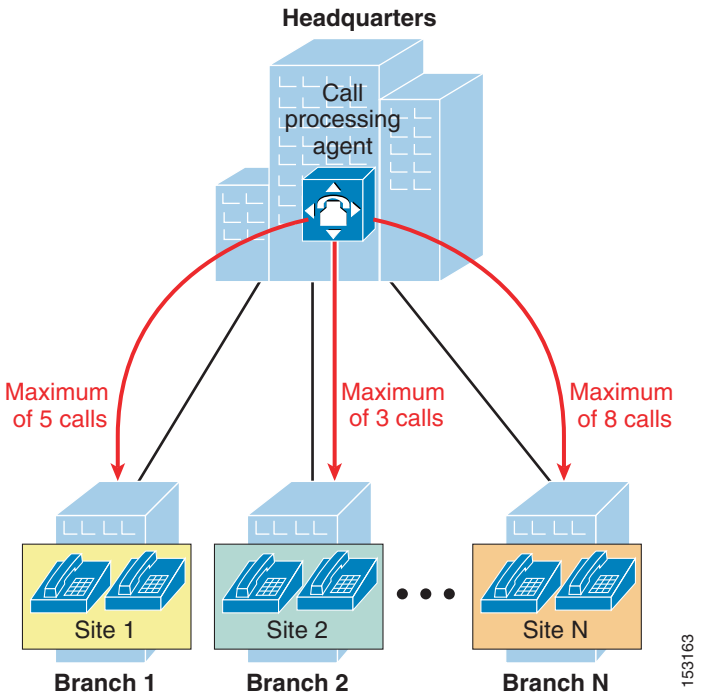
We define as topology-unaware call admission control any mechanism that is based on a static configuration within a call processing agent or IP-based PBX, aimed at limiting the number of simultaneous calls to or from a remote site connected via the IP WAN.

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.

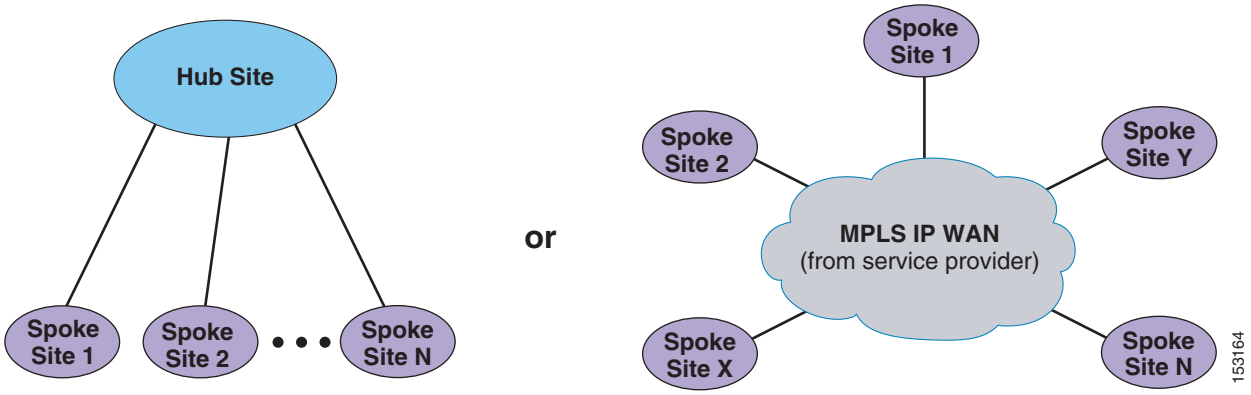
Each time a new call needs to be established, the call processing agent checks the sites to which the originating and terminating endpoints belong, and verifies whether there are available resources to place the call (in terms of number of calls or amount of bandwidth for both sites involved). If the 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.

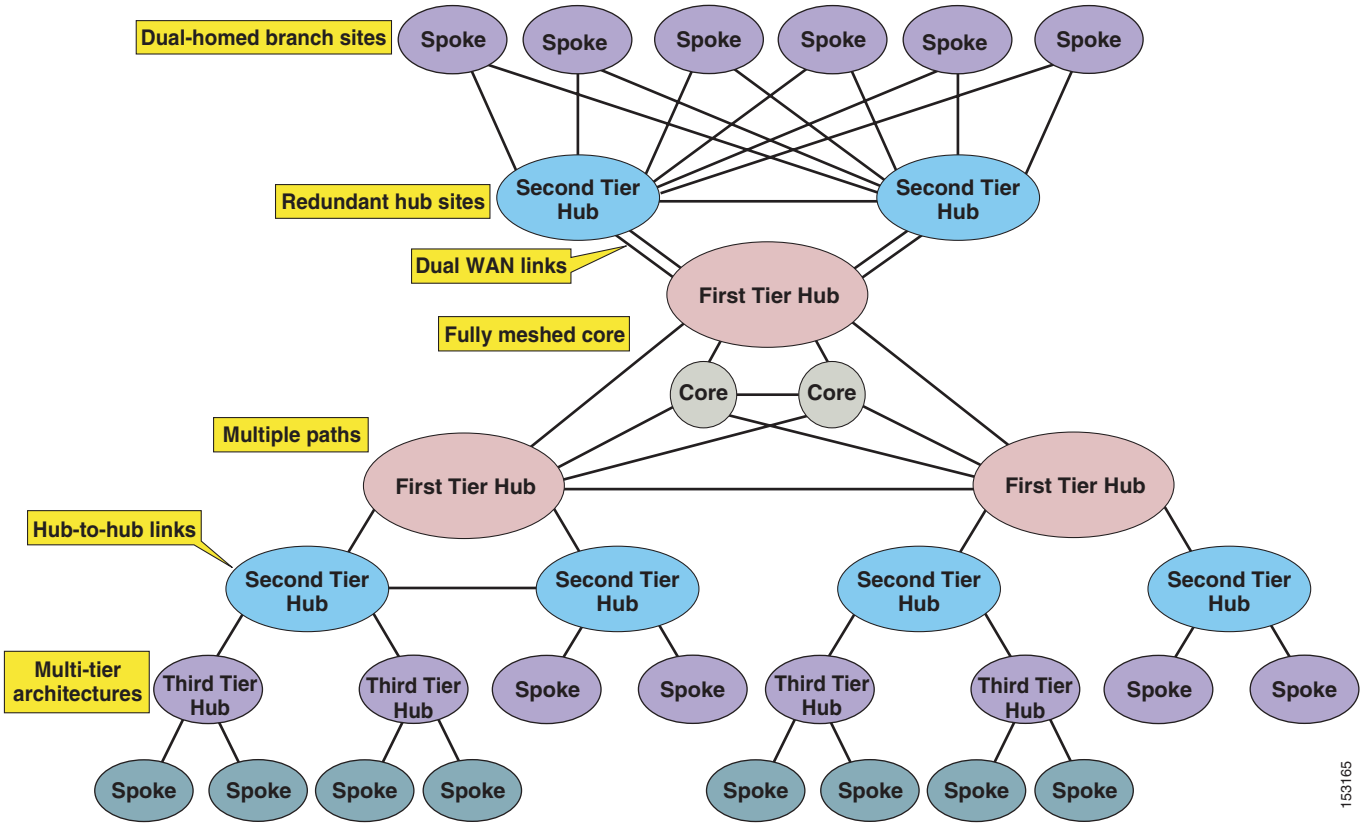
Notice the absence of redundant links from the spoke sites to the hub site and of links directly connecting two spoke sites. The next section explains why such links create problems for topology-unaware call admission control.

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

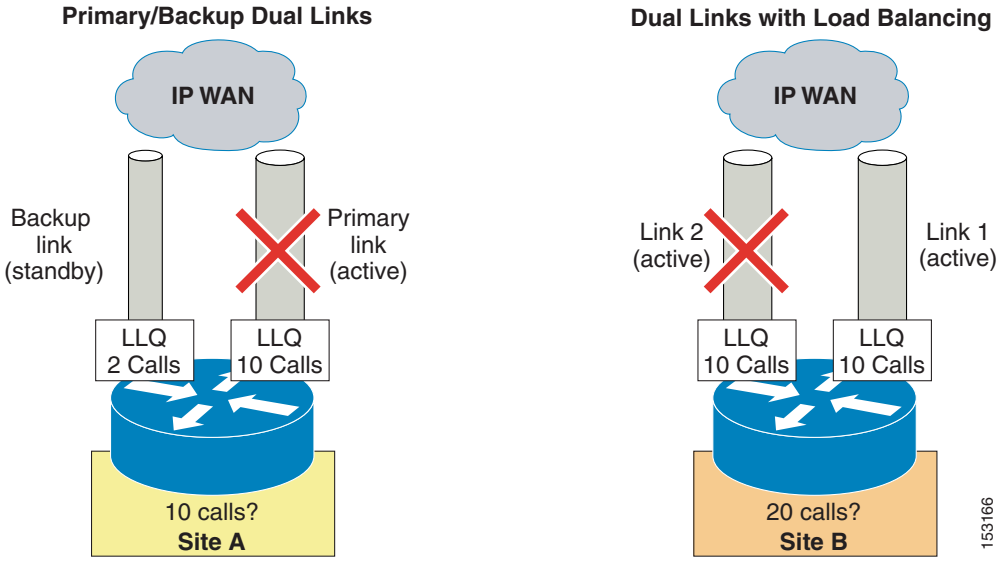


As explained in the section on Call Admission Control Design, page 9-25, it is sometimes possible to adapt a topology-unaware call admission control mechanism to a complex network topology, but there are limitations in terms of when this approach can be used and what behavior can be achieved. For example, 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.

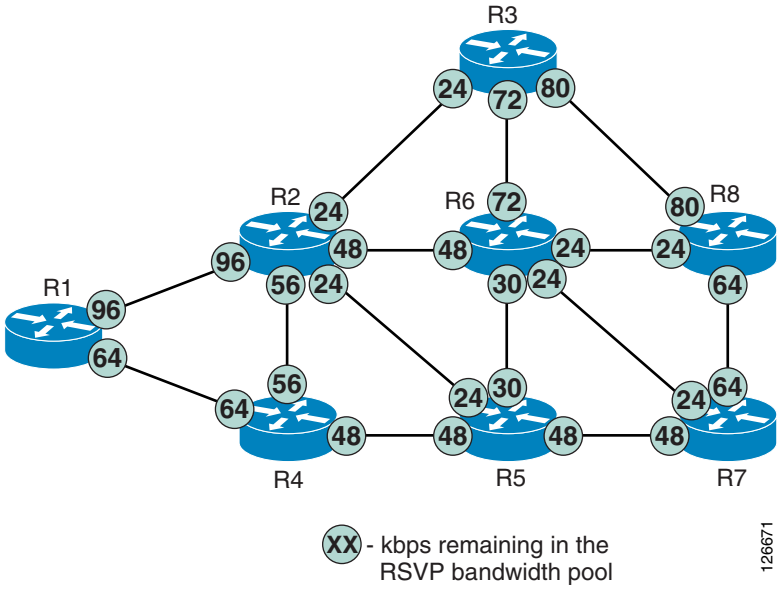
The Resource Reservation Protocol (RSVP) is the first significant industry-standard signaling protocol that enables an application to reserve bandwidth dynamically across an IP network. Using RSVP, applications can request a certain amount of bandwidth for a data flow across a network (for example, a 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 3-35.

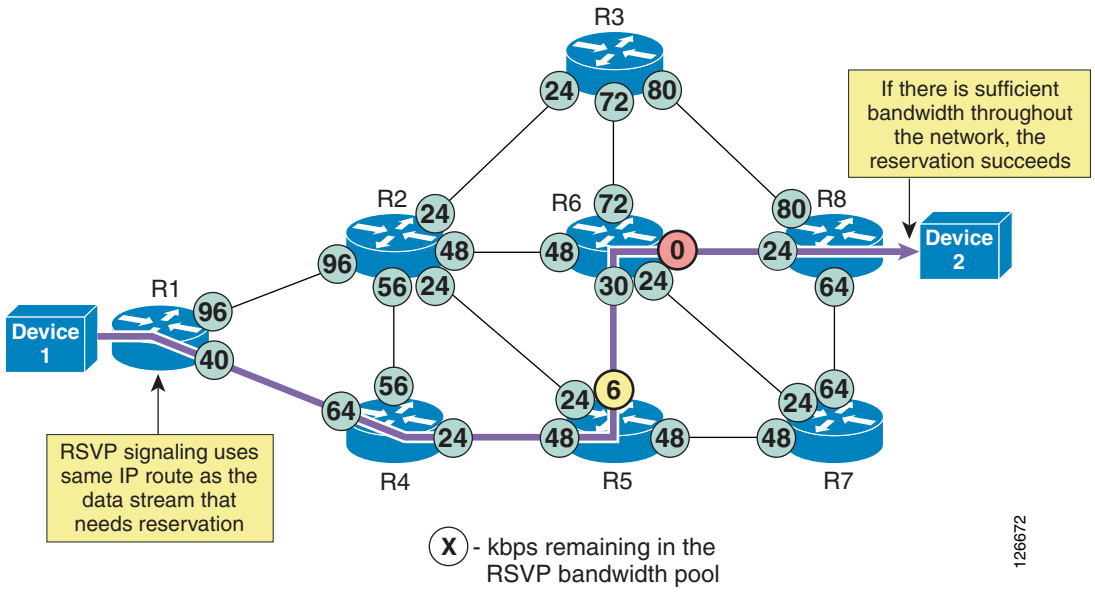
Assume that RSVP is enabled on each router interface in the network shown in Figure 9-6 and that the numbers shown in the circles represent the amount of available RSVP bandwidth remaining on each interface.

Figure 9-6 Sample Network to Show RSVP Principles



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

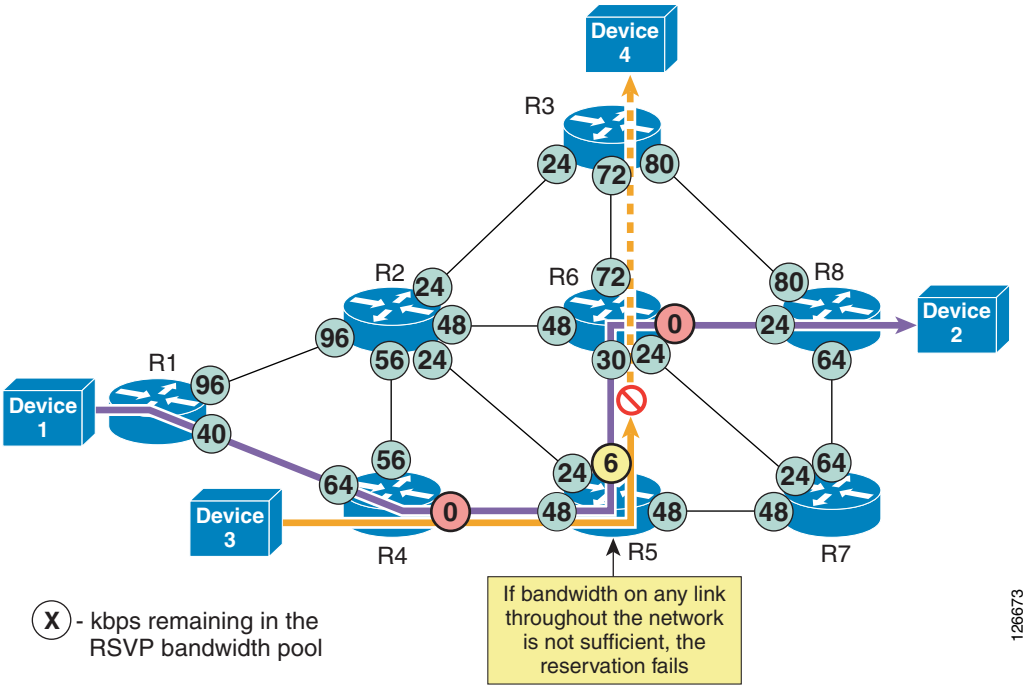


The following considerations apply to Figure 9-7:

- 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 3-35, 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:

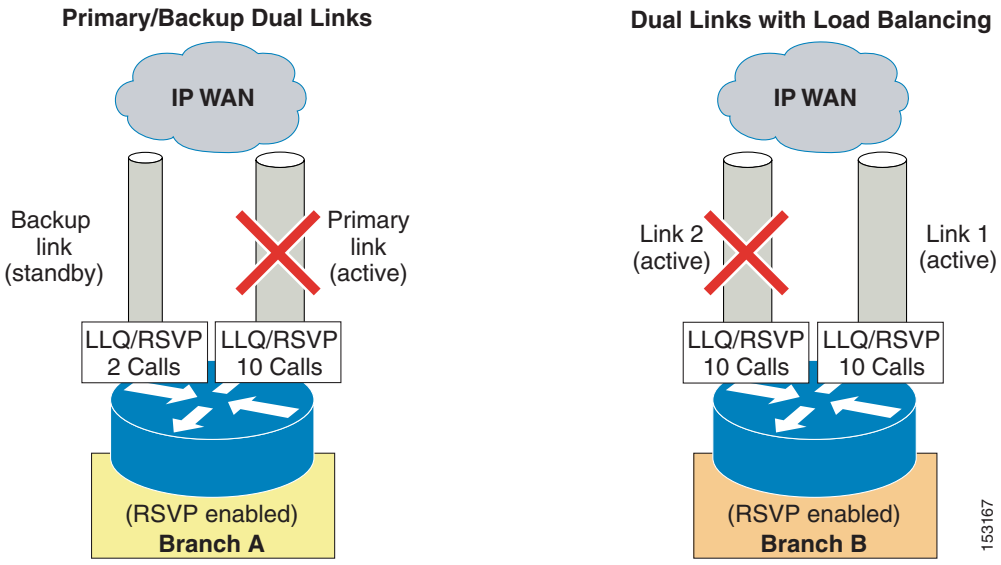
- The RSVP protocol attempts to establish an end-to-end reservation by checking for available bandwidth resources on all RSVP-enabled routers along the path from Device 3 to Device 4. As the RSVP messages progress through the network, the available RSVP bandwidth gets decremented by 24 kbps on the outbound 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 9-34.) 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.

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
  - Cisco Unified CallManager Static Locations, page 9-12
  - Cisco IOS Gatekeeper Zones, page 9-15
- Topology-aware mechanisms
  - Cisco IOS Gatekeeper and IP-to-IP Gateway with RSVP, page 9-17

Note Cisco Unified CallManager 5.0 introduces topology-aware call admission control by extending the concept of *locations*, which already existed in previous releases. Therefore, to avoid confusion, this document refers to the existing topology-unaware mechanism as *static locations* and to the new topology-aware mechanism as *RSVP-enabled* locations. Cisco Unified CallManager 4.2 provides support for static locations only.

Cisco Unified CallManager Static Locations

Cisco Unified CallManager provides a simple mechanism known as static locations for implementing call admission control across the IP WAN in centralized call processing deployments. When you configure a device in Cisco Unified CallManager, the device can be assigned to a location. A bandwidth pool is allocated to each location for calls to devices belonging to any other location. (Calls within the same location do not need call admission control.) Each time a call is set up between any two locations, the required amount of bandwidth is substracted from the corresponding location bandwidth pools, until there is no more bandwidth available in the pools. At that point, call admission control rejects subsequent calls to or from the given location.

By default, Cisco Unified CallManager assumes that each device's physical location matches the configured location. So if a device is moved from one physical location to another, the system administrator must perform a manual update on its location configuration. An alternative is to enable the Device Mobility feature, introduced in Cisco Unified CallManager Release 4.2. This feature enables Cisco Unified CallManager to assign a device dynamically to a device pool (and hence to a location) based on its IP address, and it is described in detail in the chapter on Device Mobility, page 22-1.

Note The Device Mobility feature is not available in Cisco Unified CallManager Release 5.0.

Each device is in location <None> by default. Location <None> is a special unnamed location that has unlimited audio and video bandwidth. If the devices at a given site are configured in the <None> location, all calls to and from that devices at that site will be admitted by the call admission control mechanism.

Cisco Unified CallManager 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 Cisco Unified CallManager. On the other hand, if the bandwidth values are set to a finite number of kilobits per second (kbps), Cisco Unified CallManager 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, Cisco Unified CallManager 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, Cisco Unified CallManager 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.

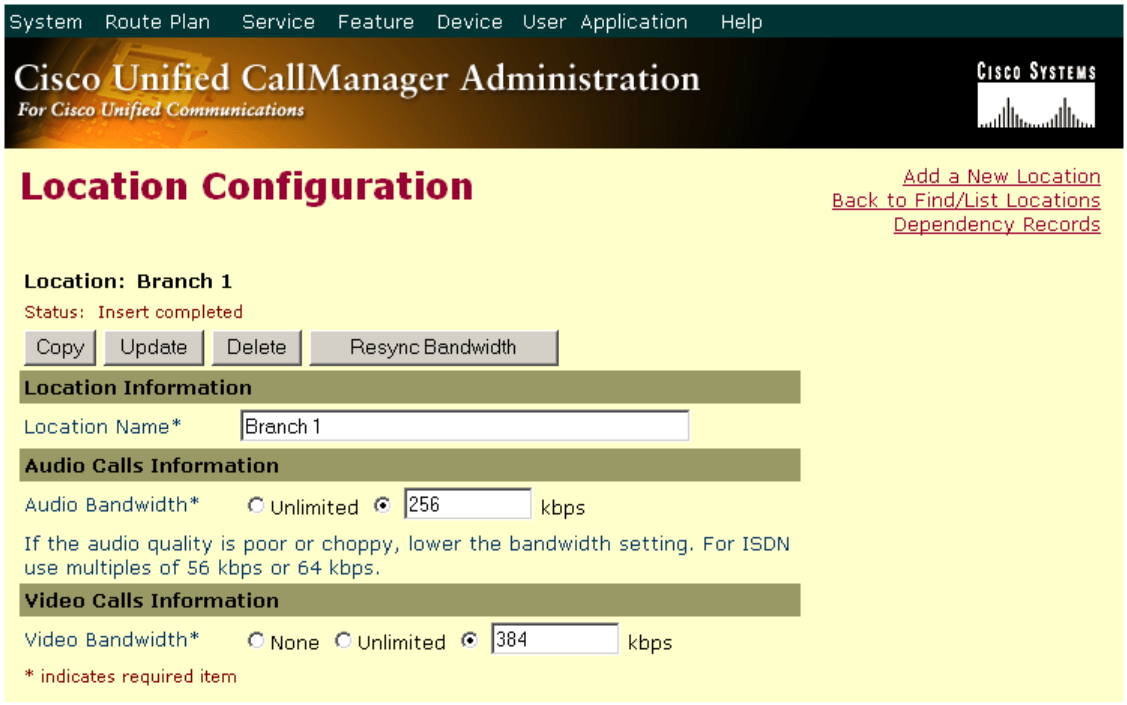
Table 9-1 lists the amount of bandwidth requested by the static locations algorithm for various call speeds. For an audio call, Cisco Unified CallManager counts the media bit rates plus the Layer 3 overhead. For example, a G.711 audio call consumes 80 kbps allocated from the location's audio bandwidth pool. For a video call, Cisco Unified CallManager 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, Cisco Unified CallManager will allocate 384 kbps from the video bandwidth pool.

Table 9-1 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

Figure 9-10 shows the configuration for the location Branch 1, with 256 kbps of available audio bandwidth and 384 kbps of available video bandwidth. 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.

Figure 9-10 Defining a Location in Cisco Unified CallManager



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, Cisco Unified CallManager deducts the appropriate amount of bandwidth from both locations. For example, a G.729 call between two locations causes Cisco Unified CallManager to deduct 24 kbps from the available bandwidth at both locations. When the call has completed, Cisco Unified CallManager returns the bandwidth to the affected locations. If there is not enough bandwidth at either branch location, the call is denied by Cisco Unified CallManager 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, Cisco Unified CallManager 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 10-22.

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 Cisco Unified CallManager. In such cases, the calls are redirected to the target specified in the Call Forward No Answer field of the called device.

Cisco IOS Gatekeeper Zones

A Cisco IOS gatekeeper can provide call routing and call admission control between devices such as Cisco Unified CallManager, Cisco Unified CallManager Express, 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 9-25, 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 8-18. For call routing considerations, refer to Call Routing in Cisco IOS with a Gatekeeper, page 10-38.

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 9-2 shows the bandwidth values used by the gatekeeper feature for some of the most popular call speeds.

Table 9-2 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

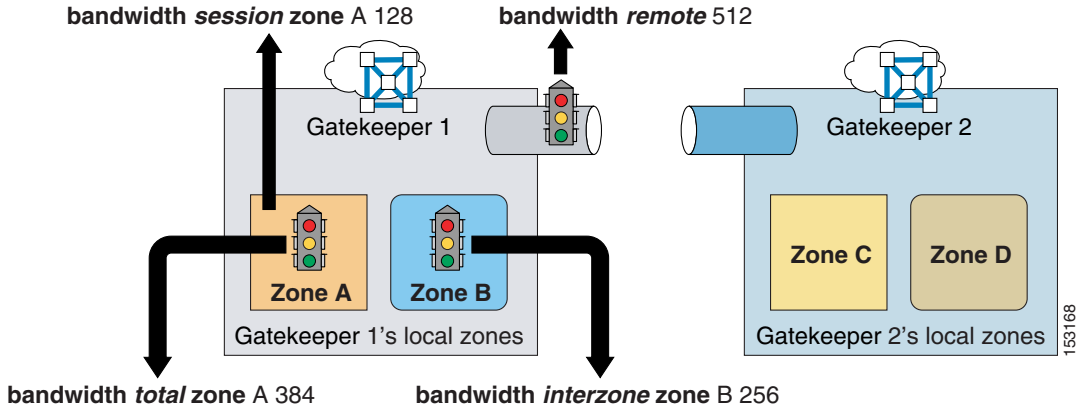
Table 9-2 Gatekeeper Bandwidth Settings for Various Call Speeds (continued)

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 3-44, 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 9-11.

Figure 9-11 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 9-11, 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.

Cisco IOS Gatekeeper and IP-to-IP Gateway with RSVP

The Cisco Multiservice IP-to-IP Gateway (also referred to as IP-IP gateway or IPIPGW) can be used to ease the restriction of hub-and-spoke topologies for the IP WAN connections between Cisco Unified CallManager clusters and/or H.323 gateways.

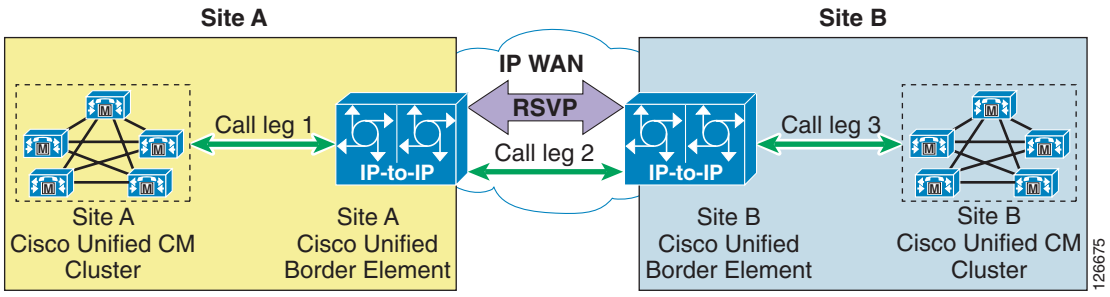
This Cisco IOS feature provides a mechanism to enable H.323 Voice over IP (VoIP) and videoconferencing calls from one IP network to another. The main purpose of the IP-IP gateway is to provide a control point and a demarcation for VoIP and video calls traversing administrative domains. This gateway performs 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 IP-IP gateway 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.

As a consequence, you can perform call admission control over any IP WAN topology by inserting two IP-IP gateways in the call flow and enabling RSVP between them. Figure 9-12 shows a basic example where two sites, A and B, each have a Cisco Unified CallManager cluster and are connected via an IP WAN that has an arbitrary topology. An IP-IP gateway is also located at each site, and the two Cisco Unified CallManager clusters are configured so that all inter-site calls are routed via a trunk that points to the local IP-IP gateway. When a call is set up between Site A and Site B, the following events occur:

- Cisco Unified CallManager at Site A sets up a call through an H.323 trunk to Site A's IP-IP gateway. (This is call leg 1 in the figure.)
- Site A's IP-IP gateway attempts to establish another call to Site B's IP-IP gateway, 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 IP-IP gateways.
- Site B's IP-IP gateway generates another call to Site B's Cisco Unified CallManager cluster. (This is call leg 3 in the figure.)

Figure 9-12 Simple Example of the IP-to-IP Gateway for RSVP Call Admission Control



The example in Figure 9-12 is a simple scenario in which all calls between Cisco Unified CallManager clusters are routed via a pair of IP-IP gateways. 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 Cisco Unified CallManager clusters, H.323 gateways, H.323 videoconferencing endpoints, and IP-IP gateways.

Note All scenarios involving IP-IP gateways described in this section apply to calls between different Cisco Unified CallManager clusters. Cisco does not recommend inserting IP-IP gateways for calls between endpoints registered to the same Cisco Unified CallManager cluster.

Via-Zone Gatekeeper

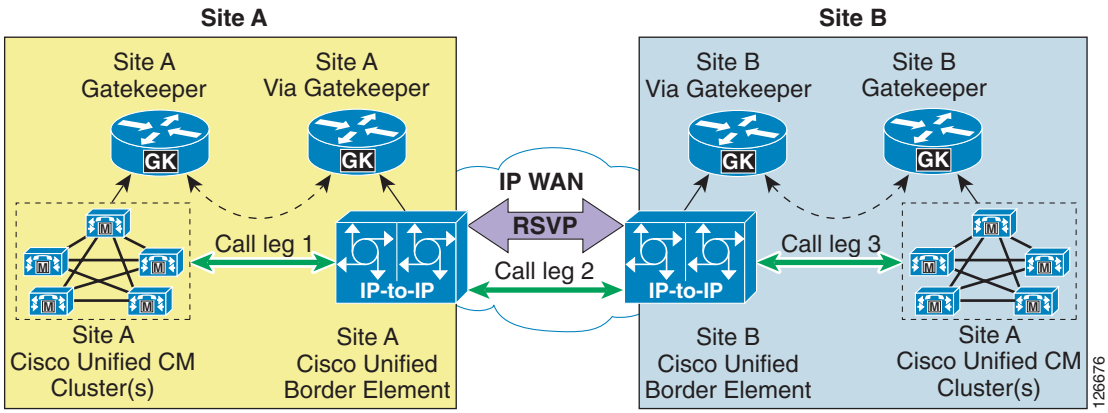
Traditional Cisco IOS gatekeeper functionality has been extended to accommodate for IP-IP gateways through the concept of a *via-zone gatekeeper*. A via-zone gatekeeper differs from legacy gatekeepers in how it uses LRQ and ARQ messages for call routing. Using via-zone gatekeepers will maintain normal 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 an IP-IP gateway in that zone by returning an ACF pointing to the IP-IP gateway. 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 9-13 shows an example of how IP-IP gateways and via-zone gatekeepers can be used in conjunction with Cisco Unified CallManager clusters and legacy gatekeepers to provide call routing and call admission control. The following considerations apply to this scenario:

- Site A's Cisco Unified CallManager 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 an IP-IP gateway for all calls coming from or destined to the Site A gatekeeper.
- The Site A IP-IP gateway attempts RSVP reservations before sending calls toward Site B's IP-IP gateway.
- The Cisco Unified CallManager clusters, gatekeepers, and the IP-IP gateway at Site B are configured in a similar fashion to their counterparts at Site A.

Figure 9-13 IP-to-IP Gateway for RSVP Using Via-Zone Gatekeepers



Design Best Practices

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

- When configuring trunks in Cisco Unified CallManager for either voice or video communication with other Cisco Unified CallManager clusters through one or more IP-IP gateways, use gatekeeper-controlled H.225 trunks. When using Cisco IOS Release 12.4(6)T or later and Cisco Unified CallManager Release 4.1 or later, MTP resources are no longer required to invoke supplementary services such as hold/resume, transfer, and conference across IP-IP gateways. In order to ensure interoperability, you must configure the following items:
  - On Cisco Unified CallManager 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 Cisco Unified CallManager Administration, under the Advanced Service Parameters page for Cisco Unified CallManager, set the **Send H225 User Info Message** field to **H225 Info For Call Progress Tone**.
  - On the IP-IP gateways, configure the following Cisco IOS commands to ensure interoperability with Cisco Unified CallManager 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 IP-IP gateway 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 IP-IP gateways if the **Media Termination Point required** option is checked on the H.225 trunks (because MTPs do not support video calls).

- Configure the IP-IP gateway as an H.323 gateway in Cisco Unified CallManager only when you want all intercluster calls to use the IP-IP gateway. In this case, the IP-IP gateway can still use a gatekeeper to resolve the remote destination.
- Configure gatekeeper-controlled intercluster trunks in Cisco Unified CallManager when you want to use a gatekeeper to resolve intercluster calls and decide whether they need to go through an IP-IP gateway or be routed directly. This approach is more flexible and allows for greater scalability.
- Compatibility with Cisco CallManager 3.3(2) or later on the IP-IP gateway 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 IP-IP gateway should have a dedicated via-zone gatekeeper.
- You can run the via-zone gatekeeper function and the IP-IP gateway function on the same router platform (co-resident); however, be aware of the scalability requirements described in the section on Redundancy, page 9-20.
- Do not use IP-IP gateways for calls between endpoints controlled by the same Cisco Unified CallManager cluster.
- Use the following options under the dial-peer configuration when enabling RSVP reservations for the IP-IP gateway:

```
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 IP-IP gateway will request an RSVP reservation using the guaranteed delay service. The fact that both the requested QoS and the acceptable QoS specify this RSVP service means that the RSVP reservation is mandatory for the call to succeed (that is, if the reservation cannot be established, the call will fail). For more information on configuration details, see Configuration Guidelines, page 9-21.

Redundancy

Redundancy and scalability can be provided by registering multiple IP-IP gateways 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 IP-IP gateways registered in the same via-zone, using a round-robin algorithm.

When an IP-IP gateway 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 IP-IP gateway so that a certain IP-IP gateway 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 IP-IP gateway:

```
ip circuit max-calls max-call-number
```

The above command specifies an aggregated session capacity of the IP-IP gateway, in terms of call legs. The default value is 1000 reserved call legs. Any call being handled by the IP-IP gateway 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 IP-IP gateway 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 IP-IP gateway. 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 IP-IP gateway, which can prevent the via-zone gatekeeper from sending a call to the IP-IP gateway 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

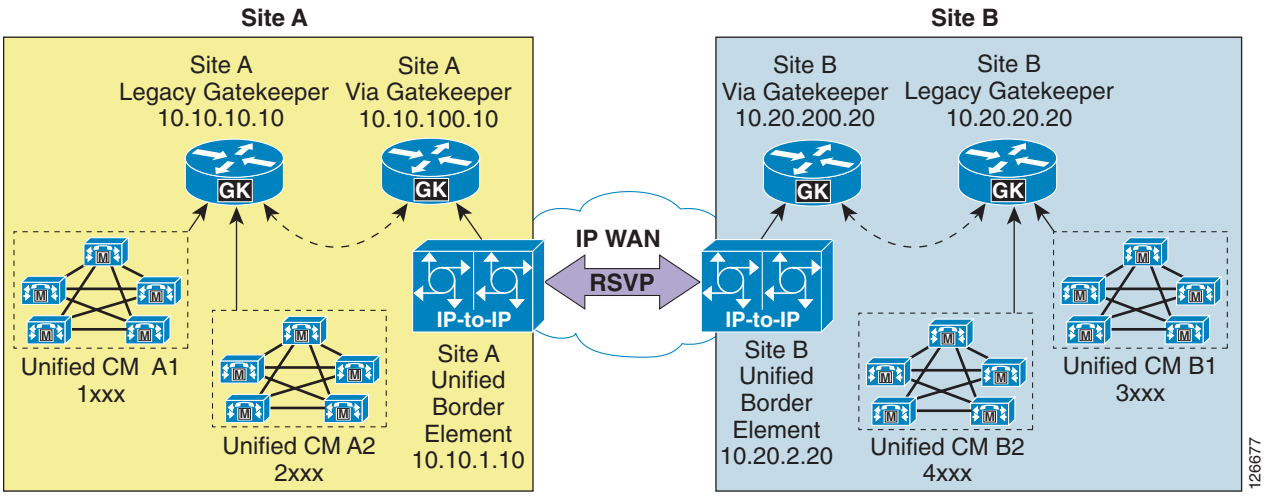
http://www.cisco.com

Configuration Guidelines

This section presents a simple configuration example based on the network diagram shown in Figure 9-14. 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 IP-IP gateways and via-zone gatekeepers is provided in the online documentation for the Cisco Multiservice IP-to-IP Gateway, available at

http://www.cisco.com

Figure 9-14 Configuration Example for IP-IP Gateway with Via-Zone Gatekeeper



For the network shown in Figure 9-14, assume that there are two Cisco Unified CallManager clusters at Site A: cluster A1, with phone extensions 1*xxx*, and cluster A2, with phone extensions 2*xxx*. There are also two Cisco Unified CallManager 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 Cisco Unified CallManager clusters (using Site A's legacy gatekeeper) while calls to Site B are routed through the two IP-IP gateways (using the respective legacy gatekeepers and via-zone gatekeepers).

Cisco Unified CallManager

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 IP-IP gateways.

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.

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 Cisco Unified CallManager clusters (extensions 3*xxx* and 4*xxx*) to Site B's via-zone gatekeeper and invokes an IP-IP gateway for calls to or from Site B. Calls directed to Site A's clusters are routed to Site A's legacy gatekeeper without invoking an IP-IP gateway. 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 lrq forward-queries
 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 an IP-IP gateway 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 IP-IP gateway is invoked for calls to or from these zones.

IP-IP Gateway

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

Example 9-3 IP-IP Gateway 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-IPIPGW
   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 transparent
!
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 transparent
!
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 transparent
!
dial-peer voice 20 voip
   destination-pattern [1-2]...
   session target ras
   codec transparent
```

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

- The **emptycapability** command enables the H.245 Empty Capabilities Set (ECS) between Cisco Unified CallManager and the IP-IP gateway to invoke supplementary services for the established calls.
- The **req-qos guaranteed-delay** [**audio** | **video**] commands specify that the IP-IP gateway 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 IP-IP gateway 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.
- Voice calls with a transparent codec for RSVP should be used only when the codecs are the same on both call legs of the IP-IP gateway.

Call Admission Control Design

This section describes how to apply the call admission control mechanisms to the various Cisco Unified CallManager deployment models and to the following IP WAN topologies:

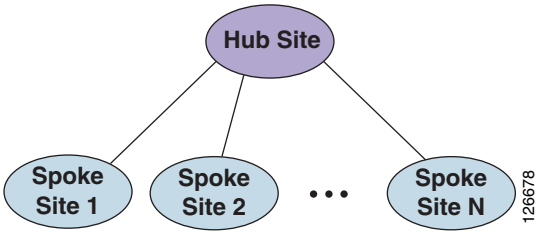
- Simple Hub-and-Spoke Topologies, page 9-25
- Two-Tier Hub-and-Spoke Topologies, page 9-29
- Simple MPLS Topologies, page 9-34
- Generic Topologies, page 9-41

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

Simple Hub-and-Spoke Topologies

Figure 9-15 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 9-15 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 9-34.

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

- Centralized Cisco Unified CallManager Deployments, page 9-26

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

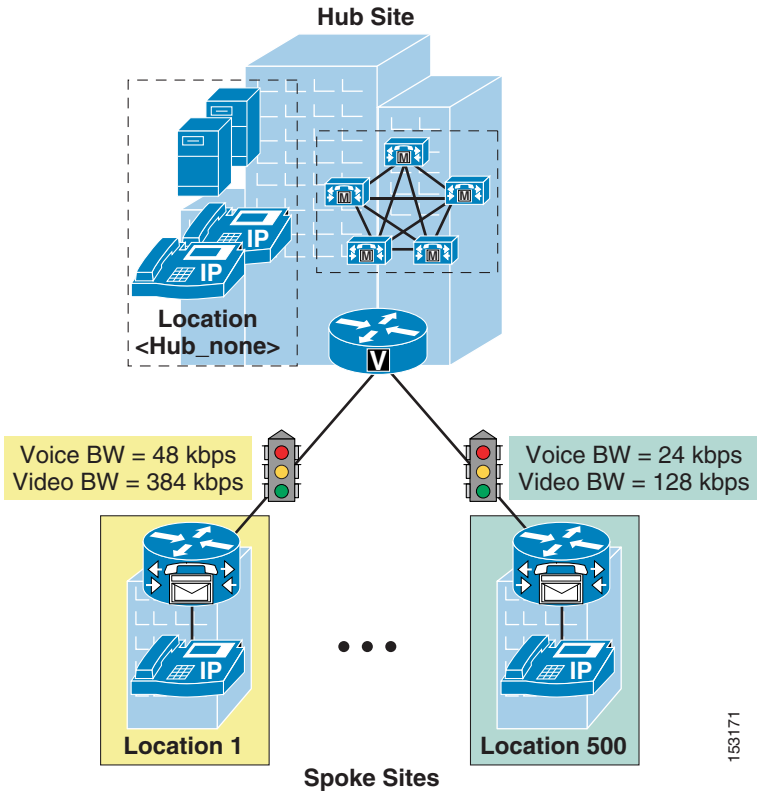
- Distributed Cisco Unified CallManager Deployments, page 9-27

  A Cisco Unified CallManager cluster or Cisco Unified CallManager Express is located at each site.

Centralized Cisco Unified CallManager Deployments

In multisite WAN deployments with centralized call processing in a simple hub-and-spoke topology, use Cisco Unified CallManager static *locations* for implementing call admission control. Figure 9-16 shows an example of how to apply this mechanism to such a topology.

Figure 9-16 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 Cisco Unified CallManager 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 9-1 for recommended bandwidth settings.)
- Assign all devices at each spoke site to the appropriate location.
- Leave devices at the hub site in the <None> location.
- If you move a device to another location, change its location configuration as well.
- Cisco Unified CallManager supports up to 500 locations.
- If you require automatic rerouting over the PSTN when the WAN bandwidth is not sufficient, configure the automated alternate routing (AAR) feature on Cisco Unified CallManager. (See Automated Alternate Routing, page 10-22.)
- If multiple Cisco Unified CallManager clusters are located at the same hub site, leave the intercluster trunk devices in the <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.

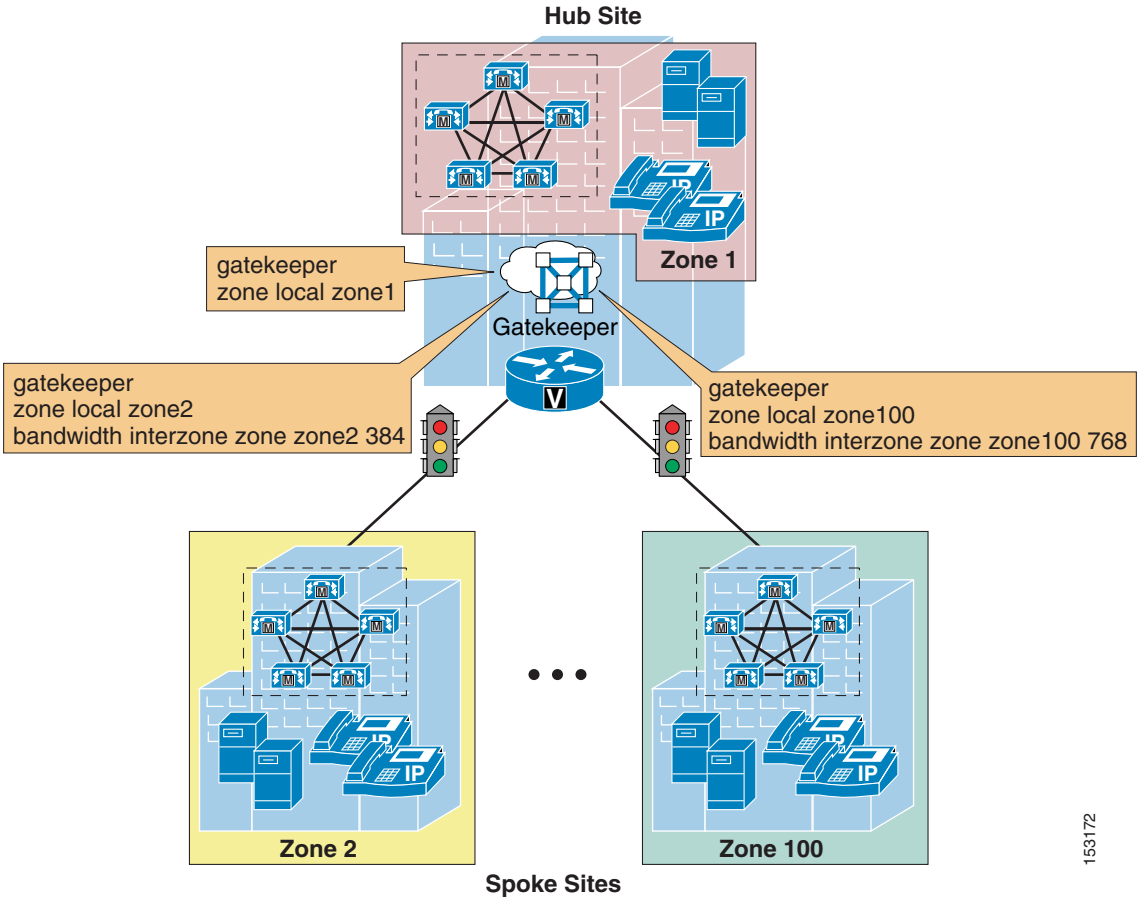
Note If one or more sites have dual connectivity to the IP WAN, you need to configure the location bandwidth according to the worst-case scenario (that is, only the smallest of the two IP WAN links is active), or upgrade to Cisco Unified CallManager 5.0 and deploy RSVP-enabled locations. See Limitations of Topology-Unaware Call Admission Control, page 9-5, for more information.

Distributed Cisco Unified CallManager 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 Cisco Unified CallManager cluster, Cisco Unified CallManager Express, 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 9-17 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 9-17 Call Admission Control for Hub-and-Spoke Topologies Using a Gatekeeper



Follow these guidelines when deploying call admission control with a gatekeeper:

- In Cisco Unified CallManager, configure an H.225 gatekeeper-controlled trunk if you have a mixed environment with Cisco Unified CallManager Express and H.323 gateways.
- In Cisco Unified CallManager, configure an intercluster gatekeeper-controlled trunk if you have an environment exclusively based on Cisco Unified CallManager clusters.
- Ensure that the zone configured in Cisco Unified CallManager matches the correct gatekeeper zone for the site.
- Each Cisco Unified CallManager subscriber listed in the device pool's CallManager Redundancy Group registers a gatekeeper-controlled trunk with the gatekeeper. (Maximum of three.)
- Calls are load-balanced across the registered trunks in the Cisco Unified CallManager cluster.
- Cisco Unified CallManager 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 Cisco Unified CallManagers, Cisco Unified CallManager Express, or H.323 gateways.

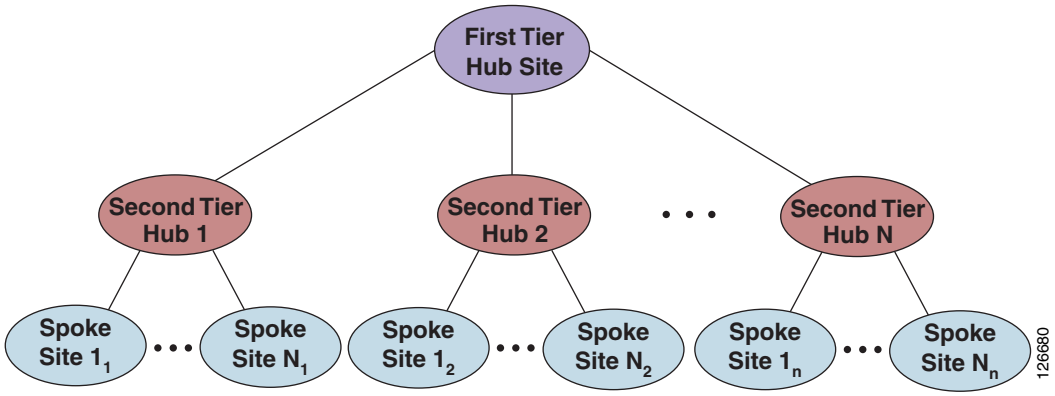
- Use the **bandwidth interzone** command on the gatekeeper to control bandwidth between Cisco Unified CallManager clusters, Cisco Unified CallManager Express servers, and H.323 devices registered directly with the gatekeeper. (See Table 9-2 for bandwidth settings by codec type.)
- A single Cisco IOS 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 using RSVP-enabled IP-IP gateways, as described in the section on Generic Topologies, page 9-41. See also Limitations of Topology-Unaware Call Admission Control, page 9-5, for more information.

Two-Tier Hub-and-Spoke Topologies

Figure 9-18 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 9-18 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 9-34.

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

- Centralized Cisco Unified CallManager Deployments, page 9-30

  One or more Cisco Unified CallManager 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.

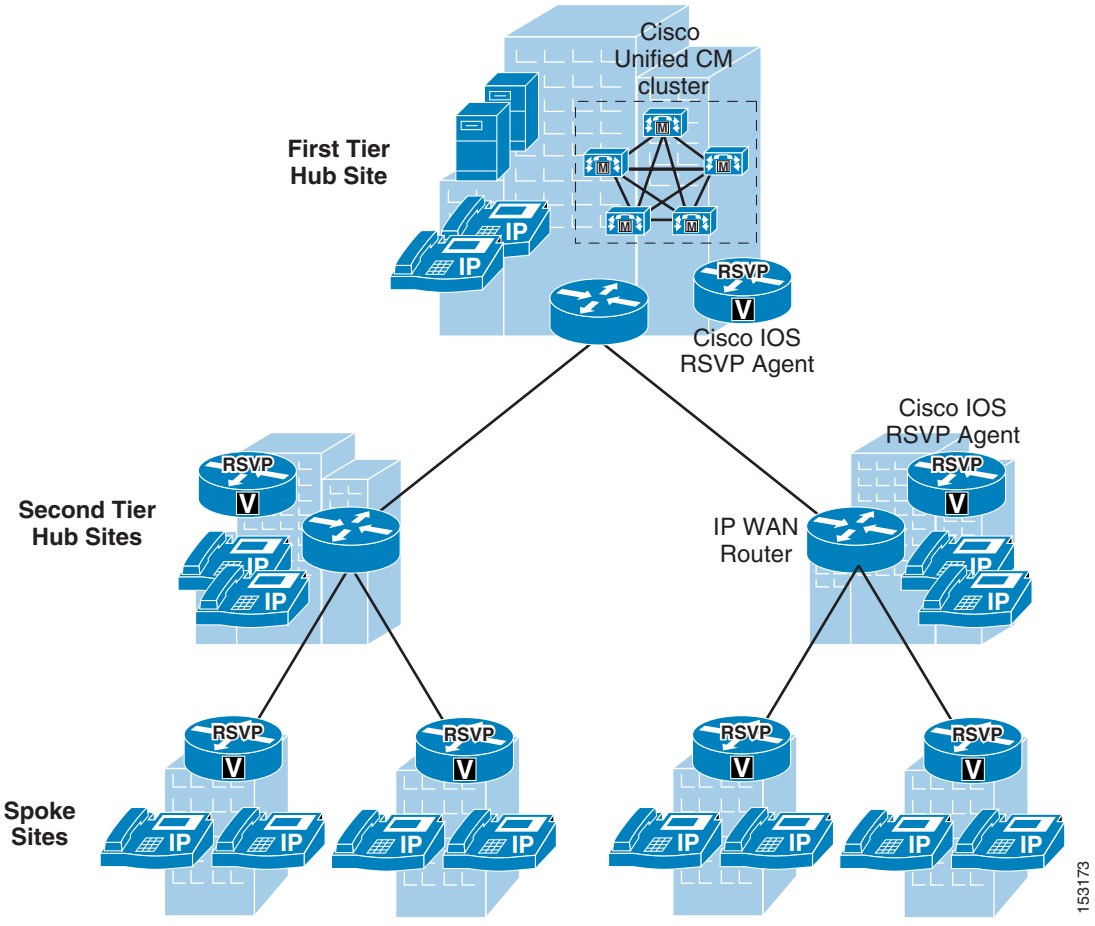
- Distributed Cisco Unified CallManager Deployments, page 9-33

  Cisco Unified CallManager 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 Cisco Unified CallManager Deployments

Figure 9-19 depicts a single centralized Cisco Unified CallManager cluster deployed in a two-tier hub-and-spoke IP WAN topology. In this scenario, the Cisco Unified CallManager cluster is located at the first-tier hub site, while all second-tier hub and spoke sites contain only endpoints and gateways.

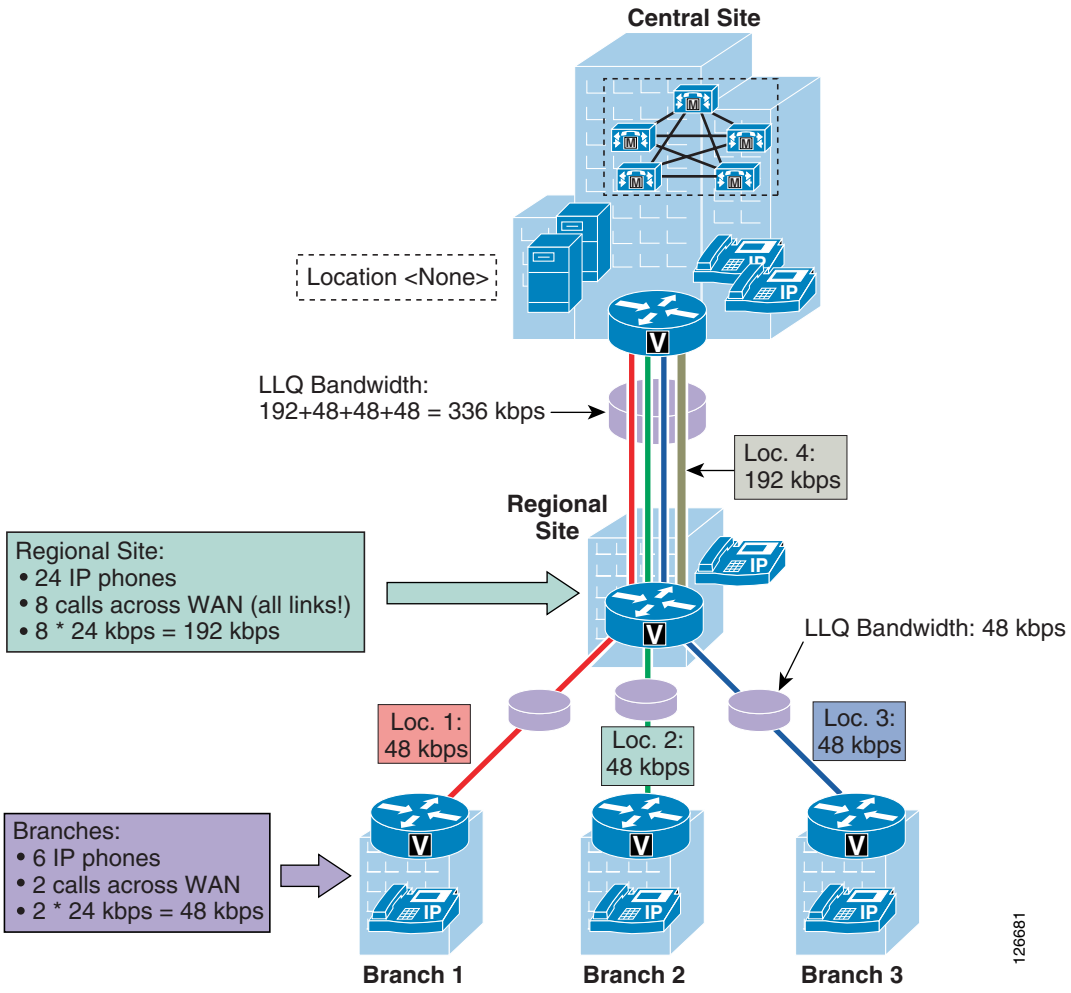
Figure 9-19 Two-Tier Hub-and-Spoke Topology with Centralized Cisco Unified CallManager



This type of topology poses a problem in terms of call admission control. The Cisco Unified CallManager static locations algorithm is the only available mechanism for call admission control in a centralized call processing deployment, and it is designed to work for simple hub-and-spoke topologies. This section describes an attempt to mitigate the issues generated by such topologies, but you should keep in mind that this workaround has limitations, and the recommended approaches to solve this problem are either to change the topology to a simple hub-and-spoke or to upgrade to Cisco Unified CallManager 5.0 and deploy RSVP-enabled locations.

Figure 9-20 shows a possible approach to enable a Cisco Unified CallManager 4.2 cluster in a centralized call processing deployment to support a two-tier hub-and-spoke topology.

Figure 9-20 Two-Tier Hub-and-Spoke in a Purely Centralized Deployment



When using this approach, adhere to the following guidelines:

- Leave the devices at the first-tier hub site (Central Site) in the <None> location.
- Place devices at each spoke site in their own location, and define the location bandwidth according to how many calls you want to allow in or out of each site. In the example of Figure 9-20, the location bandwidth for the three spoke sites (branches 1, 2 and 3) is set to 48 kbps, or enough for two calls using the G.729 codec.

- Provision the priority queue bandwidth in the LLQ configuration for the links between the second-tier hub site and the spoke sites so that it matches the spoke site location bandwidth. In this example, this bandwidth setting is 48 kbps.
- Place devices at each second-tier hub site in their own location, and define the location bandwidth according to how many calls you want to allow in or out of each site. Keep in mind that this bandwidth accounts for all calls entering or leaving the second-tier hub site, so it includes calls toward the underlying spoke sites as well as calls toward the first-tier hub site. In the example of Figure 9-20, the location bandwidth for the second-tier hub site shown (Regional Site) is set to 192 kbps, or enough for eight calls using the G.729 codec.
- Provision the priority queue bandwidth in the LLQ configuration for the links between the first-tier hub site and each second-tier hub site so that it matches the sum of the location bandwidth of the second-tier hub site itself plus all the location bandwidths of the spoke sites connected to that second-tier hub site. In this example, this bandwidth setting is 192 + 48 + 48 + 48 = 336 kbps.

Note For simplicity, this example shows LLQ bandwidth provisioning based on Layer 3 values (that is, 24 kbps for a G.729 call). In reality, Layer 2 overhead also must be taken into account when provisioning the LLQ queues. Refer to Table 3-6 in the section on Bandwidth Provisioning, page 3-44, for a complete list of bandwidth values for the various Layer 2 WAN technologies.

As shown by the example in Figure 9-20, this approach relies on over-provisioning the priority queue bandwidth between the first-tier and the second-tier hubs to compensate for the fact that Cisco Unified CallManager locations are unaware that groups of spoke sites are actually connected via a certain second-tier hub site. As a consequence, the configuration must account for bandwidth as if communications between any two sites has to transit through the first-tier hub site.

While this approach protects voice quality under all circumstances, it also involves the following design considerations and caveats:

- As the number of spoke sites increases per second-tier hub site, the amount of priority queue bandwidth that must be provisioned between the second-tier sites and the first-tier hub site also increases significantly.
- The bandwidth utilization is suboptimal because the calculations are based on the worst-case scenario. For example, in Figure 9-20, if each of the three branches has two active calls to the Regional Site, Cisco Unified CallManager will allow only two additional calls from the Regional Site to the Central Site, even if the actual bandwidth available in the priority queue would allow for 14 additional calls.
- Call completion across the IP WAN cannot be guaranteed even if the actual bandwidth resources would allow for it. For example, in Figure 9-20, if the Regional Site has eight active calls to the Central Site, Cisco Unified CallManager will not allow any of the branches to call the Regional Site, even if the actual bandwidth available would allow for two calls from each branch. On the other hand, in the same situation the branches will be allowed to call the Central Site.

Note Cisco recommends that you do not place all spoke sites connected to a given second-tier hub site in the same location as the second-tier hub site. This solution cannot guarantee voice quality in all scenarios and is *not* supported by Cisco.

As mentioned previously, the workaround illustrated in this section has several limitations. For a better solution, Cisco recommends that you implement one of the following changes:

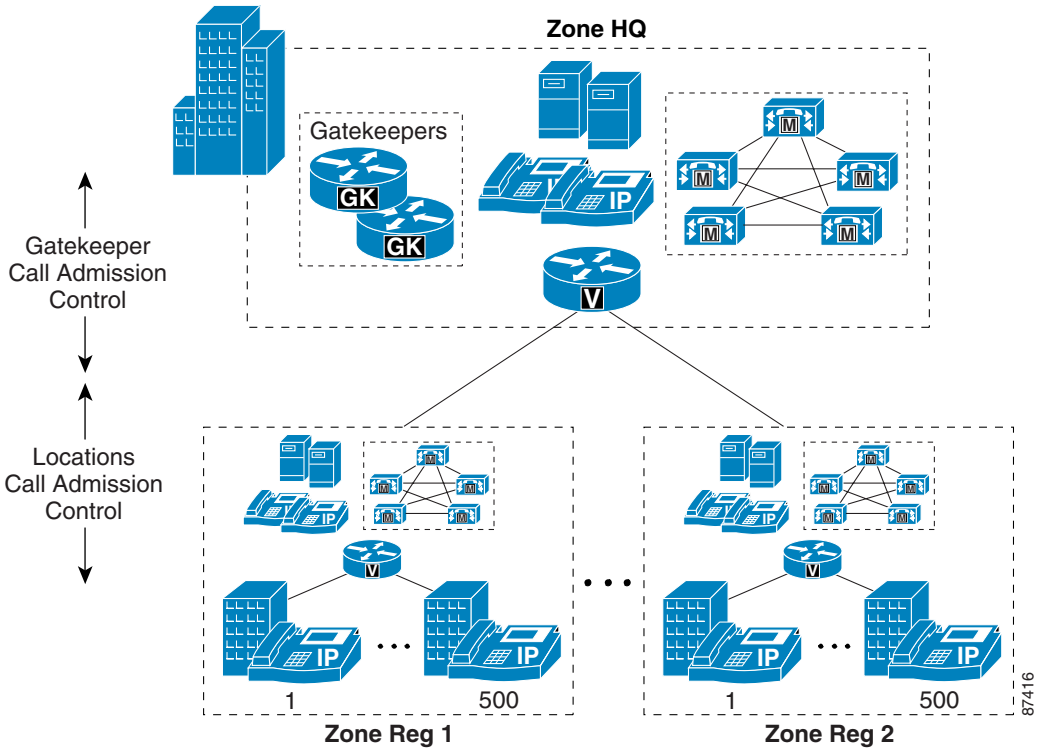
- Change the network topology to a simple hub-and-spoke by connecting all spoke sites directly to the first-tier hub site. Use Cisco Unified CallManager static locations to perform call admission control, as described in the section on Simple Hub-and-Spoke Topologies, page 9-25.
- Change the network topology to an MPLS-based network. Use Cisco Unified CallManager static locations to perform call admission control, as described in the section on Simple MPLS Topologies, page 9-34.
- Upgrade to Cisco Unified CallManager 5.0. Use Cisco Unified CallManager RSVP-enabled locations to perform call admission control, as described in the *Cisco Unified Communications SRND Based on Cisco Communications Manager 5.x*, available at

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

Distributed Cisco Unified CallManager Deployments

To provide call admission control in deployments that use a two-tier hub-and-spoke topology, with Cisco Unified CallManagers at the first-tier and second-tier hub sites, you can combine the static locations and gatekeeper zone mechanisms as illustrated in Figure 9-21.

Figure 9-21 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 Cisco Unified CallManager (that is, the spoke sites).

- Use gatekeeper-based call admission control between Cisco Unified CallManager clusters (that is, between the first-tier hub site and the second-tier hub sites).
- For each site without a local Cisco Unified CallManager, configure a location for that site in the Cisco Unified CallManager 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 9-1 and Table 9-2 for bandwidth settings.)
- Assign each device configured in Cisco Unified CallManager to a location. If you move a device to another location, change its location configuration as well.
- Cisco Unified CallManager supports up to 500 locations.
- Each Cisco Unified CallManager cluster registers a gatekeeper-controlled trunk with the gatekeeper.
- On the gatekeeper, configure a zone for each Cisco Unified CallManager cluster, and use the **bandwidth interzone** command to control the number of calls to and from each cluster.

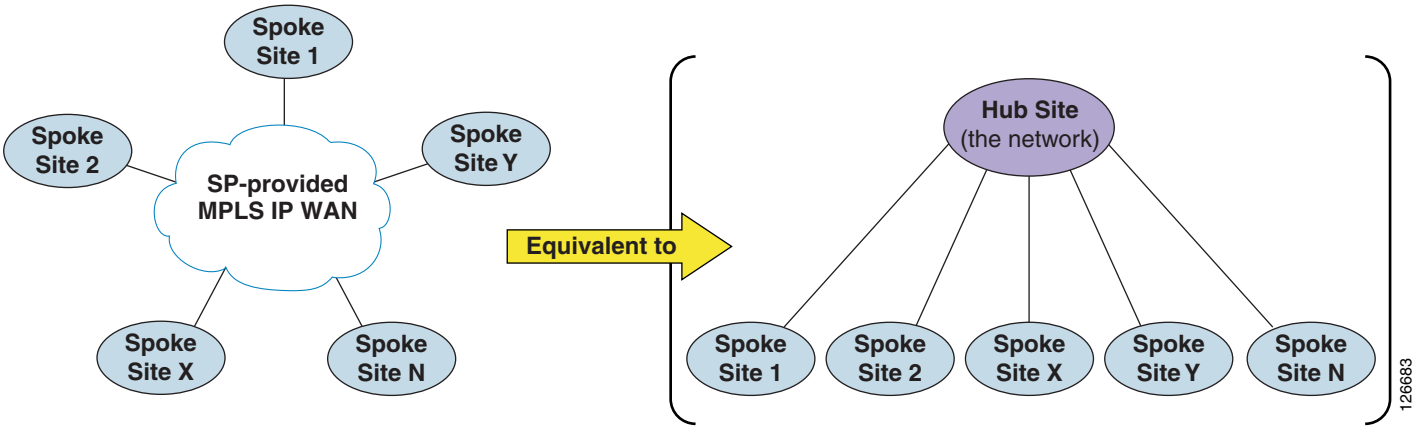
Note If one or more sites have dual connectivity to the IP WAN, you need to configure the location bandwidth according to the worst-case scenario (that is, only the smallest of the two IP WAN links is active), or upgrade to Cisco Unified CallManager 5.0 and deploy RSVP-enabled locations. See Limitations of Topology-Unaware Call Admission Control, page 9-5, for more information.

Simple MPLS Topologies

Figure 9-22 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 9-22 MPLS IP WAN from a Service Provider, and Its Topology Equivalent



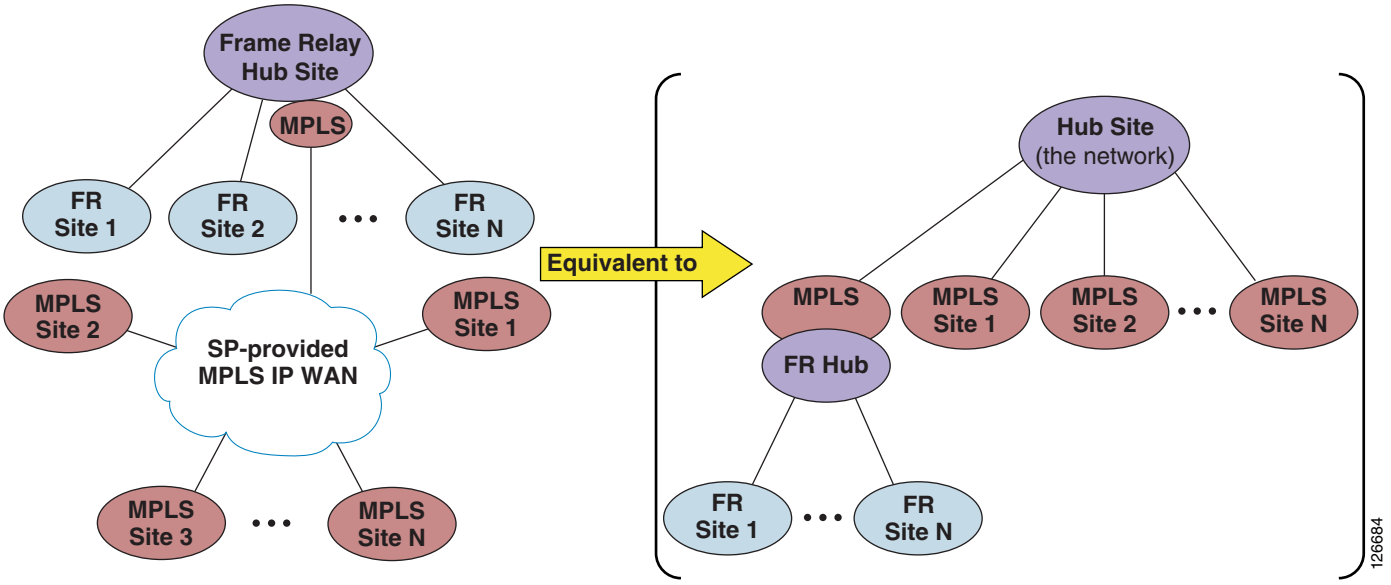
Based on these considerations, it is easy to see that, from a call admission control perspective, a service-provider IP WAN service based on MPLS is in reality equivalent to a hub-and-spoke topology without a hub site. (See Figure 9-22.) In fact, the network itself could be considered as the hub site, 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 9-23, 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 9-29.

Figure 9-23 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 Cisco Unified CallManager deployment model adopted:

- Centralized Cisco Unified CallManager Deployments, page 9-36

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

- Distributed Cisco Unified CallManager Deployments, page 9-39

  Cisco Unified CallManager 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, you need to configure the location bandwidth according to the worst-case scenario (that is, only the smallest of the two IP WAN links is active), or upgrade to Cisco Unified CallManager 5.0 and deploy RSVP-enabled locations. See Limitations of Topology-Unaware Call Admission Control, page 9-5, for more information.

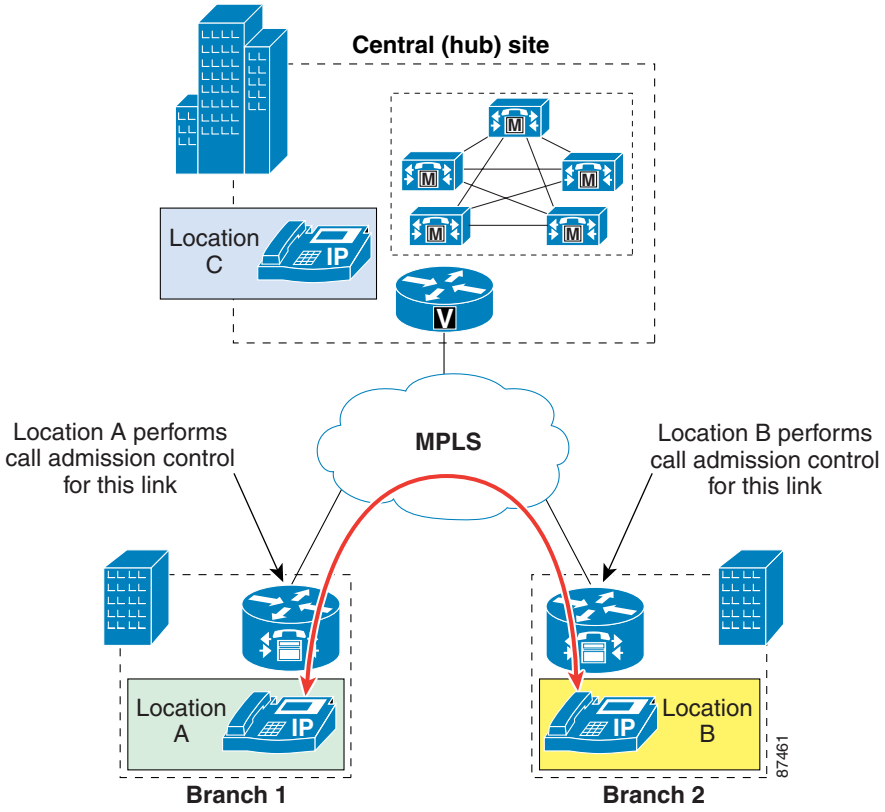
Centralized Cisco Unified CallManager Deployments

In multisite WAN deployments with centralized call processing in an MPLS topology, use Cisco Unified CallManager 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 Cisco Unified CallManager locations configuration, devices at the central site are left in the <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 9-24 illustrates a spoke-to-spoke call between two branch sites in this type of deployment.

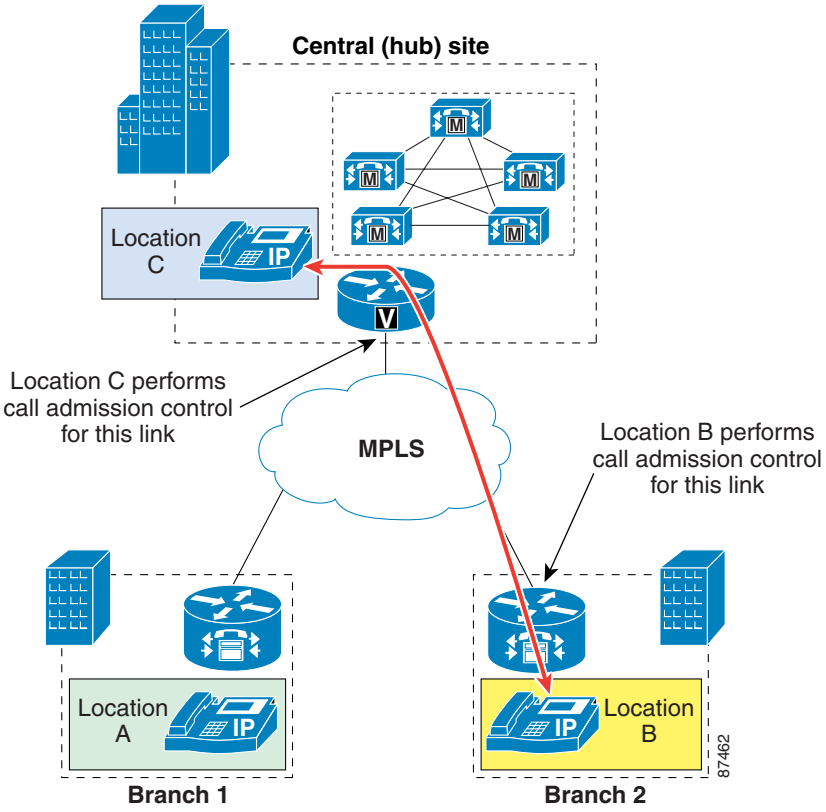
Figure 9-24 Spoke-to-Spoke Calls in an MPLS Deployment



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 <None> location), this configuration requires that call admission control be performed on the central site link independently of the branch links. (See Figure 9-25.)

Note Some devices such as trunks do not terminate media and are normally left in the <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 <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 9-25 Calls to and from the Hub in an MPLS Deployment

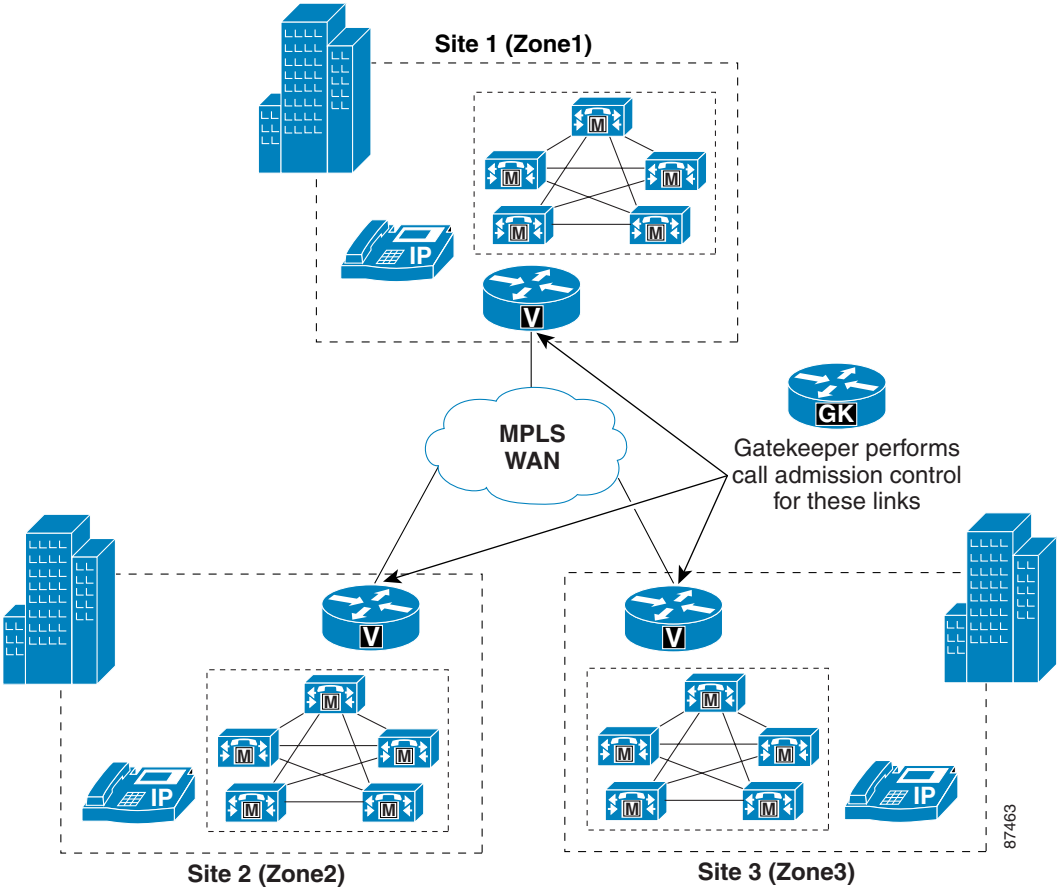


When all the available bandwidth for a particular site has been utilized, you can provide automatic failover to the PSTN by using the automated alternate routing (AAR) feature within Cisco Unified CallManager. (For more information on AAR, see Automated Alternate Routing, page 10-22.)

Distributed Cisco Unified CallManager Deployments

In multisite deployments where a Cisco Unified CallManager 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 9-26.)

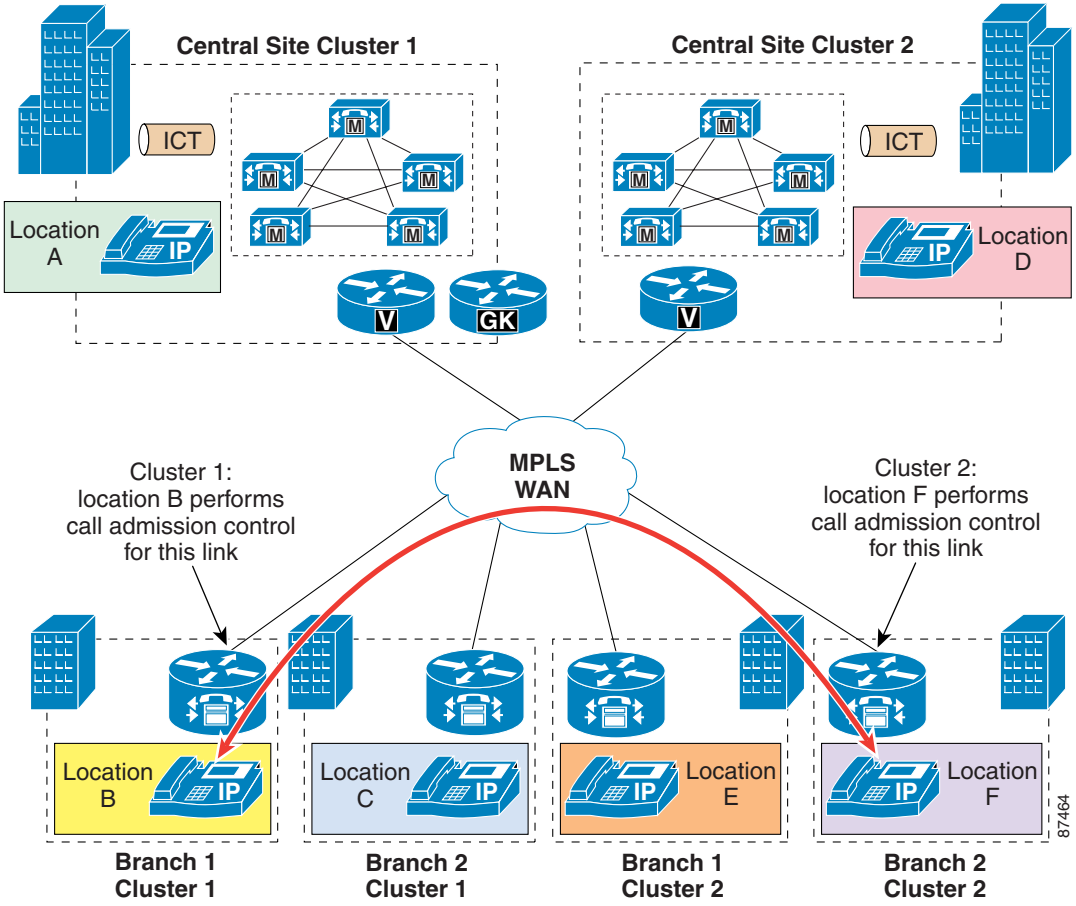
Figure 9-26 Gatekeeper Call Admission Control in a Distributed Deployment with MPLS



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 9-27.)

Figure 9-27 Multiple Clusters Connected by Intercluster Trunks (ICTs)



As in the centralized Cisco Unified CallManager 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 <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:

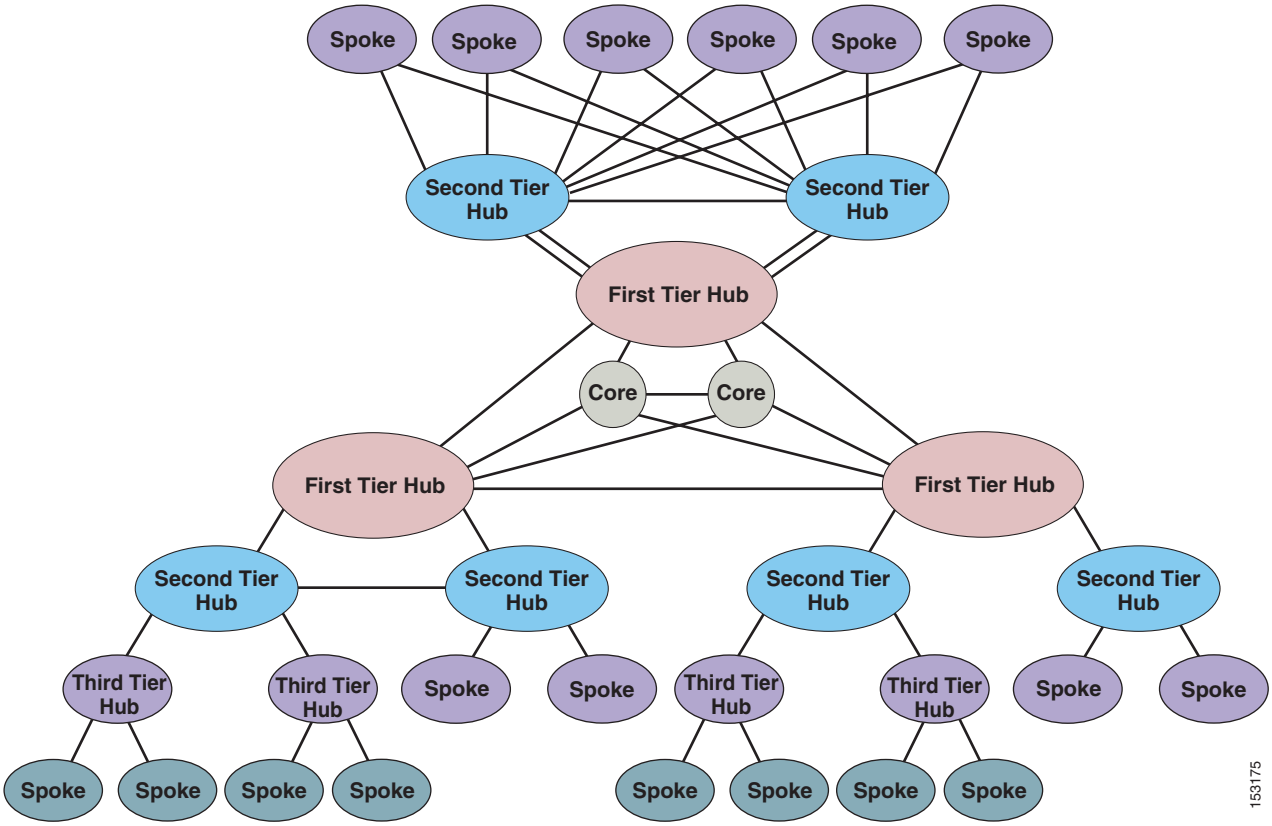
- The route list and route group construct for calls across multiple Cisco Unified CallManager clusters
- The automated alternate routing (AAR) feature for calls within a Cisco Unified CallManager cluster (For more information on AAR, see Automated Alternate Routing, page 10-22.)

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 9-28 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 9-28 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

Cisco Unified CallManager 5.0 introduces support for topology-aware call admission control using RSVP-enabled locations, therefore Cisco recommends upgrading to this release for any deployment using generic topologies.

However, even with Cisco Unified CallManager 4.*x* releases, you can introduce topology-aware call admission control for calls across different clusters by using the Cisco IP-to-IP Gateway, as described in the section on Cisco IOS Gatekeeper and IP-to-IP Gateway with RSVP, page 9-17.

Depending on the deployment details, it might prove unfeasible to add IP-IP gateway functionality to all sites within a large network. However, assuming that certain parts of the network can be simplified to a hub-and-spoke topology (either by altering the network connectivity or by making assumptions based on the available bandwidth), it is possible to combine the topology-unaware call admission control mechanisms such as Cisco Unified CallManager static locations and Cisco IOS gatekeeper zones with topology-aware call admission control provided by RSVP across IP-to-IP gateways in the same network.

The remainder of this section presents a detailed case study based on a hypothetical large customer network whose topology cannot be simplified to a hub-and-spoke. This example combines the following call admission control mechanisms to provide an end-to-end solution:

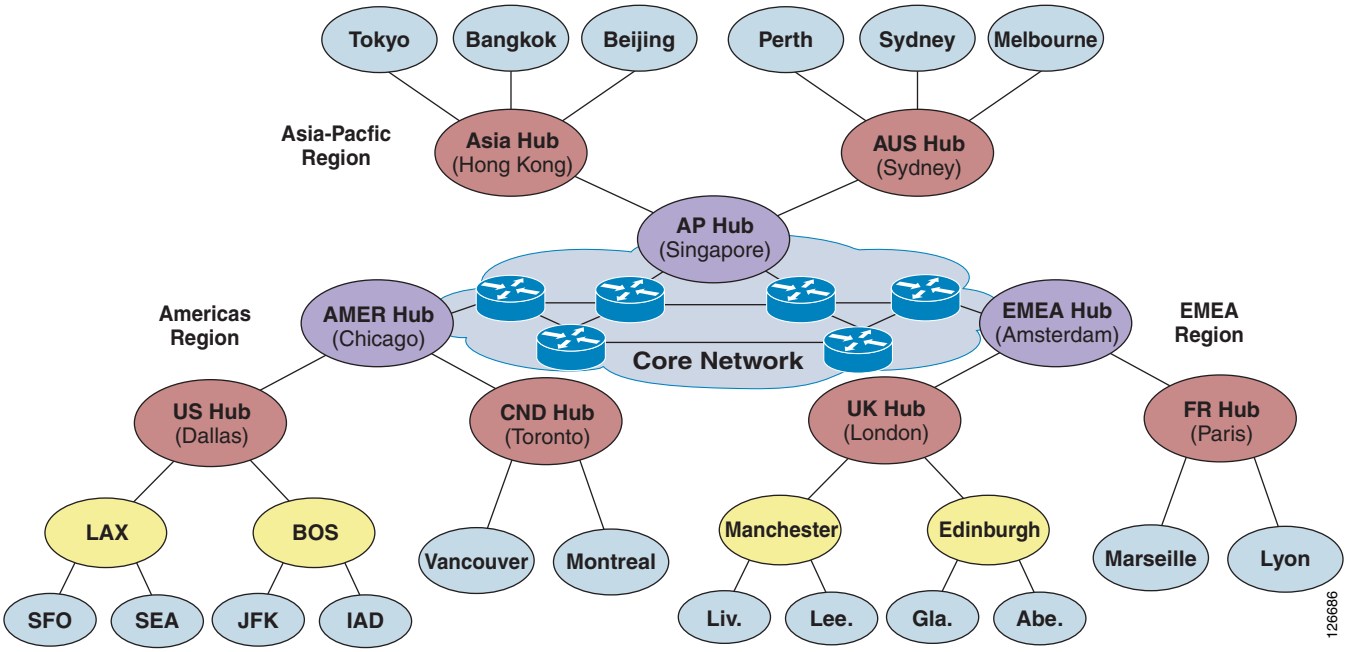
- Cisco Unified CallManager locations
- Cisco IOS Gatekeeper
- RSVP (using the Cisco IP-to-IP Gateway)

Case Study

Figure 9-29 shows the high-level topology of a hypothetical large customer network. The customer sites are divided into three main regions: the Americas (AMER), Europe-Middle East-Africa (EMEA), and Asia-Pacific (AP). Within each of these regions, the network topology is a three-tiered hub-and-spoke,

with a regional hub and several country hubs, each of which can have one or two lower levels of smaller sites. The core network interconnects the three regions across an arbitrary network topology, possibly containing multiple equal-cost paths between any two destinations.

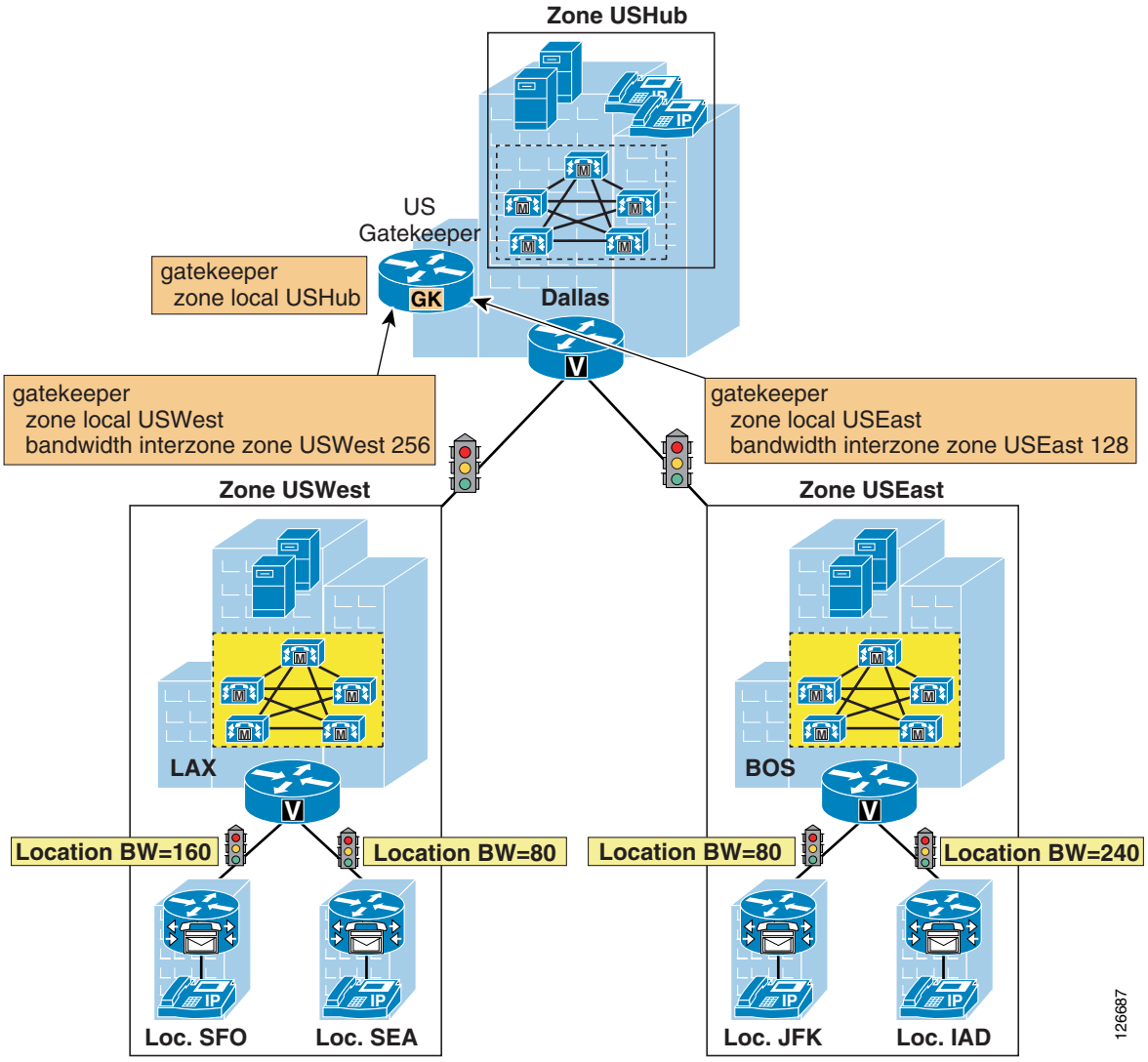
Figure 9-29 Hypothetical Customer Network



This complex topology illustrates where to apply the various call admission control mechanisms and how to combine them. To provide a better understanding of the overall solution, this section analyzes different portions of the network separately.

First, consider the AMER region. As illustrated in Figure 9-29, there are two country hubs that connect to the region hub: the US hub site, located in Dallas, and the Canada hub site, located in Toronto. Figure 9-30 shows how to implement call admission control for the US portion of the network.

Figure 9-30 Call Admission Control for the First Two Levels of the Customer Network in the US



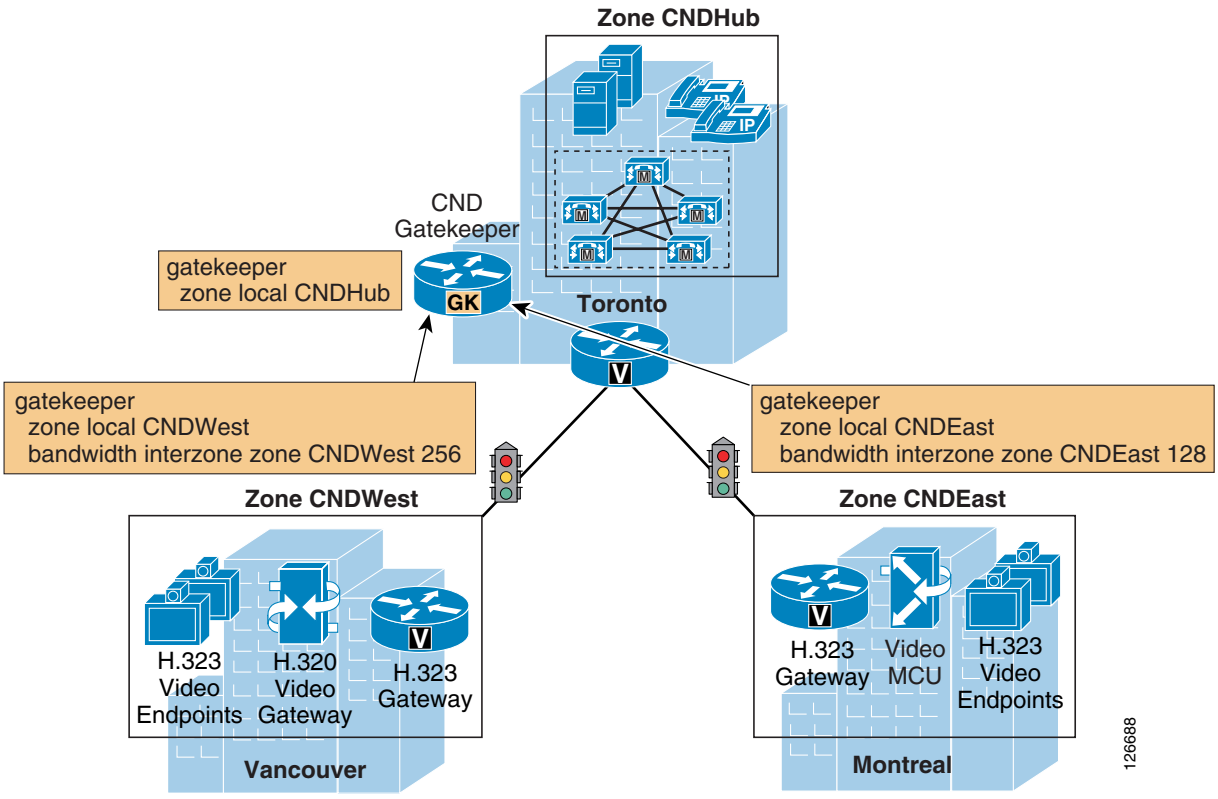
A Cisco IOS gatekeeper is placed at the US hub site in Dallas. A number of first-level spoke sites (up to 100 per gatekeeper) can connect to this hub site. In the example shown in Figure 9-30, these sites are Los Angeles (LAX) and Boston (BOS). By defining a local zone for each site in the gatekeeper (USHub, USWest, and USEast in the example), you can perform call admission control on the links between the first-level spoke sites and the US hub site using the Cisco IOS gatekeeper command **bandwidth interzone zone** *zone-name bandwidth-value*. This command ensures that the bandwidth used by voice and video calls in and out of each zone does not exceed the configured value.

In this figure, the first-level spoke sites (LAX and BOS) each host a Cisco Unified CallManager cluster. Because Cisco Unified CallManager uses its own call admission control mechanism (the locations construct), it is possible to add a number of second-level spoke sites (up to 500 per cluster) to each of the first-level spoke sites. In this example, two additional sites are controlled by each Cisco Unified CallManager cluster: San Francisco (SFO) and Seattle (SEA) for the Los Angeles cluster and New York (JFK) and Washington, DC (IAD), for the Boston cluster. The bandwidth on the links between these second-level spoke sites and their respective first-level site is therefore controlled by the Cisco Unified CallManager locations.

Note As far as the gatekeeper is concerned, the second-level spoke sites are considered to belong to the same zone as their "parent" first-level spoke site because they are controlled by the same Cisco Unified CallManager cluster. This analysis is correct because calls from any of the second-level spoke sites directed to an on-net destination not controlled by the Cisco Unified CallManager cluster will have to traverse the link between the "parent" first-level spoke site and the country hub site.

Figure 9-31 shows the Canada (CND) sites, which provide another example of how call admission control is implemented at this level in the network.

Figure 9-31 Call Admission Control for the First Level of the Customer Network in Canada



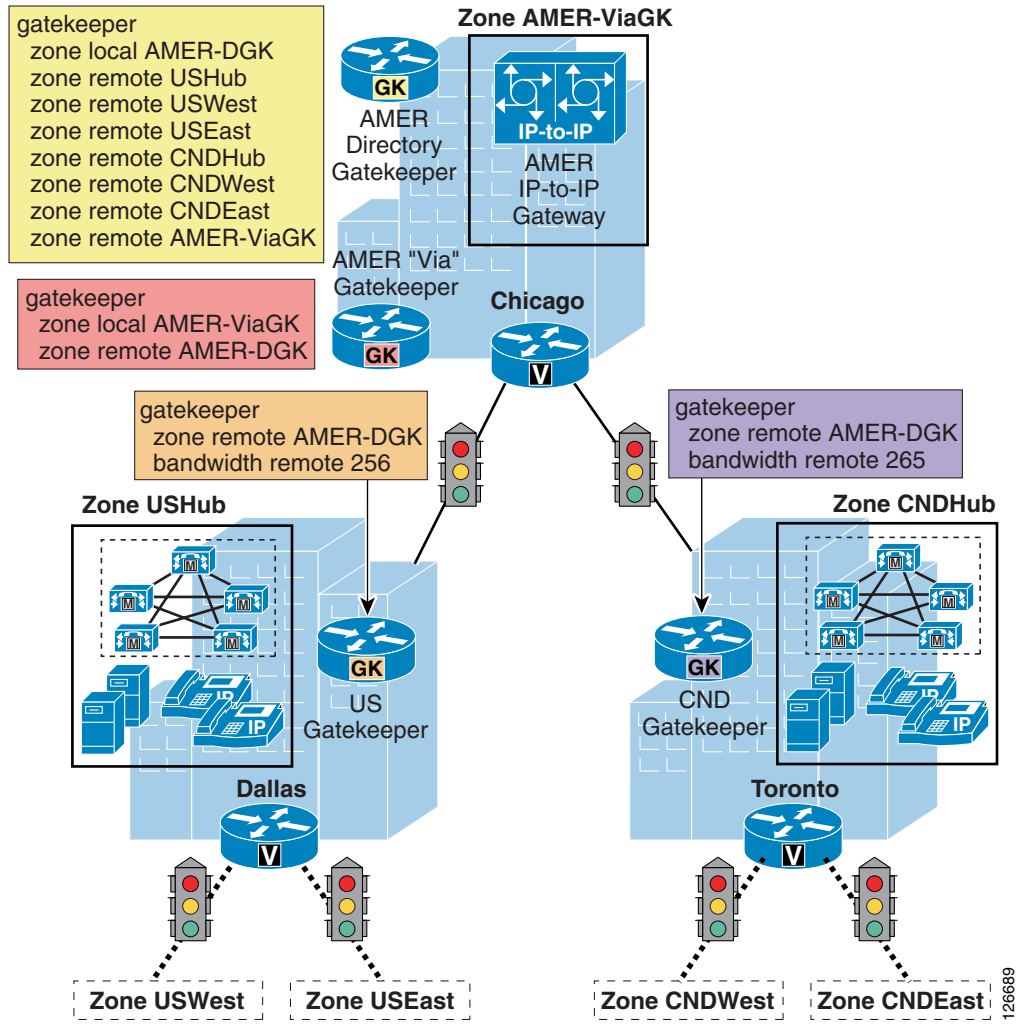
Another Cisco IOS gatekeeper is placed at the Canada hub site in Toronto, and a number of spoke sites connect to the hub site. In the example, the two spoke sites are Vancouver and Montreal. Note that these spoke sites do not contain a Cisco Unified CallManager cluster but do contain a variety of H.323-based devices such as voice gateways, video endpoints, Multipoint Control Units (MCUs), and H.320 video gateways. For each site, a local zone (CNDHub, CNDWest, and CNDEast in the example) is defined within the gatekeeper. The bandwidth on the links between the Canada hub site and the spoke sites can again be controlled using the Cisco IOS gatekeeper command **bandwidth interzone zone** *zone-name bandwidth-value*. This command ensures that the bandwidth used by voice and video calls in and out of each zone does not exceed the configured value.

Because the H.323 devices at the spoke sites cannot perform call admission control, it is not possible to add a second level of spoke sites in this case. However, because all these devices can also be controlled by Cisco Unified CallManager, it would be possible to add a Cisco Unified CallManager cluster to the first-level spoke sites and distribute these H.323 devices across multiple second-level spoke sites.

Also note that, within the sites controlled by the same gatekeeper, it is possible to mix and match Cisco Unified CallManager spoke sites and H.323 spoke sites.

Now consider the next level up in the network hierarchy. Figure 9-32 illustrates the call admission control mechanisms used between the AMER region hub site in Chicago and the two country hub sites for the US and Canada, located in Dallas and Toronto respectively.

Figure 9-32 Call Admission Control for the Second Level of the Customer Network (Region Hubs)



This part of the network is also a hub-and-spoke, but in this case the approach used to perform call admission control is different from that used between the country hub sites and the first-level spoke sites.

As discussed earlier, the country hub sites contain a Cisco IOS gatekeeper that controls bandwidth to the first-level spoke sites using local gatekeeper zones and the **bandwidth interzone** Cisco IOS command. The AMER region hub site contains a gatekeeper used as a directory gatekeeper, whose main purpose is to handle call routing between the country gatekeepers.

To control bandwidth usage on the links between the country hub sites and the region hub site, you can use the **bandwidth remote** *bandwidth-value* Cisco IOS command on the two gatekeepers located at the country hub sites. This command ensures that the bandwidth used by voice and video calls between all the local zones and all non-local zones does not exceed the configured value. Because all calls to

destinations that do not belong to a local zone must traverse the link between the country hub site and the region hub site, you can use the **bandwidth remote** command to perform call admission control effectively on this link.

In Figure 9-32, also notice the presence of a via-zone gatekeeper at the AMER region hub site. The directory gatekeeper is configured (configuration not shown in Figure 9-32) to route calls to destinations in other regions (such as AP and EMEA) to the AMER via-zone gatekeeper.

Next consider the core section of the network, which connects the three region hub sites. Figure 9-33 and Figure-35 illustrate how call routing and call admission control are performed across the core network.

Figure 9-33 Call Routing Across the Core of the Customer Network Using Via-Zone Gatekeepers

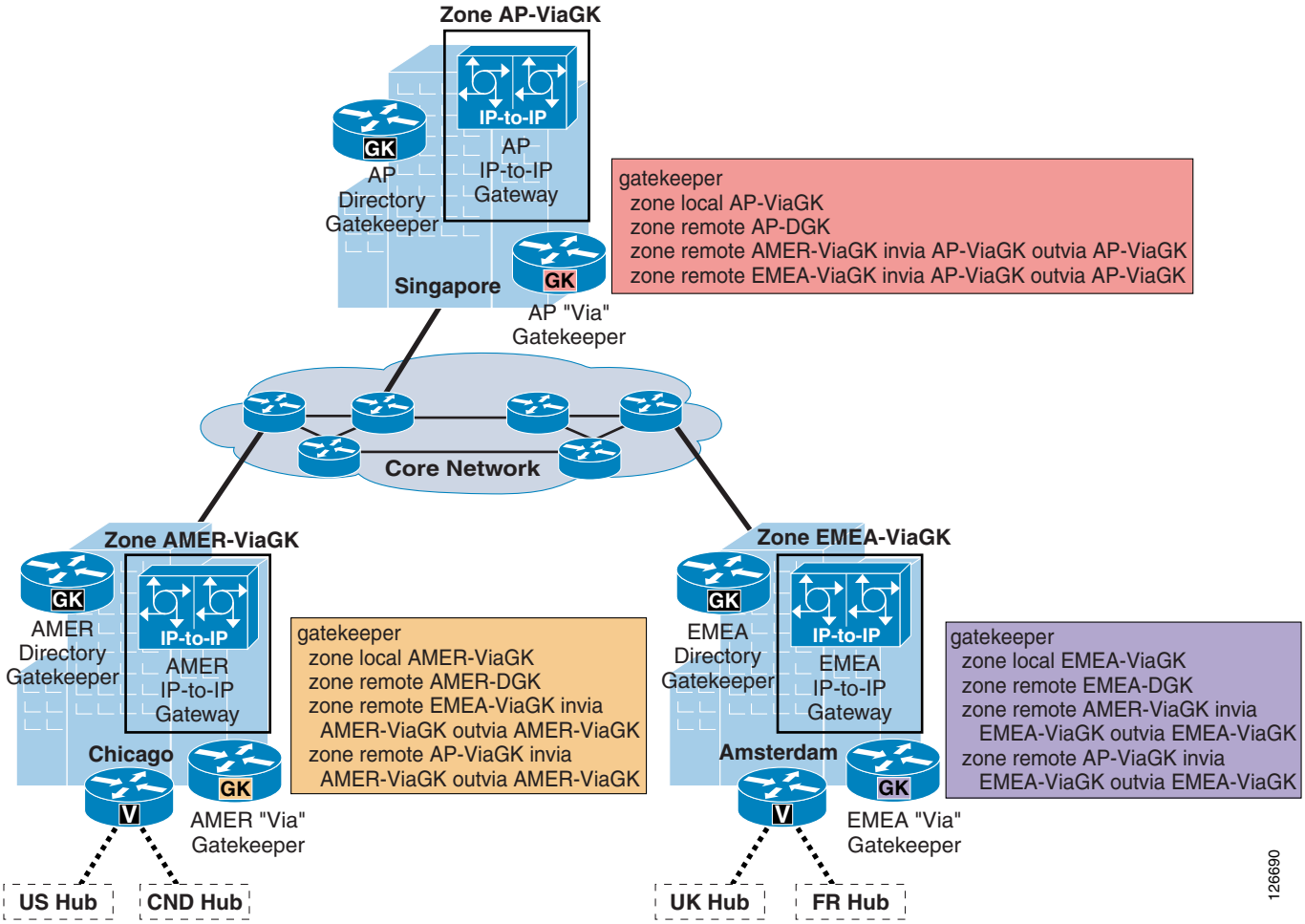


Figure 9-33 shows partial configurations of the three via-zone gatekeepers located in each of the three region hub sites. Each region's via-zone gatekeeper is configured so that all calls destined to or coming from the other two regions will cause the gatekeeper to insert an IP-to-IP gateway into the call.

For example, the configuration for the AMER region's via-zone gatekeeper defines the following zones:

- A single local zone called AMER-ViaGK
- The remote zone AMER-DGK, used to reach the AMER directory gatekeeper and, hence, all destinations within the AMER region

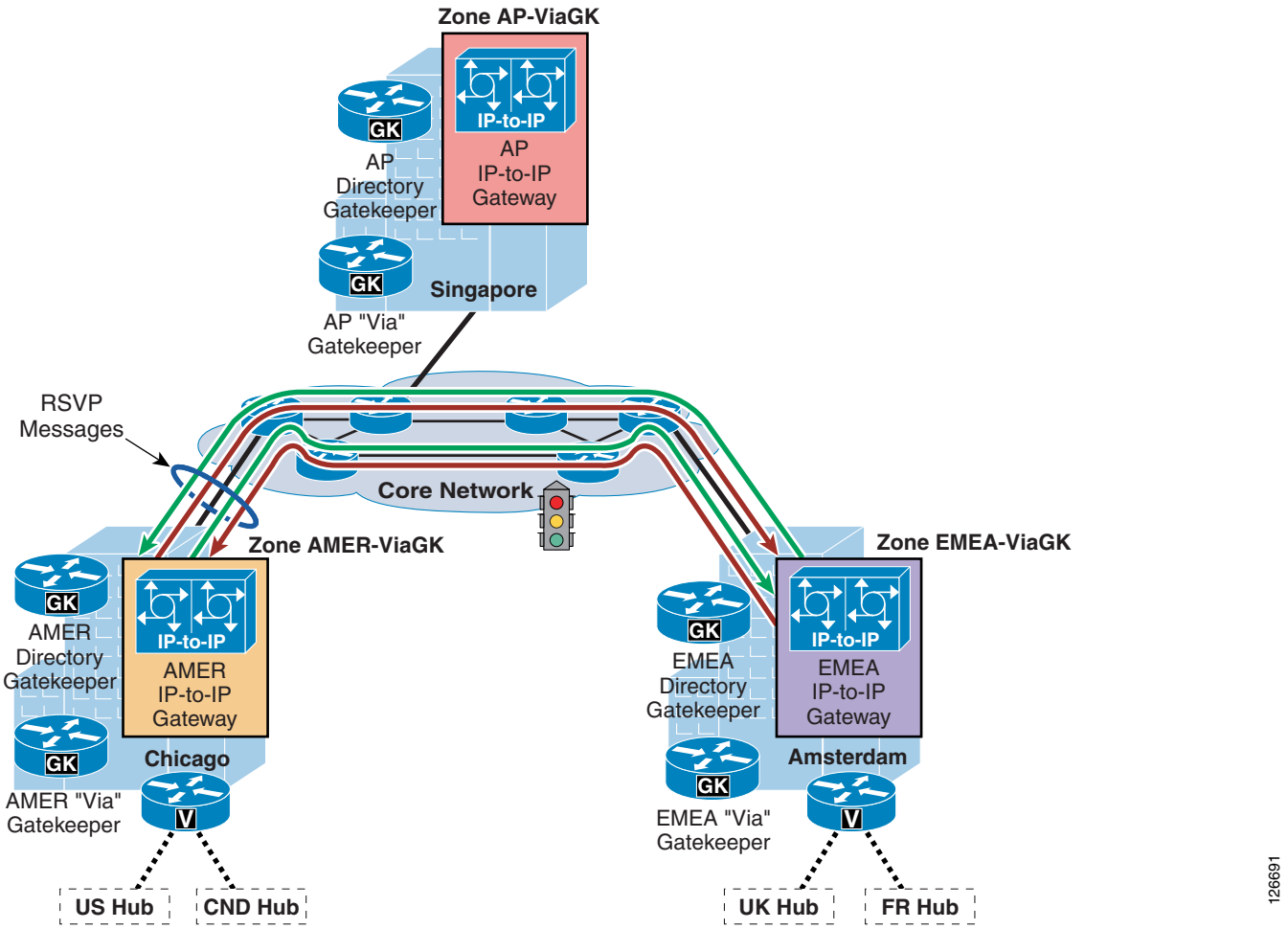
- The remote zone EMEA-DGK, used to reach destinations in the EMEA region
- The remote zone AP-DGK, used to reach destinations in the AP region

The definitions of the EMEA-DGK and AP-DGK remote zones also specify that the local zone AMER-ViaGK should be used as the **invia** and **outvia** zone to reach these destinations. This means that all calls destined to or coming from these remote zones will cause the AMER via-zone gatekeeper to insert an IP-to-IP gateway into the call. The specific IP-to-IP gateway resource will be selected from among all those registered with the gatekeeper in the local zone AMER-ViaGK.

Note Inserting an IP-to-IP gateway into a call effectively splits the call into two call legs. In this example, because an IP-to-IP gateway is inserted at each region hub site, a call between an endpoint in the AMER region and one in the EMEA region would consist of three call legs: the first leg from the AMER endpoint to the AMER IP-to-IP gateway, the second leg between the AMER IP-to-IP gateway and the EMEA IP-to-IP gateway, and the third leg between the EMEA IP-to-IP gateway and the EMEA endpoint.

Figure 9-34 illustrates the call admission control mechanism used in the core network.

Figure 9-34 RSVP-Based Call Admission Control Across the Core of the Customer Network



The IP-to-IP gateways are inserted into the call flow by the via-zone gatekeepers whenever the call needs to traverse the core network. Each IP-to-IP gateway is in turn configured to use the via-zone gatekeeper to route its calls and to use RSVP call admission control for the calls it originates and terminates across the core network. Because RSVP reservations are unidirectional, each voice call generates two RSVP reservations, one for each direction.

Figure 9-34 shows a call between the AMER IP-to-IP gateway and the EMEA IP-to-IP gateway, assuming that multiple paths exist between the two regional hubs across the core network. If the routes taken are asymmetrical, the two RSVP reservations will travel along different paths. However, for a given reservation, the RSVP messages will always travel along the same path because RSVP uses reverse hop-by-hop routing to establish the reservation. RSVP ensures that the bandwidth used by voice and video calls across the core network does not exceed the configured values on each router's interface.

In summary, end-to-end call admission control is achieved by combining different techniques in different portions of the network.

Referring to the network depicted in Figure 9-29, for example, a call between two IP phones located in San Francisco in the US and Liverpool in the UK would use the following mechanisms:

- Cisco Unified CallManager locations for the link between San Francisco and Los Angeles
- Cisco IOS gatekeeper **bandwidth inter-zone** command for the link between Los Angeles and Dallas
- Cisco IOS gatekeeper **bandwidth remote** command for the link between Dallas and Chicago
- RSVP between Chicago and Amsterdam (the two regional hub sites)
- Cisco IOS gatekeeper **bandwidth remote** command for the link between Amsterdam and London (the country hub site for the UK)
- Cisco IOS gatekeeper **bandwidth inter-zone** command for the link between London and Manchester (the hub site for England)
- Cisco Unified CallManager locations for the link between Manchester and Liverpool