Bandwidth Management

Revised: February 6, 2017

This chapter describes the bandwidth management strategy for the Cisco Preferred Architecture (PA) for Enterprise Collaboration.

Certain requirements might put your deployment outside the PA design guidelines and recommendations, in which case you might have to use other documentation such as the Cisco Collaboration SRND and related product documentation for a more customized architecture.

The first part of this chapter provides an architectural overview and introduces some fundamental design concepts, while the second part covers deployment procedures. The Architecture section discusses topics such as identification and classification, queuing and scheduling, provisioning and admission control, using the hypothetical customer topology presented in the examples throughout this document. The next portion of this chapter is the Bandwidth Management Deployment section. The deployment examples in that section help you to understand the implementation of certain design decisions more clearly than an abstract discussion of concepts can. The order of the topics in the Bandwidth Management Deployment section follows the recommended order of configuration.

What's New in This Chapter

Table 8-1 lists the topics that are new in this chapter or that have changed significantly from previous releases of this document.

Table 8-1 New or Changed Information Since the Previous Release of This Document

New or Revised Topic	Described in:	Revision Date
Cisco Meeting Server replaced Cisco TelePresence Conductor and TelePresence Server	Various sections of this chapter	February 6, 2017
DSCP settings for application servers	Application Server QoS, page 8-49	February 6, 2017
Minor updates for Cisco DX70 and DX80 endpoints with CE Software	Various sections of this chapter	February 6, 2017

Core Components

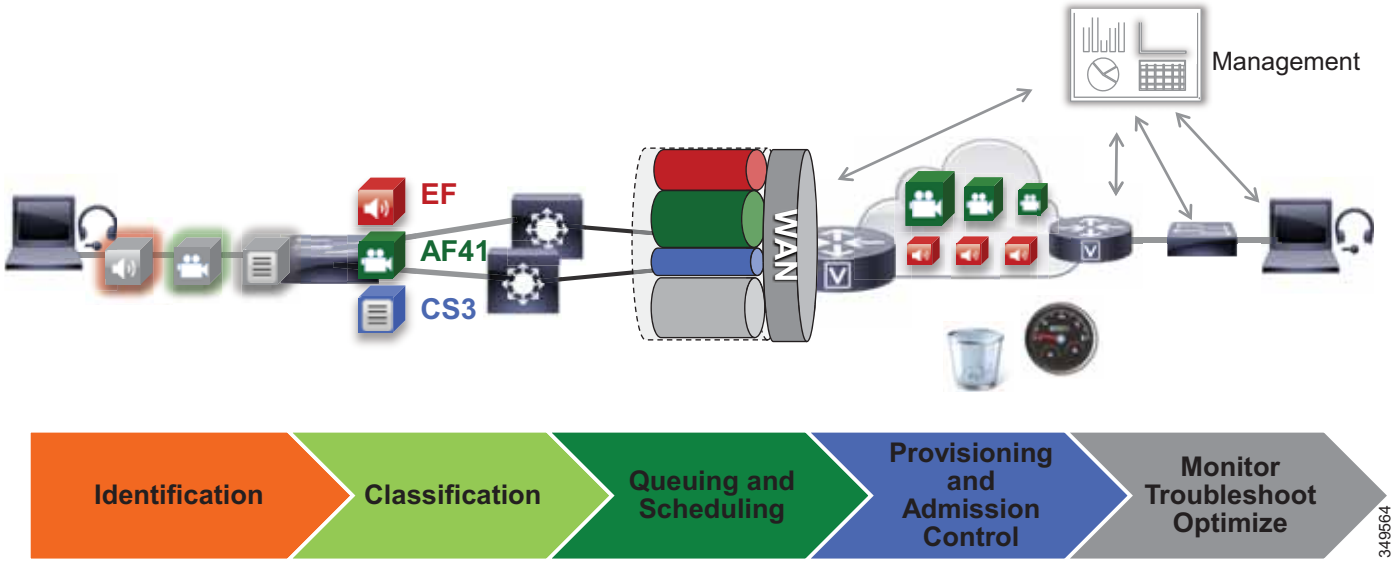
The Quality of Service (QoS) architecture contains these key components:

- Cisco Unified Communications Manager
- Cisco endpoints
- Cisco Expressway
- Cisco Unity Connection
- Cisco Meeting Server
- Network infrastructure:
  - Cisco routers
  - Cisco switches

Figure 8-1 illustrates the design approach to QoS used in the Cisco PA for Enterprise Collaboration. This approach consists of the following phases:

- **Identification and classification** — Refers to concepts of trust and techniques for identifying media and call signaling for endpoints and applications. It also includes the process of mapping the identified traffic to the correct DSCP to provide the media and signaling with the correct per-hop behavior end-to-end across the network.
- **Queuing and scheduling** — Consists of general WAN queuing and scheduling, the various types of queues, and recommendations for ensuring that collaboration media and signaling are correctly queued on egress to the WAN.
- **Provisioning and admission control** — Refers to provisioning the bandwidth in the network and determining the maximum bit rate that groups of endpoints will utilize. This is also where call admission control can be implemented in areas of the network where it is required.
- **Monitoring, troubleshooting, and optimization** — Ensures the proper operation and management of voice and video across the network. Cisco Prime Collaboration offers a suite of tools to perform these functions. Monitoring, troubleshooting and optimization are not covered in the Preferred Architectures but are part of the overall approach.

Figure 8-1 Architecture for Bandwidth Management



Recommended Deployment

- Identify media and SIP signaling traffic from the endpoints.
- Classify and mark traffic at the access switch edge.
  - Mark all audio with Expedited Forwarding class EF (includes all audio of voice-only and video calls).
  - Mark all critical desktop and room system video with an Assured Forwarding class of AF41.
  - Mark all Jabber, mobile and remote access (MRA), and edge video with an Assured Forwarding class of AF42.
  - Mark all call signaling with CS3.
  - Configure QoS on all media originating and terminating applications and MCUs across the solution.
- Apply simplified WAN edge policies for identifying, classifying, marking, and queuing collaboration traffic:
  - WAN edge ingress re-marking policy
  - WAN edge egress queuing and scheduling policy
- Group video endpoints into classes according to maximum video bit rate, to limit bandwidth consumption based on endpoint type and usage within the solution.
- Deploy Enhanced Locations Call Admission Control and limit video calling only in areas of the network where bandwidth resources are restricted.

Key Benefits

This deployment of bandwidth management provides the following benefits:

- Provides prescriptive recommendations to simplify deployment with a simplified QoS architecture
- Makes more efficient use of network resources
- Supports mobile and multi-media Collaboration devices
- Takes into account "unmanaged" network segments (Internet)
- Is "future-proof" because it facilitates introduction of new services, features, and endpoints

Architecture

Bandwidth management is about ensuring the best possible user experience end-to-end for all voice and video endpoints, clients, and applications in the Collaboration solution. This chapter provides a holistic approach to bandwidth management, incorporating an end-to-end Quality of Service (QoS) architecture with call admission control and video rate adaptation and resiliency mechanisms to ensure the best possible user experience for deploying pervasive video over managed and unmanaged networks.

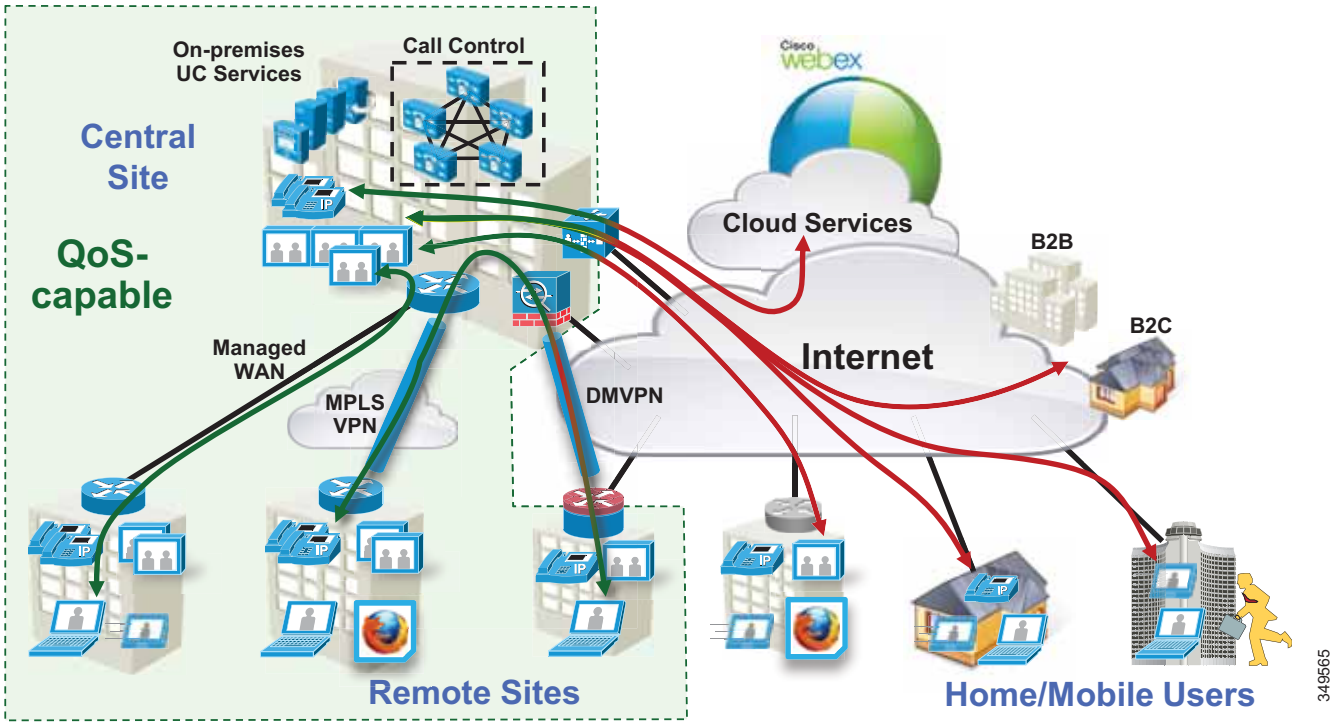
This section starts with a discussion of collaboration media, the differences between audio and video, and the impact this has on the network. Next this section outlines an end-to-end QoS architecture for collaboration that includes: identification and classification of collaboration media and SIP signaling for endpoints, clients, and applications; WAN queuing and scheduling; and bandwidth provisioning and admission control. The next section on Bandwidth Management Deployment explains the steps involved in implementing this architecture in both the collaboration and network architecture.

Note The Network Infrastructure chapter of the Collaboration SRND lays the foundation for QoS in the LAN and WAN. If you are unfamiliar with the concepts of QoS, it is important to read that chapter to fully understand the concepts discussed therein. This chapter assumes an understanding of QoS.

Introduction

In this Preferred Architecture, usage of the Internet and cloud-based services such as WebEx are an important aspect of the solution, which means that some of the Collaboration infrastructure is located outside of the managed enterprise network and located in the cloud. The enterprise office connectivity options also range from remote sites and mobile users connected over managed leased lines directly connected to MPLS or other technologies, to being connected over the Internet through technologies such as Dynamic Multipoint VPN (DMVPN), for example. Figure 8-2 illustrates the convergence of a traditional on-premises Collaboration solution in a managed (capable of QoS) network with cloud services and sites located over an unmanaged (not capable of QoS) network such as the Internet. On-premises remote sites are connected over this managed network, where administrators can prioritize collaboration media and signaling with QoS, while other remote sites and branches connect into the enterprise over the Internet, where collaboration media and signaling cannot be prioritized or prioritized only outbound from the site. Many different types of mobile users and teleworkers also connect over the Internet into the on-premises solution. So the incorporation of the Internet as a source for connecting the enterprise with remote sites, home and mobile users, as well as other businesses and consumers, has an important impact on bandwidth management and user experience.

Figure 8-2 Managed versus Unmanaged Network



This section presents a strategy for leveraging smart media techniques in Cisco video endpoints, building an end-to-end QoS architecture, and using the latest design and deployment recommendations and best practices for managing bandwidth to achieve the best user experience possible based on the network resources available and the various types of networks that collaboration media traverse.

Collaboration Media

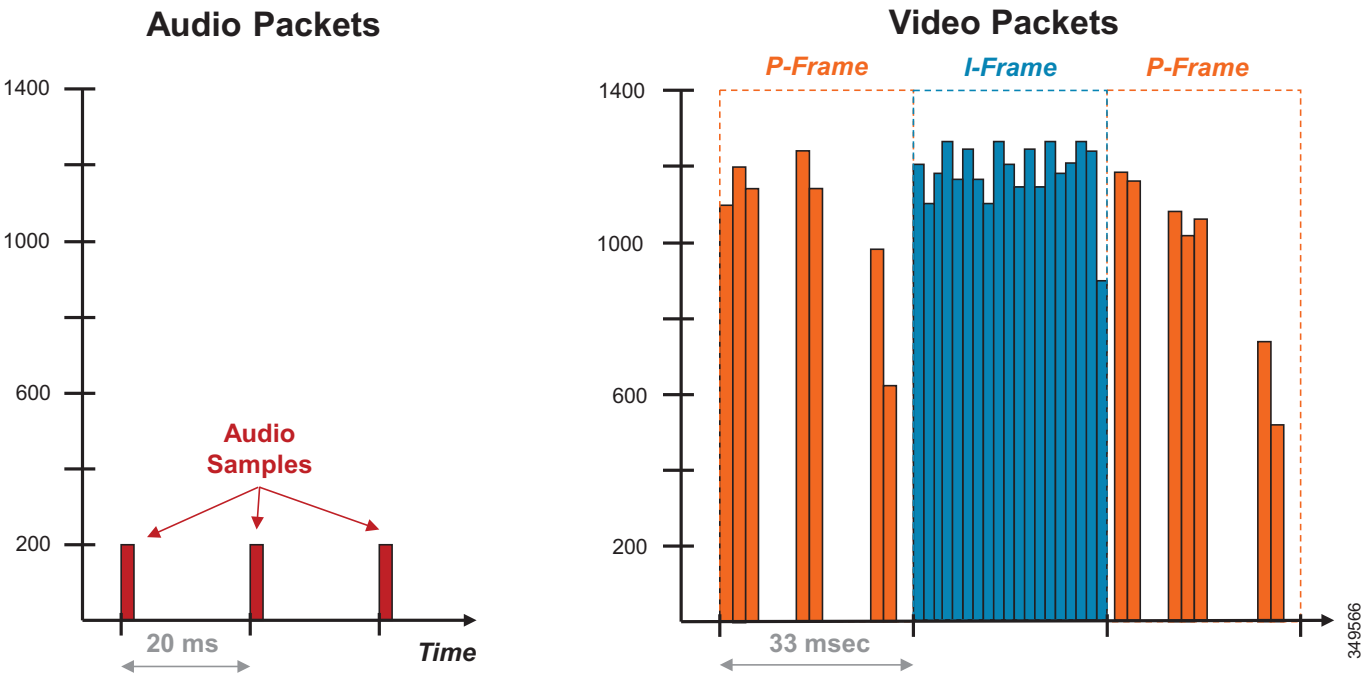
This section covers the characteristics of audio and video streams in real-time media as well as the smart media techniques that Cisco video endpoints employ to ensure high fidelity video in the event of packet loss, delay, and jitter.

Audio versus Video

Voice and video are often thought of as quite similar, and although they are both real-time protocol (RTP) applications, the similarities stop there. Voice is generally considered well behaved because each packet is a fixed size and fixed rate. Video frames are spread over multiple packets that travel as a group. Because one lost packet can ruin a P-frame, and one bad P-frame can cause a persistent artifact, video generally has a tighter loss requirement than audio. Video is asymmetrical. Voice can also be asymmetrical but typically is not. Even on mute, an IP phone will send and receive the same size flow.

Video increases the average real-time packet size and has the capacity to quickly alter the traffic profile of networks. Without planning, this could be detrimental to network performance. Figure 8-3 shows the difference between a series of audio packets and video packets sent during a specific time interval.

Figure 8-3 Audio versus Video



As illustrated in Figure 8-3, the audio packets are the same size, sent at exactly the same time intervals, and represent a very smooth stream. Video, on the other hand, sends a larger group of packets over fixed intervals and can vary greatly from frame to frame. Figure 8-3 shows the difference in the number of packets and packet sizes for an I-Frame compared to P-frames. This translates to a stream of media that is very bursty in nature when compared to audio. This burstiness is illustrated in Figure 8-4, which shows the bandwidth profile over time of an HD video stream. Note the large bursts when I-Frames are sent.

Figure 8-4 Bandwidth Usage: High-Definition Video Call

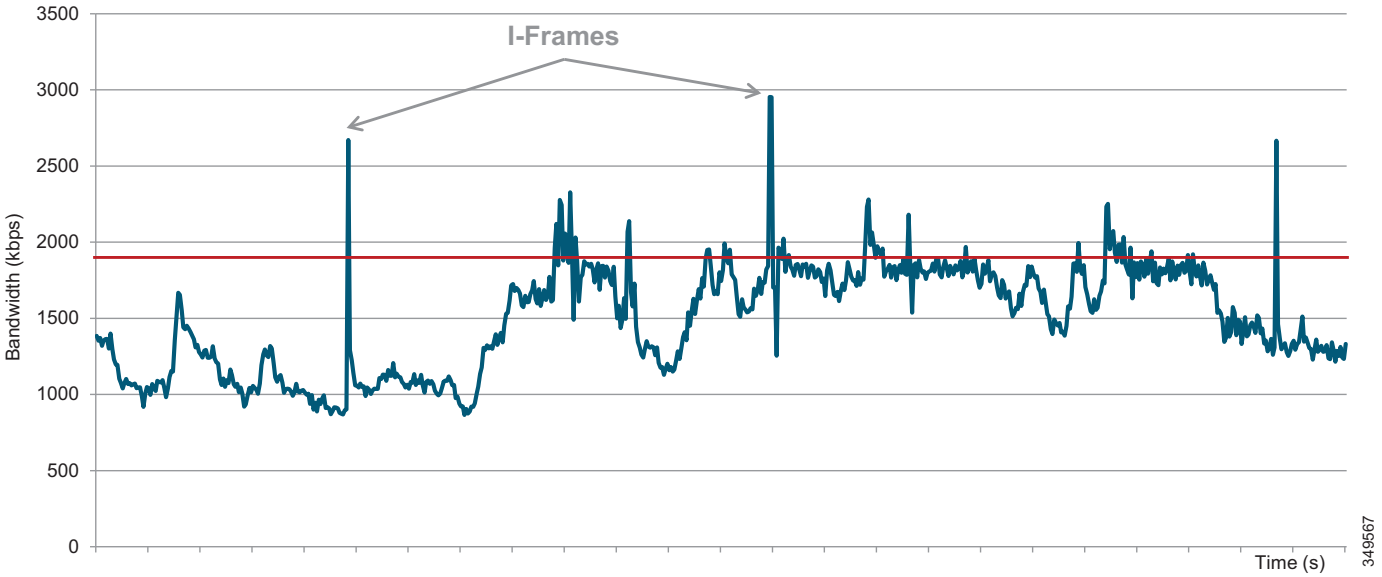
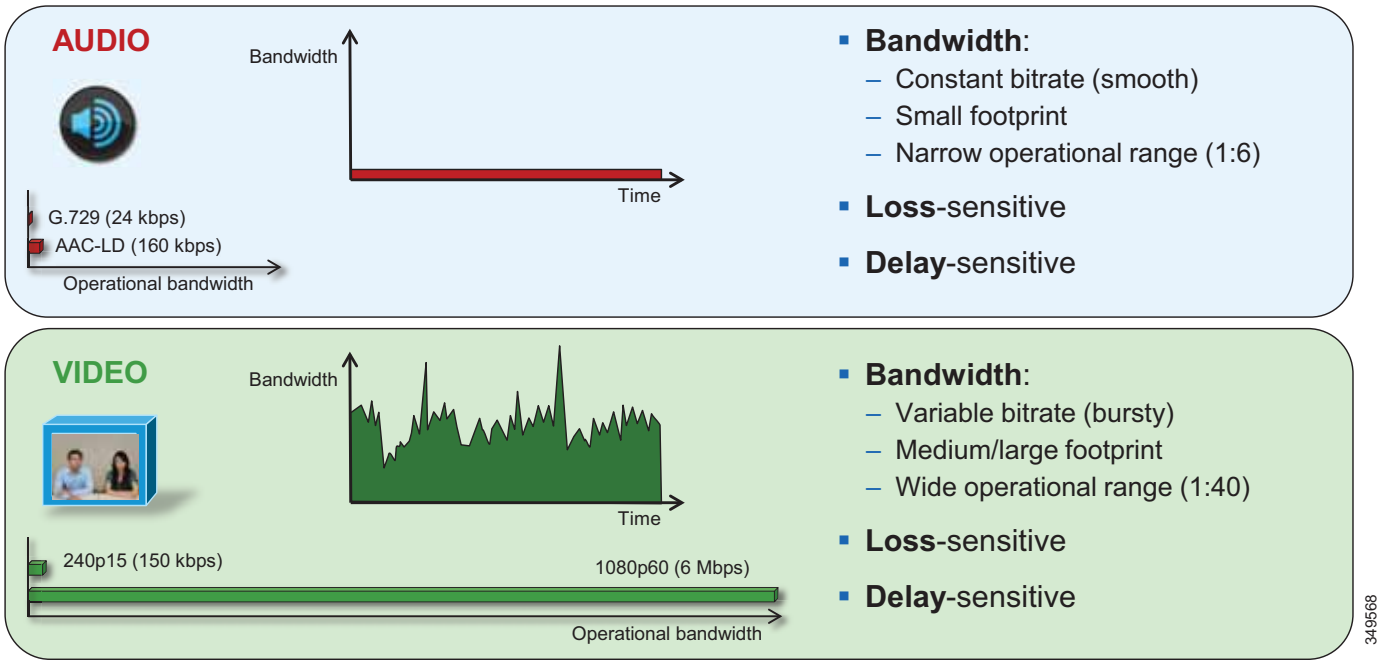


Figure 8-4 shows an HD video call at 720p at 30 fps and 1920 kbps (1792 kbps video + 128 kbps audio). The red line indicates the average bit rate for the duration of the call.

While audio and video are both transported over UDP and are sensitive to loss and delay, they are quite different in their network requirements and profile. As shown in Figure 8-5, audio is a constant bit rate, has a smaller footprint compared to video, and has a narrower operational range of 1:6 ratio when comparing the lowest bit rate audio codec to one of the highest bit rate codecs. Video, on the other hand, has a variable bit rate (is bursty), has a medium to large footprint when compared to audio, and has a wide operational range of 1:40 (250p at 15 fps up to 1080p at 60 fps).

Figure 8-5 Video Traffic Requirements and Profiles



The main point here is that audio and video, while similar in transport and sensitivity to loss and delay, are quite different in the methods employed to manage their bandwidth requirements in the network. Also, while video is pertinent to a full collaboration experience, audio is critical. For example, during a video call, if video is lost or distorted due to a network outage or some other network related event, communication can continue provided that audio is not lost during that outage. This is a critical concept in bandwidth management in the PA.

"Smart" Media Techniques (Media Resilience and Rate Adaptation)

When deploying video pervasively across an organization, administrators will inevitably encounter insufficient bandwidth to handle the load of video required during the busy hour in some bottleneck areas of the Wide Area Network (WAN). In light of this it is important to prioritize video correctly, to ensure that audio is not affected by any video packet loss that may occur, and to ensure that certain types of video can leverage video rate adaptation to manage the amount of bandwidth used during times of congestion. The media resilience and rate adaptation techniques allow for an optimized video experience in the face of congestion and packet loss over managed and unmanaged networks, but that is not all. These techniques, when used as a strategy coupled with QoS, can offer the ability for an organization to

deploy video pervasively by allowing the endpoints to reduce their bit rate and thus their bandwidth utilization during congestion and packet loss, while also allowing the endpoints to increase their bit rate and thus bandwidth utilization during more idle times of the day outside of the busy hour, thereby maximizing video quality.

Every Cisco video endpoint employs a number of smart media techniques to avoid network congestion, recover from packet loss, and optimize network resources. The following smart media techniques are some of the techniques employed by Cisco video endpoints:

- Media resilience techniques
  - Encoder pacing
  - Gradual Decoder Refresh (GDR)
  - Long-Term Reference Frame (LTRF) with Repair
  - Forward Error Correction (FEC)
- Rate adaptation

Media Resilience Techniques

Table 8-2 summarizes examples of the media resilience techniques supported on various Cisco video endpoints.

Table 8-2 Media Resilience Support in Cisco Collaboration Video Endpoints

Video Endpoint or Bridge	Encoder Pacing	Rate Adaptation	FEC	LTRF Repair
Cisco IP Phone 8800 Series	Yes	No	No	No
Cisco Jabber	Yes	Yes	Yes	Yes
Cisco DX70 and DX80 (CE Software)	Yes	Yes	No	Yes
Cisco TelePresence MX Series	Yes	Yes	Yes	Yes
Cisco TelePresence SX Series	Yes	Yes	Yes	Yes
Cisco TelePresence IX Series	No	No	No	No
Cisco Meeting Server	Yes	Yes	Yes	Yes

Encoder Pacing

Encoder pacing is a simple technique used to spread the packets as evenly as possible in order to smooth out the peaks of the bursts of bandwidth.

Gradual Decoder Refresh (GDR)

GDR is a method of gradually refreshing the picture over a number of frames, giving a smoother, less bursty bit stream.

Long Term Reference Frame (LTRF)

A Long Term Reference Frame (LTRF) is a reference frame stored in the encoder and decoder, which allows the video endpoints to recover more efficiently from packet loss with less bandwidth utilization over the network path in order to resend lost frames.

Forward Error Correction (FEC)

Forward error correction (FEC) provides redundancy to the transmitted information by using a predetermined algorithm. The redundancy allows the receiver to detect and correct a limited number of errors occurring anywhere in the message, without the need to ask the sender for additional data. FEC gives the receiver an ability to correct errors without needing a reverse channel (such as RTCP) to request retransmission of data, but this advantage is at the cost of a fixed higher forward channel bandwidth (more packets sent). FEC protects the most important data (typically the repair P-frames) to make sure the receiver is receiving those frames. The endpoints do not use FEC on bandwidths lower than 768 kbps, and there must also be at least 1.5% packet loss before FEC is introduced. Endpoints typically monitor the effectiveness of FEC; and if FEC is not efficient, they make a decision not to do FEC.

Rate Adaptation

Rate adaptation or dynamic bit rate adjustments adapt the call rate to the variable bandwidth available, down-speeding or up-speeding the video bit rate based on the packet loss condition. An endpoint will reduce bit rate when it receives messages from the receiver indicating there is packet loss; and once the packet loss has decreased, up-speeding of the bit rate will occur.

Opportunistic Video and Prioritized Audio

Opportunistic Video and Prioritized Audio is a concept and a QoS strategy combined. It consists of defining a number of video endpoints that opportunistically utilize available video bandwidth. During the busy hour these endpoints rate adapt or throttle down their bit rate to accommodate limited bandwidth availability, all the while not impacting prioritized video. During the idle hour they utilize available bandwidth to optimize video quality by increasing the video bit rate. Prioritized audio for both audio-only and audio of video calls ensures that all audio is prioritized in the network and is thus not impacted by any loss that can occur in the video queues. Prioritizing voice from all types of collaboration media ensures that even during times of extreme congestion when video is experiencing packet loss and adjusting to that loss, the audio streams are not experiencing packet loss and are allowing the user to carry on an uninterrupted audio experience. Prioritizing audio from both voice-only and video calls is a paradigm shift from the previous historic model where audio and video of video calls were always marked with the same QoS. Opportunistic video with prioritized audio maintains an acceptable video experience while simultaneously ensuring that voice media for voice-only and video calls is not compromised. This of course applies to the managed network, since an unmanaged network such as the Internet is not QoS enabled and thus provides no guarantees with regard to packet loss. Nevertheless, the media resiliency and rate adaptation mechanisms also attempt to ensure that media over unmanaged networks has the best possible quality in the face of packet loss, delay, and jitter.

QoS Architecture for Collaboration

Quality of Service (QoS) ensures reliable, high-quality voice and video by reducing delay, packet loss, and jitter for media endpoints and applications. QoS provides a foundational network infrastructure technology, which is required to support the transparent convergence of voice, video, and data networks. With the increasing amount of interactive applications (particularly voice, video, and immersive applications), real-time services are often required from the network. Because these resources are finite, they must be managed efficiently and effectively. If the number of flows contending for such priority resources were not limited, then as these resources become oversubscribed, the quality of all real-time traffic flows would degrade, eventually to the point of futility. "Smart" media techniques, QoS, and admission control ensure that real-time applications and their related media do not oversubscribe the network and the bandwidth provisioned for those applications. These smart media techniques coupled

with QoS and, where needed, admission control, are a powerful set of tools used to protect real-time media from non-real-time network traffic and to protect the network from over-subscription and the potential loss of quality of experience for end users of voice and video applications.

Identification and Classification

QoS Trust and Enforcement

The enforcement of QoS is crucial to any real-time audio, video, or immersive video experience. Without the proper QoS treatment (classification, prioritization, and queuing) through the network, real-time media can potentially incur excessive delay or packet loss, which compromises the quality of the real-time media flow. In the QoS enforcement paradigm, the issue of trust and the trust boundary is equally important. Trust is the permitting or the "trusting" of QoS marking (Layer 2 CoS or Layer 3 IP DSCP) of the traffic by the endpoint or device, to allow the traffic to continue through the network. The trust boundary is the place in the network where the trust occurs. It can occur at any place in the network; however, we recommend enforcing trust at the network edge such as the LAN access ingress or the WAN edge, or both, as is feasible and applicable.

There are three main categories of trust:

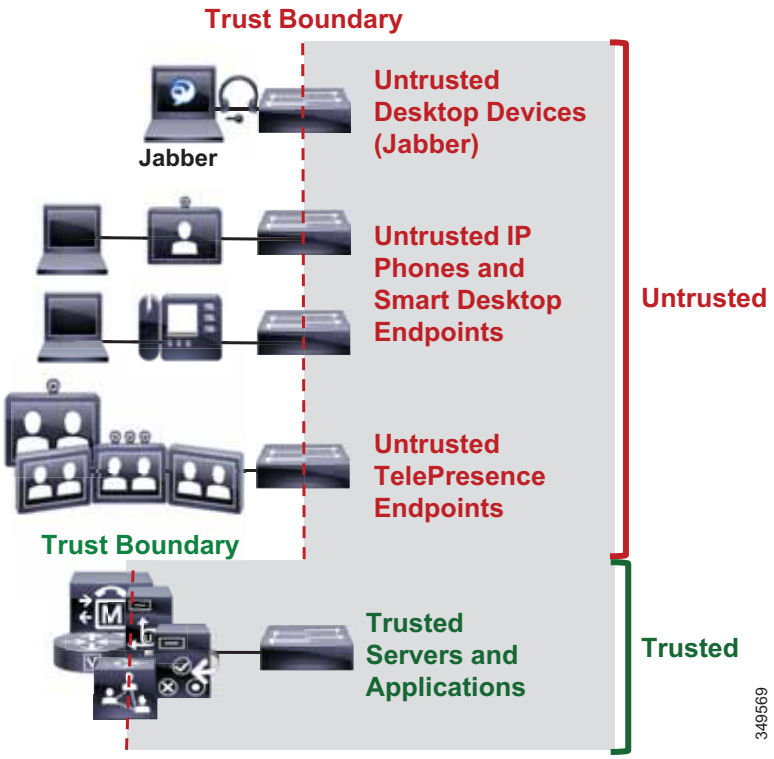
- **Untrusted** — These devices include unsecure PCs, Macs, or hand-held mobile devices running Jabber clients, IP phones, and smart desktop endpoints.
- **Trusted** — These devices can include secure PCs and servers, video conferencing endpoints, analog and video conferencing gateways, PSTN gateways, Cisco Unified Border Element, trusted application servers, and other similar devices.
- **Conditionally trusted** — These devices typically include endpoints that support Cisco Discovery Protocol (CDP). Although Cisco TelePresence, IP Phone and Smart Desktop phones support CDP, they are not conditionally trusted in the PA.

In the PA, trusted and untrusted switch ports are used but conditionally trusted ports are not used. Conditional trust is not recommended in the PA for the following reasons:

- Complexity across a variety of switches — Enabling conditional trust across a variety of switch types can become complex. Some of the older switch types do not trust by default, while newer switches do trust by default. Furthermore, the commands for enabling trust and the process of trust enforcement are different across platforms.
- Even more important is the lack of Layer 3 DSCP re-marking on PC ports of IP phones and smart desktop endpoints. The endpoints re-mark only Layer 2 CoS. Because of this and the inability to correctly re-mark PC traffic at the DSCP level, using access lists to re-mark IP phones and smart desktop endpoints is a preferred method in the PA.
- A single ACL that maps directly to all switch ports is easier to manage than specifying only a limited number of ports for trust.

Figure 8-6 illustrates the types of trust used in the PA, and which devices are trusted and untrusted.

Figure 8-6 Trust Boundaries in the Preferred Architecture



Classification and Marking

This section discusses classification and marking for endpoints.

All Cisco endpoints derive their DSCP marking from Unified CM. Unified CM houses the QoS configuration for endpoints in two places, in the Service Parameters for the CallManager service (**Clusterwide Parameters (System - QoS)**) and in the SIP Profile (applicable only to SIP devices). The SIP Profile configuration of QoS settings overrides the Service Parameter configuration. This allows Unified CM administrators to set different QoS policies for groups of endpoints. Unified CM passes this QoS configuration to the endpoints in a configuration file over TFTP during endpoint registration. This configuration file contains the QoS parameters as well as a number of other endpoint specific parameters. For QoS purposes there are two categories of video endpoints: TelePresence endpoints (any endpoint with TelePresence in the phone type name) and all other non-TelePresence video endpoints (referred to as "UC Video endpoints").

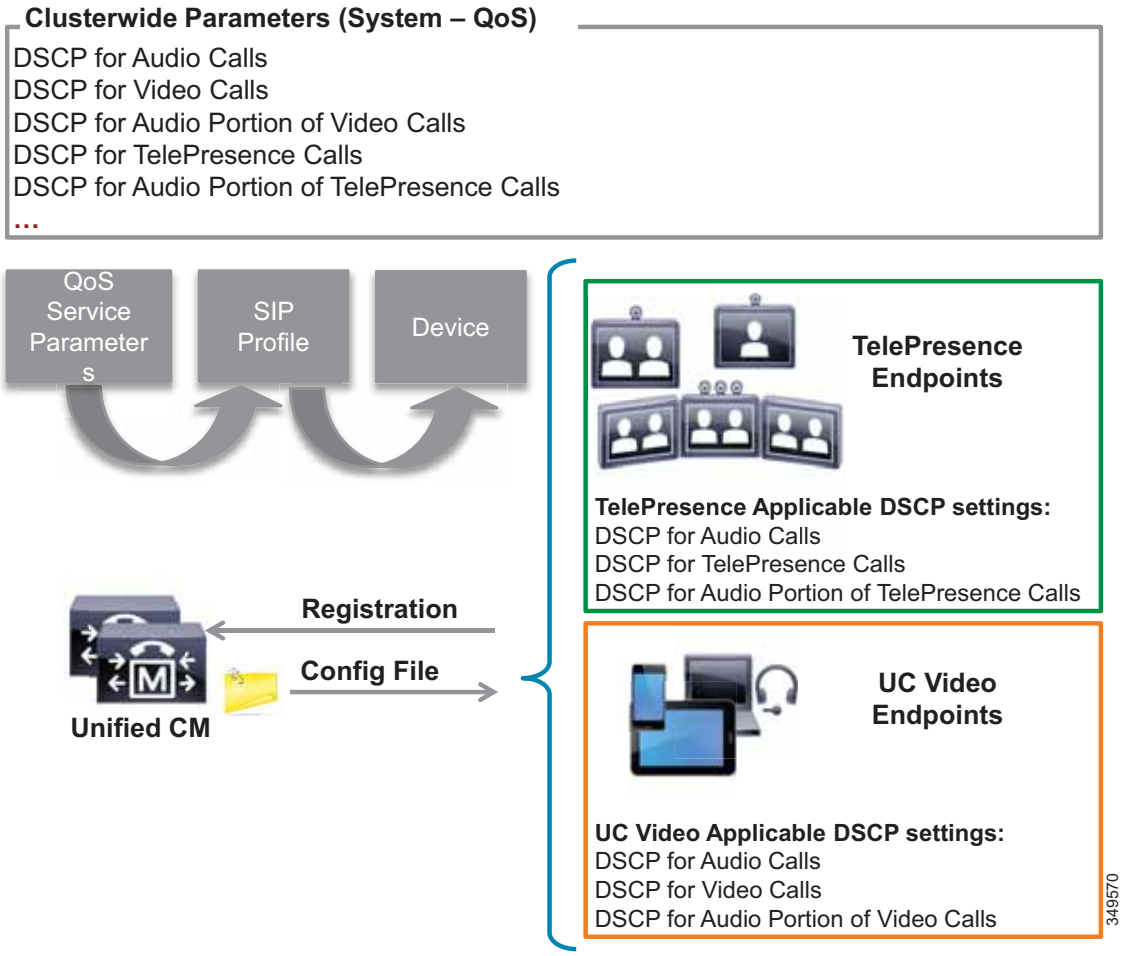
Table 8-3 shows the Preferred Architecture endpoints and their classification.

Table 8-3 PA Video Endpoints

Endpoint	TelePresence Endpoint	UC Video Endpoint
Cisco IP Phone 8800 Series		X
Cisco Jabber		X
Cisco DX70 and DX80 (CE Software)	X	
Cisco TelePresence MX Series	X	
Cisco TelePresence SX Series	X	
Cisco TelePresence IX Series	X	

Figure 8-7 illustrates how the two categories of Cisco video endpoints derive DSCP. These categories apply only to QoS and call admission control (CAC).

Figure 8-7 How Cisco Endpoints Derive DSCP



The configuration file is populated with the QoS parameters from the CallManager service parameters or the SIP Profile, when configured, and sent to the endpoint upon registration. The endpoint then uses the correct DSCP parameters for each type of media stream, depending on which category of endpoint it is. Table 8-4 lists the DSCP parameters, the type of endpoint, and the type of call flow determining the DSCP marking of the stream.

Table 8-4 DSCP for Basic Call Flows

DSCP Parameter	TelePresence Endpoint	UC Video Endpoint	Call Flow
DSCP for Audio Calls	X	X	Voice-only
DSCP for Video Calls		X	Video – Audio and video stream of a video call, unless the endpoint supports the **DSCP for Audio Portion of Video Calls** parameter (see Table 8-5)
DSCP for Audio Portion of Video Calls		X	Audio stream of a video call – Applicable only to endpoints that support the parameter
DSCP for TelePresence Calls	X		Immersive video – Audio and video stream of an immersive video call, unless the endpoint supports the **DSCP for Audio Portion of TelePresence Calls** parameter (see Table 8-5)
DSCP for Audio Portion of TelePresence Calls	X		Audio stream of a video call – Applicable only to endpoints that support the parameter

Table 8-5 Endpoint Support for DSCP for Audio Portion of Video and TelePresence Calls

Video Endpoint	DSCP for Audio Portion of Video Call	DSCP for Audio Portion of TelePresence Call
IP Phone 8845 and 8865 Series	Yes	No
DX70 and DX80 (CE Software)	No	Yes
TelePresence IX Series	No	Yes
TelePresence SX and MX Series (CE 8.0 or later firmware)	No	Yes

Trusted Core Devices and Applications

Like endpoints, devices and applications in the collaboration portfolio source and terminate media and signaling streams. These trusted applications require the appropriate configuration on the application itself as well as the switch to which the application is connected in order to transparently pass the QoS marking of the media and signaling.

Core trusted devices and applications:

- Cisco Unified Communications Manager and IM and Presence Service
- Cisco Expressway
- Cisco Unity Connection
- Cisco Meeting Server
- Cisco IOS SIP Gateway and Cisco Unified Border Element

It is important to ensure that DSCP Trust is enabled on the switch ports to which these endpoints and application servers are connected. QoS DSCP trust is typically enabled by default on all newer Cisco switches; however, it is important to verify each switch platform to determine if this QoS trust is enabled, since some platforms do not trust DSCP by default.

Endpoints and Clients

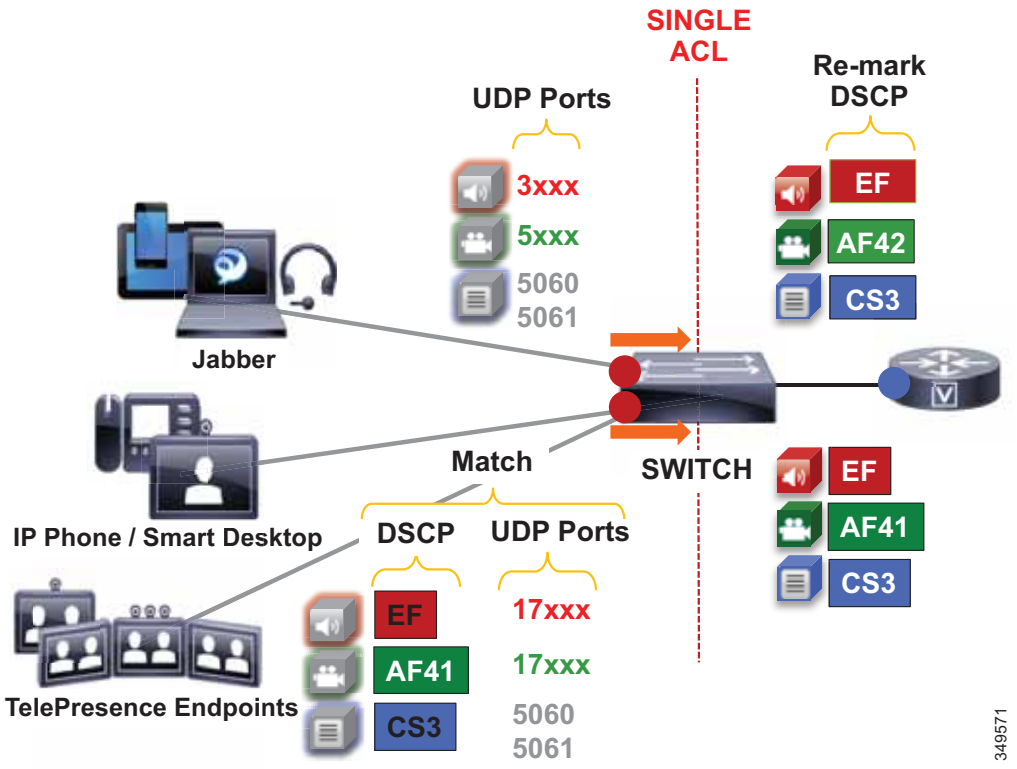
For the endpoints, the DSCP marking of packets on ingress into the switch needs to be re-marked using network access control lists (ACLs) to ensure that collaboration media and SIP signaling are marked appropriately and that other PC data traffic is marked to DSCP of Best Effort (DSCP 0) or as is appropriate based on the QoS data policies set forth by the enterprise.

The method used here consists of mapping identifiable media and signaling streams based on specific protocol ports, such as UDP and TCP ports, then making use of network access lists to remark QoS of the signaling and media streams based on those protocol port ranges. This method applies to all Cisco Jabber clients (Cisco Jabber for Windows, Cisco Jabber for Mac OS, Cisco Jabber for iPhone, Cisco Jabber for iPad, and Cisco Jabber for Android) because they all behave similarly when allocating media and signaling port ranges. Unlike Cisco Jabber clients, endpoints such as IP phones, 8800 Series IP and video phones, and DX Series with PC ports require an additional measure of matching on DSCP as well as UDP ports. The reason for this is that the IP phones and video endpoints use the same UDP port range for both audio and video, and thus in order to differentiate audio and video, matching on both UDP port range and DSCP allow for the proper identification of media traffic.

The concept is simple. An access list is used in the network access layer equipment (switch) to identify the media and signaling streams based on UDP port ranges and DSCP matching, and then it is set to re-mark them to the appropriate DSCP values. Although this technique is easy to implement and can be widely deployed, it is however not a 100% secure method and this point should be noted. This PA assumes that other security measures will be implemented to ensure the correct access to the network as well as any securing of the operating systems (OS) on the PCs and Macs used for Jabber to impede user tampering of OS related QoS settings.

Figure 8-8 illustrates the use of network access control lists (ACLs) to map identifiable media and signaling streams to DSCP for Jabber clients.

Figure 8-8 Endpoint Marking



Example 8-1 Switch ACL-Based QoS Policy for Untrusted Endpoints in Figure 8-8:

- Jabber clients
  - Match UDP Port Range 3xxx –> Re-mark to DSCP EF
  - Match UDP Port Range 5xxx –> Re-mark to DSCP AF42
  - Match TCP Port 5060 or 5061 –> Re-mark to DSCP CS3
- IP phones and video endpoints
  - Match UDP Port Range 17xxx with DSCP EF –> Re-mark to DSCP EF
  - Match UDP Port Range 17xxx with DSCP AF41 –> Re-mark to DSCP AF41
  - Match TCP Port 5060 or 5061 –> Re-mark to DSCP CS3
- Generic matching
  - Matches the rest of the traffic and sets DSCP to 0 (Best Effort or BE) using a default class-map

Endpoints send and receive other data and signaling such as ICMP, DHCP, TFTP, BFCP, LDAP, XMPP, FECC, CTI, and so forth. The QoS values for this traffic should follow the enterprise best practices for each type of traffic. Without doing this step, all other traffic apart from media and SIP signaling will be set to a DSCP of BE (DSCP 0) by the class-default in this configuration. We recommend either passing through the traffic marking by matching on DSCP and then re-marking the DSCP to the same value, or else using the TCP and UDP ports for each protocol that the endpoints use for communications.

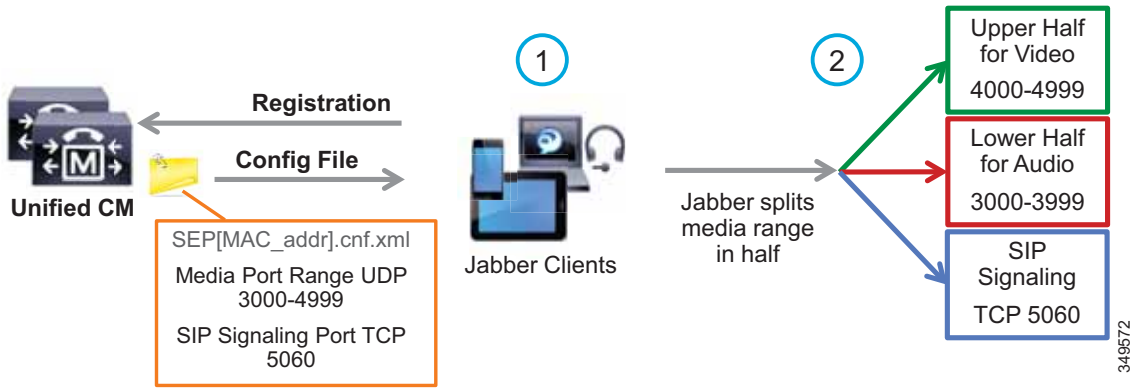
QoS for Cisco Jabber Clients

As discussed, this method involves classifying media and signaling by identifying the various streams from the Jabber client based on IP address, protocol, and/or protocol port range. Once identified, the signaling and media streams can be classified and re-marked with a corresponding DSCP. The protocol port ranges are configured in Unified CM and are passed to the endpoint to use during device registration. The network can then be configured via access control lists (ACLs) to classify traffic based on IP address, protocol, and protocol port range, and then to re-mark the classified traffic with the appropriate DSCP as discussed in the preceding section.

Cisco Jabber provides identifiable media streams based on UDP protocol port ranges and identifiable signaling streams based on TCP protocol port ranges. In Unified CM, the signaling port for endpoints is configured in the SIP Security Profile, while the media port range is configured in the SIP Profile of the Unified CM administration pages.

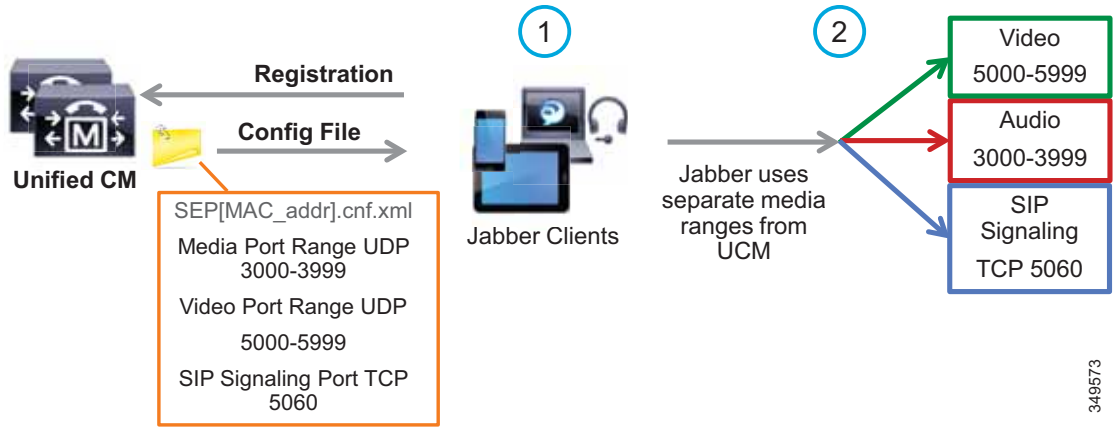
For the media port range, all endpoints and clients use the SIP profile parameter **Media Port Ranges** to derive the UDP ports used for media. By default, media port ranges are configured with **Common Port Range for Audio and Video**. When Jabber clients receive this port range in their configuration file, they split the port range in half and use the lower half for the audio streams of both voice and video calls and the upper half for the video streams of video calls. This is illustrated in Figure 8-9.

Figure 8-9 Media and Signaling Port Range – Common



Jabber can also use the **Media Port Ranges** > **Separate Port Range for Audio and Video** configuration. In this configuration the Unified CM administrator can specify a non-contiguous audio and video port range, as illustrated in Figure 8-10.

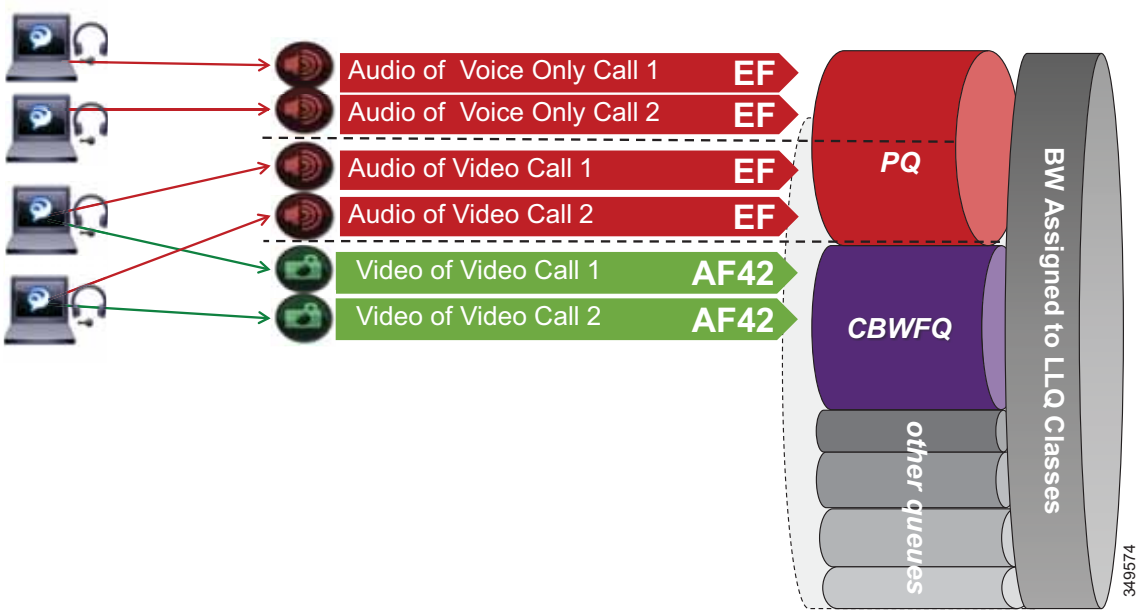
Figure 8-10 Media and Signaling Port Range – Separate



Caution **Security Alert**: If you use identifiable media streams for QoS classification at the network level, the trust model does *not* extend to the application itself. Apart from prioritizing streams from the intended application, other applications *could* potentially be configured to use the same identification criteria (media port range) and marking, and therefore achieve network prioritization. Because this unintended traffic would not be accounted for in call admission control or in the provisioning of the network, severe overall impact to real-time conversations could occur. It is for this reason that it is a good practice to define restricted port ranges whenever possible to identify the media streams.

When utilizing this technique, it is important to ensure that the audio portion of these video calls that will be re-marked to the audio traffic class (EF), and the video portions that will be re-marked to the video traffic class (AF4), are provisioned in the network accordingly. Figure 8-11 is an example of placing audio traffic into a Priority Queue (PQ) and video traffic into a Class Based Weighted Fair Queue (CBWFQ). The combination of PQ and CBWFQ is often referred to as low latency queuing (LLQ). Note that, because it is not possible to use port ranges in Cisco Jabber endpoints to differentiate the audio portion of voice-only calls from the audio portion of video calls, all audio using this technique will be re-marked to EF. It is important to provision the PQ adequately to support voice-only calls and the audio portion of video calls. An example of such provisioning is illustrated in Figure 8-11. For more information on the design and deployment recommendations for provisioning queuing and scheduling in the network, see the WAN Queuing and Scheduling section.

Figure 8-11 Provisioning Jabber QoS in the Network



According to RFC 3551, when RTCP is enabled on the endpoints, it uses the next higher odd port. For example, a device that establishes an RTP stream on port 3500 would send RTCP for that same stream on port 3501. This function of RTCP is also true with all Jabber clients. RTCP is common in most call flows and is typically used for statistical information about the streams and to synchronize audio and video in video calls to ensure proper lip-sync. In most cases, video and RTCP can be enabled or disabled on the endpoint itself or in the common phone profile settings.

Utilizing the Network for Classification and Marking

Based on the identifiable media and signaling streams created by the endpoints, common network QoS tools can be used to create traffic classes and to re-mark packets according to those classes.

These QoS mechanisms can be applied at different layers, such as the access layer (access switch), which is closest to the endpoint and the router level in the distribution, core, or services WAN edge. Regardless of where classification and re-marking occurs, we recommend using DSCP to ensure end-to-end per-hop behaviors.

As previously mentioned, Cisco Unified CM allows the port range utilized by SIP endpoints to be configured in the SIP Profile. As a general rule, a port range of a minimum of 100 ports (for example, 3000 to 3099) is sufficient for most scenarios. A smaller range could be configured as long as there are enough ports for the various audio, video, and associated RTCP ports (RTCP runs over the odd ports in the range), as well as ensuring against port conflict with other applications on the operating system of the device that may be using these ports since this could cause port collisions.

Access Layer (Layer 2 Definitions)

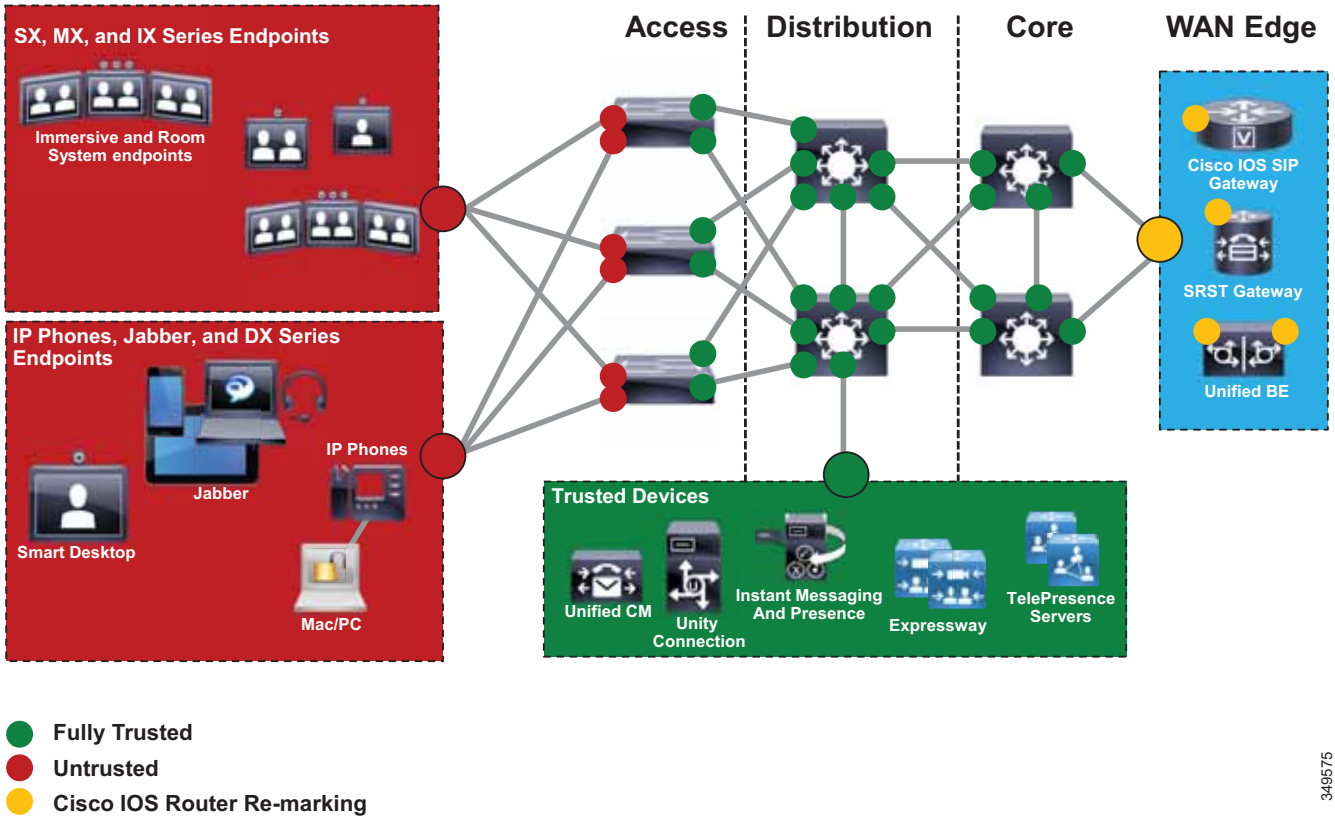
When utilizing the access layer to classify traffic, the classification occurs at the ingress of traffic into the network, thus allowing the flows to be identified as they enter. In environments where QoS policies are applied not only in the WAN but also within the LAN, all upstream components can rely on traffic markings when processing. Classification at the ingress allows different methods to be utilized based on different types of endpoints.

Configuring QoS policies in the access layer of the network could result in a significant amount of devices that require configuration, which can create additional operational overhead. The QoS policy configurations should be standardized across the various switches of the access layer through templates. You can use configuration deployment tools to relieve the burden of manual configuration. The PA simplifies this process by using a single group of ACLs that can be used across the various switching platforms.

Distribution/Core/Services WAN Edge (Layer 3 Definitions)

Another location where QoS marking can take place is at the Layer 3 routed boundary. In a campus network Layer 3 could be in the access, distribution, core, or services WAN edge layers. The recommendation is to classify and re-mark at the access layer, then trust through the distribution and core of the network, and finally re-classify and re-mark at the WAN edge if and when needed. For smaller networks such as branch offices where there are no Layer 3 switching components deployed, QoS marking can be applied at the WAN edge router. At Layer 3, QoS policies are applied to the Layer 3 routing interfaces. In most campus networks these would be VLAN interfaces, but they could also be Fast Ethernet or Gigabit Ethernet interfaces. Figure 8-12 illustrates the areas of the network where the various types of trust are applied in relation to the places in the network – access, distribution, core, or WAN edge.

Figure 8-12 Trust and Enforcement – Places in the Network



Endpoint Identification and Classification Considerations and Recommendations

Summary of design and deployment considerations and recommendations:

- Use DSCP markings whenever possible because these are IP layer end-to-end, more granular, and more extensible than Layer 2 markings.
- Mark as close to the endpoint as possible, preferably at the LAN switch level.
- When trying to minimize the number of media ports used by the Cisco Jabber client, a minimum range of 100 ports is recommended. This is to ensure that there are enough ports for all of the streams, such as RTCP, RTP for audio and video, BFCP, and RTP for secondary video for desktop sharing sessions, as well as to avoid any overlap with other applications on the same computer.
- Ensure a QoS policy includes other pertinent collaboration traffic to be re-marked, otherwise a value of 0 (Best Effort, or BE) will be placed on all remaining traffic.

WAN Queuing and Scheduling

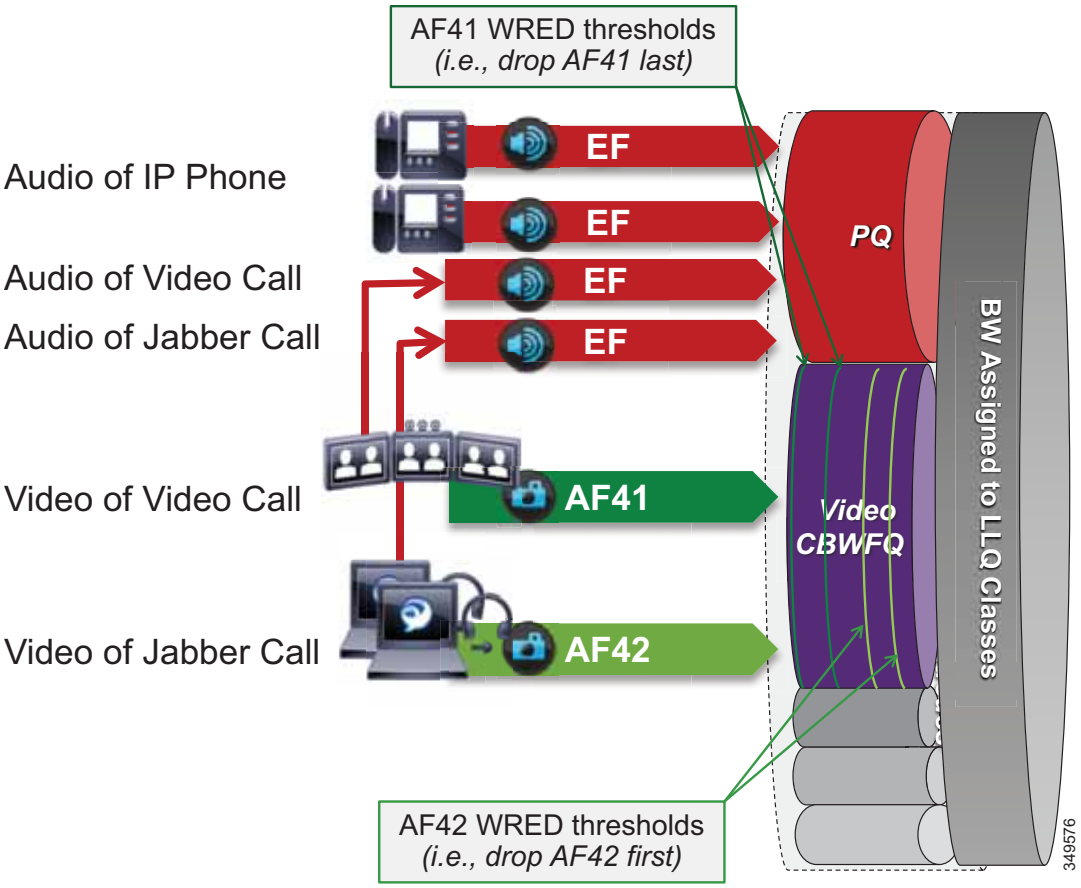
As discussed in the Identification and Classification section, Unified CM has the ability to differentiate the video endpoint types as well as the their media streams. This provides the network administrator the ability to treat the video from different endpoints differently in the network. The recommended approach in the PA is to use AF42 DSCP markings for Jabber clients and AF41 for all smart desktop, room system, and immersive video endpoints. These values are in line with RFC 4594.

PA Queuing and Scheduling Approach (Single Video Queue)

A single rate-based queue with multiple DSCPs with differing drop probabilities is used in the PA for managing multiple types of video across an integrated collaboration media and data network. In this approach to scheduling video traffic in the WAN, the single video queue is configured with two AF4 drop probabilities using AF41 and AF42, where AF42 has a higher drop precedence or probability than AF41. The premise behind a single video queue with this service class with hierarchical drop precedence is that, when one class of video is not using the bandwidth within the queue, the rest of the queue bandwidth is available for the other class of video. This approach dedicates bandwidth reserved for video to be available only to the video queue during times of congestion. Other approaches such as dual rate-based video queues sub-optimally allocate excess video bandwidth from one queue to all queues on the interface equally.

Although different strategies for optimized video bandwidth utilization can be designed based on this single video queue with hierarchical DSCP drop probabilities, the PA approach is illustrated in Figure 8-13.

Figure 8-13 Single Video Queue Approach



In Figure 8-13 the audio of a voice call is marked as EF and placed into a Priority Queue (PQ) with a strict policer on how much bandwidth the PQ can allocate to this traffic. Video calls are separated into two classes, AF41 for prioritized video and AF42 for opportunistic or Jabber video. Using a CBWFQ with Weighted Random Early Detection (WRED), the administrator can adjust the drop precedence of AF42 over AF41, thus ensuring that during times of congestion when the queue is filling up, AF42 packets are dropped from the queue at a higher probability than AF41. See the *WAN Quality of Service* section in the Network Infrastructure chapter of the Collaboration SRND for more details on the function of WRED.

The above example illustrates how an administrator using a single CBWFQ with DSCP-based WRED for all video can protect one type of video (immersive video) from packet loss during periods of congestion over another type of video (desktop). With this "single video queue approach," unlike the "dual video queue approach," when one type of video is not using bandwidth in the queue, the other type of video gains full access to the entire queue bandwidth if and when needed. This is a significant improvement when deploying pervasive video.

Achieving this holistically across the entire solution depends on a number of conditions. Below is a list of conditions required to achieve marking all audio to a DSCP of EF:

- The customer equipment (CE) or service provider (SP) owned WAN equipment must support AF4 QoS containing both AF41 and AF42 QoS markings as well as Weighted Random Early Detection (WRED).
- Enhanced Locations Call Admission Control (ELCAC) can be implemented in conjunction with marking all audio as EF. ELCAC relies on the correct DSCP setting in order to ensure the protection of the queues that voice and video CAC pools represent. Changing the DSCP of audio streams of the video calls requires updating how ELCAC deducts bandwidth for video calls. This can be done by setting the service parameter **Deduct Audio Bandwidth from Audio Pool for Video Call**, under the Call Admission Control section of the CallManager service, to **True**. This parameter can be set to true or false:
  - **True** (recommended) — Cisco Unified CM splits the audio and video bandwidth allocations for video calls into separate pools. The bandwidth allocation for the audio portion of a video call is deducted from the audio pool, while the video portion of a video call is deducted from the video pool.
  - **False** (default)— Cisco Unified CM applies the legacy behavior, which is to deduct the audio and video bandwidth allocations of video call from the video pool. This is the default setting.

Opportunistic Video

When video is deployed pervasively across the organization, bandwidth constraints typically determine the video resolution that can be achieved during the busiest hour of the day based on the bandwidth available and the number of video calls during that busy hour. To address this challenge, the PA has targeted a group of endpoints whose video is treated opportunistically by the network by using a single video queue with DSCP-based WRED coupled with a strategy for identification and classification of the Jabber clients' collaboration media.

Opportunistic video is achieving the best video quality based on the WAN bandwidth resources available at any given time. To achieve this, a number of requirements must be met:

- Select a group of video endpoints to be opportunistic. In the case of the PA, Jabber clients are used as the opportunistic video endpoints.
- Ensure the WAN is configured with a single video queue using DSCP-based WRED with AF4 DSCP class servicing with drop precedence of AF41 and AF42. (While AF43 could be used, only two DSCP values are necessary in the PA.)
- Identify and classify the video of opportunistic endpoints with AF42.
- Identify and classify all other video endpoints with AF41.

Provisioning and Admission Control

Provisioning bandwidth and ensuring the correct bit rate is negotiated between various groups of endpoints are important aspects of bandwidth management. In a Unified CM environment, bit rate is negotiated through Unified CM, which uses a concept of regions to set maximum audio and maximum video bit rates for any given call flow. This section focuses on the maximum bit rate for video calls.

Unified CM locations (see Enhanced Locations Call Admission Control) work in conjunction with regions to define the characteristics of a call flow. Regions define the type of compression or bit rate (8 kbps or G.729, 64 kbps or G.722/G.711, and so forth) that is used between any two devices. Location

links define the amount of available bandwidth for the path between devices. Each device and trunk in the system is assigned to both a region (by means of a device pool) and a location (by means of a device pool or by direct configuration on the device itself):

- Regions allow the bandwidth of video calls to be set. The audio limit on the region can result in filtering out codecs with higher bit rates. However, for video calls, the video limit constrains the quality (resolution and transmission rate) of the video.
- Locations define the amount of total bandwidth available for all calls on that link. When a call is made on a link, the regional value for that call must be subtracted from the total bandwidth allowed for that link.

Building a region matrix to manage maximum video bit rate (video resolution) for groups of devices can assist in ensuring that certain groups of devices do not over-saturate the network bandwidth. The following guidelines apply to creating a region matrix:

- Group devices into maximum video bit rate categories.
- The smaller the number of groups, the easier it is to calculate bandwidth requirements.
- Consider the default region settings to simplify the matrix and provide intra-region and inter-region defaults.
- Use a single audio codec across the entire organization, both LAN and WAN, to simplify the region matrices.

For more information about region settings, see the section on Enhanced Locations Call Admission Control.

Table 8-6 lists an example of a maximum video session bit rate region matrix for three groups of devices.

Note Table 8-6 is only an example of how to group devices and what maximum bit rate might be suggested for a general resolution between the groups of devices.

Table 8-6 Example Group Region Matrix for Three Groups of Devices

Endpoint Groupings	Video_1.5MB	Video_2.5MB	Video_20MB
Video_1.5MB	1,500 kbps	1,500 kbps	1,500 kbps
Video_2.5MB	1,500 kbps	2,500 kbps	2,500 kbps
Video_20MB	1,500 kbps	2,500 kbps	20,000 kbps

For the example in Table 8-6, the three groups are:

- Video_1.5MB

  These devices would typically be the largest group of deployed video capable endpoints and thus would benefit from the opportunistic video approach. Classified as opportunistic video, they can go up to a maximum of 1,500 kbps (720p @ 30 fps) and will rate-adapt downward based on packet loss.

- Video_2.5MB

  These devices would be room systems such as the Cisco TelePresence MX or SX Series as well as smart desktop endpoints such as the Cisco DX Series. At 2,500 kbps maximum video bit rate, these endpoints would typically be capable of 720p @ 30 fps.

- Video_20MB

  This class is for the larger Cisco TelePresence IX Series endpoints as well as Cisco Meeting Servers and MCUs set to a maximum of 20 Mbps to allow for endpoints capable of it to run at 1080p @ 60 fps. (The IX Series, for example, requires 18 MB to do 1080p @ 60 fps on three screens). Single-screen systems use much less bandwidth, but this group would be for devices utilizing their maximum bit rate capacity.

To simplify the configuration of the regions, it is important to standardize on one audio codec to be used throughout the entire organization. The first consideration is to decide whether to have a lower bit rate codec for audio calls between sites. Historically as part of managing bandwidth, enterprises have used a lower bit rate codec such as G.729 over the WAN while using a higher bit rate, wider band codec such as G.722 for calls within the LAN or MAN. Typically when deploying video at 1 to 2.5 MB per call, audio (even at 80 kbps per call) consumes so much less bandwidth that many enterprises have moved to using a higher bit rate, better quality codec such as G.722 across the entire organization (LAN and WAN). This decision has an impact on the region matrix and whether per-site regions are required or not. The concept here is that if inter-region audio or video bit rates are to be different, then per-site regions will be required. This augments the configuration of regions to the number of sites (N) multiplied by the number of video groups (X):

Number of regions required on average = N∗X

If audio bit rates will be the same across the WAN and LAN, then only the regions for the video groups are required (X).

Enhanced Locations Call Admission Control

The call admission control function is an important component of a Collaboration system, especially when multiple sites are connected through an IP WAN and limited bandwidth resources are available for audio and video calls.

Call Admission Control Architecture

Unified CM Enhanced Locations Call Admission Control

Cisco Unified CM provides Enhanced Locations Call Admission Control (ELCAC) to support complex WAN topologies as well as distributed deployments of Unified CM for call admission control where multiple clusters manage devices in the same physical sites using the same WAN up-links.

To support more complex WAN topologies, Unified CM implements a location-based network modeling functionality. This provides Unified CM with the ability to support multi-hop WAN connections between calling and called parties. This network modeling functionality has also been incrementally enhanced to support multi-cluster distributed Unified CM deployments. This allows each cluster to "share" locations by enabling the clusters to communicate with one another to reserve, release, and adjust allocated bandwidth for the same locations across clusters.

Network Modeling with Locations, Links, and Weights

Enhanced Locations CAC is a model-based static CAC mechanism. ELCAC involves using the administration interface in Unified CM to configure locations and links to model the "routed WAN network" in an attempt to represent how the WAN network topology routes media between groups of endpoints for end-to-end audio and video calls. Although Unified CM provides configuration and serviceability interfaces in order to model the network, it is still a "static" CAC mechanism that does not take into account network failures and network protocol rerouting. Therefore, the model needs to be

updated when the WAN network topology changes or bandwidth allocations across the WAN are increased or decreased. Enhanced Locations CAC is also call oriented, and bandwidth deductions are per-call not per-stream, so asymmetric media flows where the bit-rate is higher in one direction than in the other will always deduct for the highest bit rate bi-directionally. In addition, unidirectional media flows will be deducted as if they were bidirectional media flows.

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

- Locations — A location represents a LAN. It could contain endpoints or simply serve as a transit location between links for WAN network modeling. For example, an MPLS provider could be represented by a location.
- Links — Links interconnect locations and are used to define bandwidth available between locations. Links logically represent the WAN link and are configured in the Location user interface (UI).
- Weights — A weight provides the relative priority of a link in forming the effective path between any pair of locations. The effective path is the path used by Unified CM for the bandwidth calculations, and it has the least cumulative weight of all possible paths. Weights are used on links to provide a "cost" for the "effective path" and are pertinent only when there is more than one path between any two locations.
- Path — A path is a sequence of links and intermediate locations connecting a pair of locations. Unified CM calculates least-cost paths (lowest cumulative weight) from each location to all other locations and builds a map of the various paths. Only one "effective path" is used between any pair of locations.
- Effective Path — The effective path is the path with the least cumulative weight and is the bandwidth accounting path that is always used between any two locations.
- Bandwidth Allocation — Is the amount of bandwidth allocated in the model for each type of traffic: audio, video, and immersive video.
- Location Bandwidth Manager (LBM) — Is the active service in Unified CM that assembles a network model from configured location and link data in one or more clusters. It determines the effective paths between pairs of locations, determines whether to admit calls between a pair of locations based on the availability of bandwidth for each type of call, and deducts (reserves) bandwidth for the duration of each call that is admitted.
- Location Bandwidth Manager Hub — Is a Location Bandwidth Manager (LBM) service that has been designated to participate directly in intercluster replication of fixed locations, links data, and dynamic bandwidth allocation data. LBMs assigned to an LBM hub group discover each other through their common connections and form a fully-meshed intercluster replication network. Other LBM services in a cluster with an LBM hub participate indirectly in intercluster replication through the LBM hubs in their cluster.

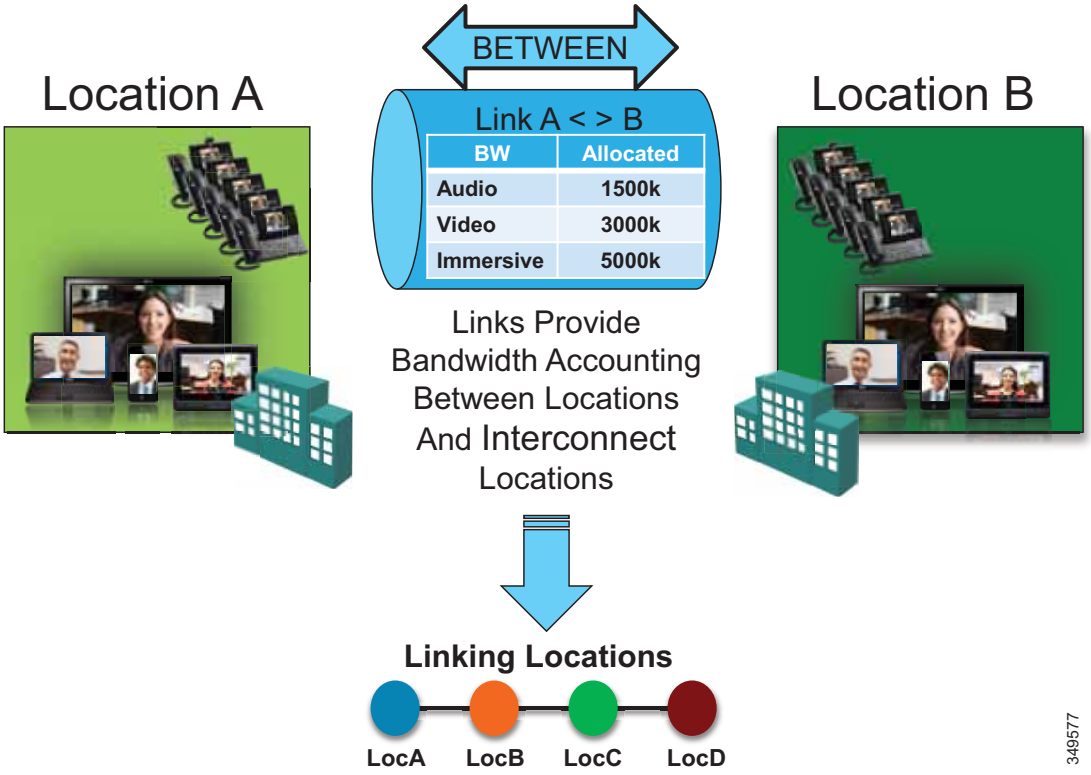
Locations, Links, and Weight

Unified CM uses the concept of locations to represent a physical site and to create an association with media devices such as endpoints, voice messaging ports, trunks, gateways, and so forth, through direct configuration on the device itself, through a device pool, or even through device mobility. Locations logically represent the Local Area Network (LAN). Unified CM also uses a configuration parameter called links to interconnect locations and to define bandwidth available between locations. Links logically represent the Wide Area Network (WAN). This section describes locations and links and how they are used (see Figure 8-14).

The location configuration itself consists of three main parts: links, intra-location bandwidth parameters, and RSVP locations settings. The intra-location bandwidth parameters are set to unlimited by default and should remain that way because there is little or no reason to limit bandwidth within a location (LAN). The RSVP locations settings are not considered here for Enhanced Location CAC because they apply only to RSVP implementations.

The link bandwidth parameters allow the administrator to characterize the provisioned bandwidth for audio, video, and immersive calls between "adjacent locations" (that is, locations that have a link configured between them). This feature offers the administrator the ability to create a string of location pairings in order to model a multi-hop WAN network.

Figure 8-14 Locations and Links



Weight is configurable on the link and provides the ability to force a specific path choice when multiple paths between two locations are available. When multiple paths are configured, only one will be selected based on the cumulative weight, and this path is referred to as the *effective path*. This weight is static and the effective path does not change dynamically. When two paths have equal weight, one path is randomly chosen; therefore it is important to ensure that only one path exists or has the least cumulative weight to ensure that it is the effective path for the Location Bandwidth Manager (LBM). This is especially important in multi-cluster environments.

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

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

Unified CM allows the administrator to define separate voice, video, and immersive video bandwidth pools for each link between locations. In the PA, only voice and video bandwidth pools are used. Typically the link between locations is set to a finite number of kilobits per second (kbps) to match the provisioned amount of bandwidth available for audio and video in the WAN links between physical sites. Some WANs do not require any limitations because they are over-provisioned for the expected amount of traffic. If the bandwidth values are set to a finite number of kilobits per second (kbps), Unified CM will track all calls within the location and all calls that use the location as a transit location (a location that is in the calculation path but is not the originating or terminating location in the path).

The following devices must be configured in a location:

- Endpoints
- Conference bridges
- Gateways
- SIP trunks
- Music on hold (MoH) servers
- Annunciator (via device pool)

Table 8-7 lists the amount of bandwidth requested for various call speeds. For all audio (both audio-only calls as well as audio of a video call), Unified CM counts the media bit rates plus the IP and UDP overhead. For example, a G.711 or G.722 audio call consumes 80 kbps (64 kbps bit rate + 16 kbps for IP/UDP headers) deducted from the audio bandwidth allocation of the location and link. For a video call, Unified CM counts only the payload (no IP/UDP header overhead) for video streams, but the audio portion is calculated with IP and UDP overhead. For example, for a video call at a bit rate of 384 kbps where audio is set to use a 64 kbps bit rate, Unified CM will allocate 320 kbps from the video bandwidth allocation and take 64 kbps for the audio and add the 16 kbps for IP/UDP headers to derive 80 kbps for the audio pool deduction. For the same call where the audio is set to use a 8 kbps bit rate, Unified CM will allocate 376 kbps from the video bandwidth allocation and take 8 kbps for the audio and add the 16 kbps for IP/UDP headers to derive 24 kbps for the audio pool deduction.

Table 8-7 Amount of Bandwidth Requested by the Locations and Links Bandwidth Deduction Algorithm in the PA Configuration[1]

Call Speed (Session Bit Rate)	Audio Pool Bandwidth	Video Pool Bandwidth
G.711 or G.722 audio call (64 kbps)	80 kbps	N/A
G.729 audio call (8 kbps)	24 kbps	N/A
512 kbps video call with G.729 audio (8 kbps)	24 kbps	504 kbps
512 kbps video call with G.711 or G.722 audio (64 kbps)	80 kbps	448 kbps
768 kbps video call with G.729 audio (8 kbps)	24 kbps	760 kbps
768 kbps video call with G.711 or G.722 audio (64 kbps)	80 kbps	704 kbps
1,024 kbps video call with G.729 audio (8 kbps)	24 kbps	1,016 kbps
1,024 kbps video call with G.711 or G.722 audio (64 kbps)	80 kbps	960 kbps

1. Only 8 kbps and 64 kbps are used in these examples, but the same principle also applies to other audio bit rate codecs. The audio bit rate (payload only) is subtracted the video bit rate to get the adjusted video bit rate value.

For example, assume that the link configuration for the location Branch 1 to Hub_None allocates 256 kbps of available audio bandwidth and 1,024 kbps of available video bandwidth. In this case the path from Branch 1 to Hub_None can support up to three G.711 audio calls (at 80 kbps per call) or ten G.729 audio calls (at 24 kbps per call), or any combination of both that does not exceed 256 kbps. The link between locations can also support different numbers of video calls, depending on the video and audio codecs being used (for example, one video call requesting 1,024 kbps of bandwidth or two video calls with each requesting 512 kbps of bandwidth).

When a call is placed from one location to the other, Unified CM deducts the appropriate amount of bandwidth from the effective path of locations and links from one location to another. When the call has completed, Unified CM returns the bandwidth to those same links over the effective path. If there is not enough bandwidth at any one of the links over the path, the call is denied by Unified CM and the caller receives the network busy tone. If the calling device is an IP phone with a display, that device also displays the message "Not Enough Bandwidth."

When an inter-location call is denied by call admission control, Unified CM can automatically reroute the call to the destination through the PSTN connection by means of the Automated Alternate Routing (AAR) feature (see Automated Alternate Routing, page 2-47).

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

Also, AAR is applicable only to intra-cluster endpoint-to-endpoint calls. For all inter-cluster calls that fail CAC, the route groups are used to try different SIP routing paths.

Video devices can be enabled to **Retry Video Call as Audio** if a video call between devices fails CAC. This option is configured on the video endpoint or SIP trunk configuration page in Unified CM and is applicable to video endpoints or trunks placing calls. For some video endpoints, **Retry Video Call as Audio** is enabled by default and not configurable on the endpoint.

Locations, Links, and Region Settings

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

You can configure locations in Unified CM to define:

- Physical sites (for example, a branch office) or transit sites (for example, an MPLS cloud) — A location represents a LAN. It could contain endpoints or simply serve as a transit location between links for WAN network modeling.
- Link bandwidth between adjacent locations — Links interconnect locations and are used to define bandwidth available between locations. Links logically represent the WAN link between physical sites.
  - Audio Bandwidth — The amount of bandwidth that is available in the WAN link for voice and fax calls being made from devices in the location to the configured adjacent location. Unified CM uses this bandwidth value for Enhanced Locations Call Admission Control.
  - Video Bandwidth — The amount of video bandwidth that is available in the WAN link for video calls being made from devices in the location to the configured adjacent location. Unified CM uses this bandwidth value for Enhanced Locations Call Admission Control.
  - Immersive Video Bandwidth — Not used in the PA configuration.

You can configure regions in Unified CM to define:

- The maximum audio bit rate
- The maximum session bit rate for video calls (includes audio)
- The maximum session bit rate for immersive video calls (includes audio) — Not used in the PA configuration.
- Audio codec preference lists

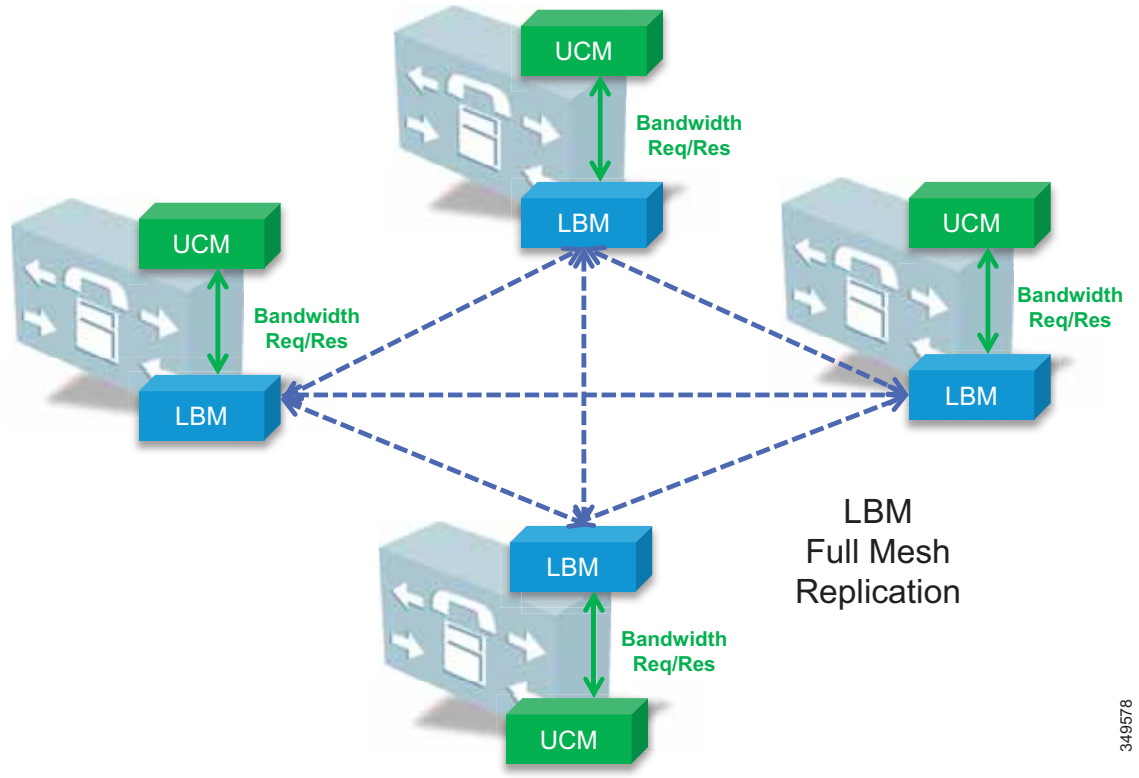
Location Bandwidth Manager

The Location Bandwidth Manager (LBM) is a Unified CM Feature Service managed from the serviceability web pages and is responsible for all of the Enhanced Locations CAC bandwidth functions. The LBM should be configured to run on each subscriber node in the cluster that is also running the Cisco CallManager service.

The LBM performs the following functions:

- Assembles topology of locations and links
- Calculates the effective paths across the topology
- Services bandwidth requests from the Cisco CallManager service (Unified CM call control)
- Replicates the bandwidth information to other LBMs (see Figure 8-15)
- Provides configured and dynamic information to serviceability
- Updates Location Real-Time Monitoring Tool (RTMT) counters

Figure 8-15 LBM Local Replication Network



By default the CallManager service communicates with the local LBM service.

Deducting All Audio from the Voice Pool

This PA utilizes a new Unified CM 11.*x* feature that allows the administrator to deduct the audio bandwidth of video calls from the voice pool. Because ELCAC relies on the correct DSCP setting in order to ensure the protection of the queues that voice and video CAC pools represent, changing how Unified CM deducts bandwidth from the video pool requires the DSCP of audio streams of the video calls to be marked the same as the audio streams of audio-only calls.

In Unified CM this feature is enabled by setting the service parameter **Deduct Audio Bandwidth from Audio Pool for Video Call** to **True** under the Call Admission Control section of the CallManager service. False is the default setting, and by default Unified CM deducts both audio and video streams of video calls from the video pool.

Multi-Cluster Considerations

This section covers the following topics:

- Intercluster ELCAC
- LBM Hub Replication Network
- Common Locations (Shared Locations) and Links
- Shadow Location
- Location and Link Management Cluster

Intercluster ELCAC

Intercluster Enhanced Locations CAC extends the concept of network modeling across multiple clusters. In intercluster Enhanced Locations CAC, each cluster manages its locally configured topology of locations and links and then propagates this local topology to other remote clusters that are part of the LBM intercluster replication network. Upon receiving a remote cluster's topology, the LBM assembles this into its own local topology and creates a global topology. Through this process the global topology is then identical across all clusters, providing each cluster a global view of enterprise network topology for end-to-end CAC. Figure 8-16 illustrates the concept of a global topology with a simplistic hub-and-spoke network topology as an example.

Figure 8-16 Example of a Global Topology for a Simple Hub-and-Spoke Network

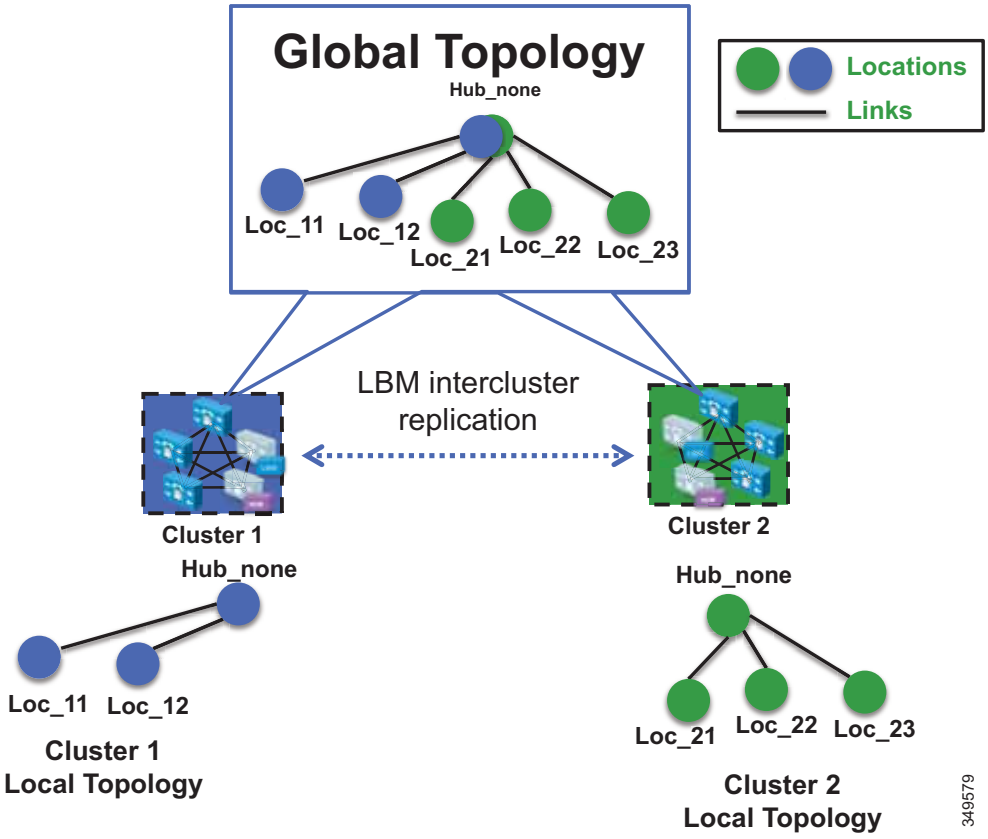


Figure 8-16 shows two clusters, Cluster 1 and Cluster 2, each with a locally configured hub-and-spoke network topology. Cluster 1 has configured Hub_None with links to Loc_11 and Loc_12, while Cluster 2 has configured Hub_None with links to Loc_21, Loc_22, and Loc_23. When intercluster Enhanced Locations CAC is enabled, Cluster 1 sends its local topology to Cluster 2, as does Cluster 2 to Cluster 1. After each cluster obtains a copy of the remote cluster's topology, each cluster overlays the remote cluster's topology over its own. The overlay is accomplished through common locations, which are locations that are configured with the same name. Because both Cluster 1 and Cluster 2 have the common location Hub_None with the same name, each cluster will overlay the other's network topology with Hub_None as a common location, thus creating a global topology where Hub_None is the hub and Loc_11, Loc_12, Loc_21, Loc_22 and Loc_23 are all spoke locations. This is an example of a simple network topology, but more complex topologies would be processed in the same way.

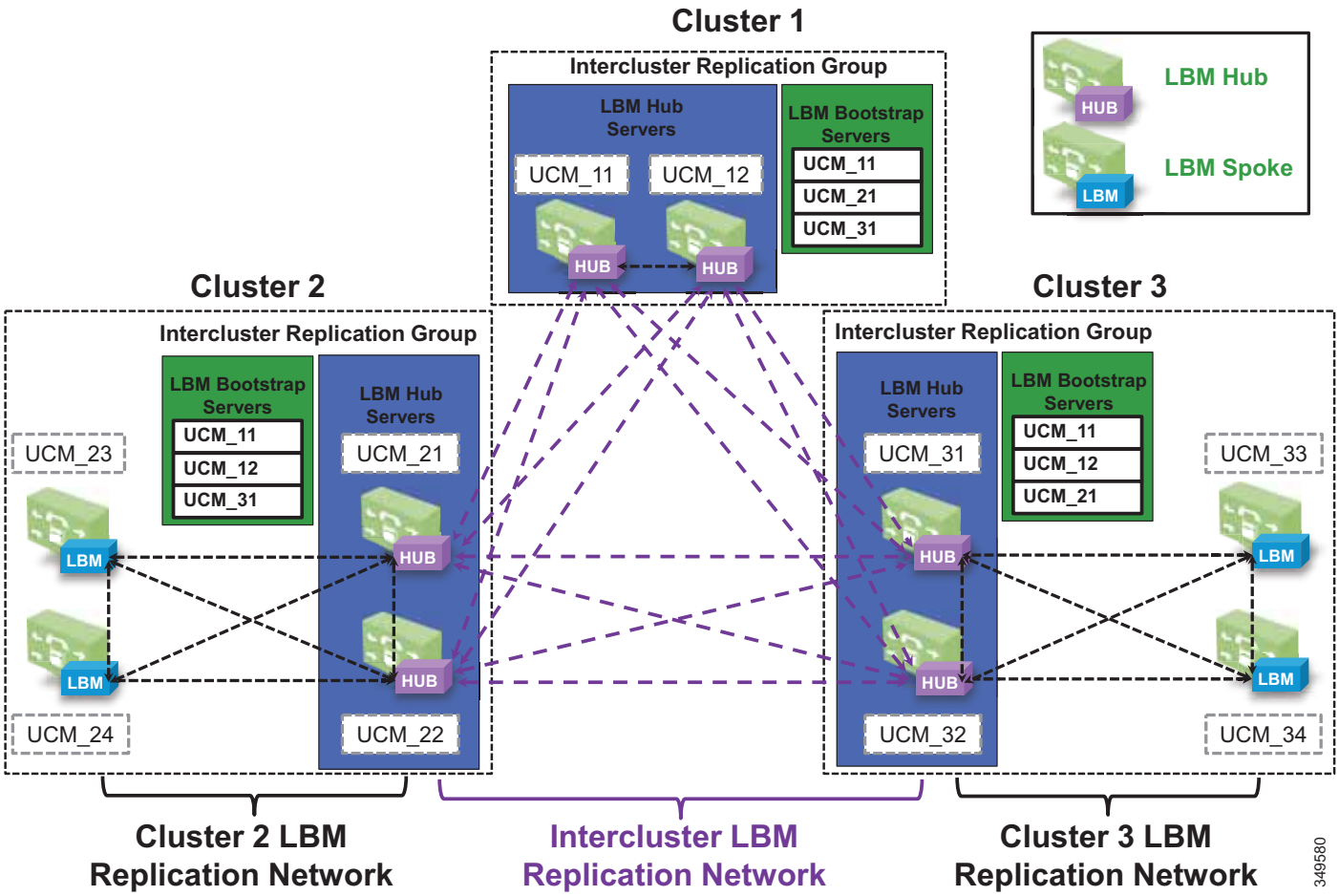
LBM Hub Replication Network

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

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

Once the LBM hub group is configured on each cluster, the designated LBM hubs will create the full-mesh intercluster replication network. Figure 8-17 illustrates an intercluster replication network configuration with LBM hub groups set up between three clusters (Cluster 1, Cluster 2, and Cluster 3) to form the intercluster replication network.

Figure 8-17 Example Intercluster Replication Network for Three Clusters



In Figure 8-17, two LBMs from each cluster have been designated as the LBM hubs for their cluster. This provides redundancy for the LBM hub role. These LBM hubs form the intercluster LBM replication network. The bootstrap LBMs configured in each LBM intercluster replication group are designated as UCM_11 and UCM_12. These two LBM hubs from Cluster 1 serve as points of contact or bootstrap LBMs for the entire intercluster LBM replication network. UCM_21 and UCM_31 in Cluster 2 and Cluster 3, respectively, serve as backup bootstrap LBM hubs when the primaries are not available (that is, when Cluster 1 is not available). Establishing the intercluster LBM replication network means that each LBM hub in each cluster connects to UCM_11, replicates its local topology, and gets the remote topology. They also get the connectivity information for the other clusters from UCM_11, connect to the other remote clusters, and replicate their topologies. This creates the full-mesh replication network. If UCM_11 is unavailable, the LBM hubs will connect to UCM_12. If Cluster 2 LBM hubs are unavailable, Cluster 2 and Cluster 3 LBM hubs will connect to UCM_31, and Cluster 3 LBM hubs will connect to UCM_21.

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

- LBM hubs (local LBMs)
  - Communicate directly to other remote hubs as part of the intercluster LBM replication network
- LBM spokes (local LBMs)
  - Communicate directly to local LBM hubs in the cluster and indirectly to the remote LBM hubs through the local LBM hubs
- Bootstrap LBMs
  - LBM hubs responsible for interconnecting all clusters' LBM hubs in the replication network
  - Can be any LBM hub(s) in the network
  - Can indicate up to three bootstrap LBM hubs per LBM intercluster replication group
- LBM hub replication network — Bandwidth deduction and adjustment messages
  - LBM optimizes the LBM messages by choosing a sender and receiver from each cluster

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

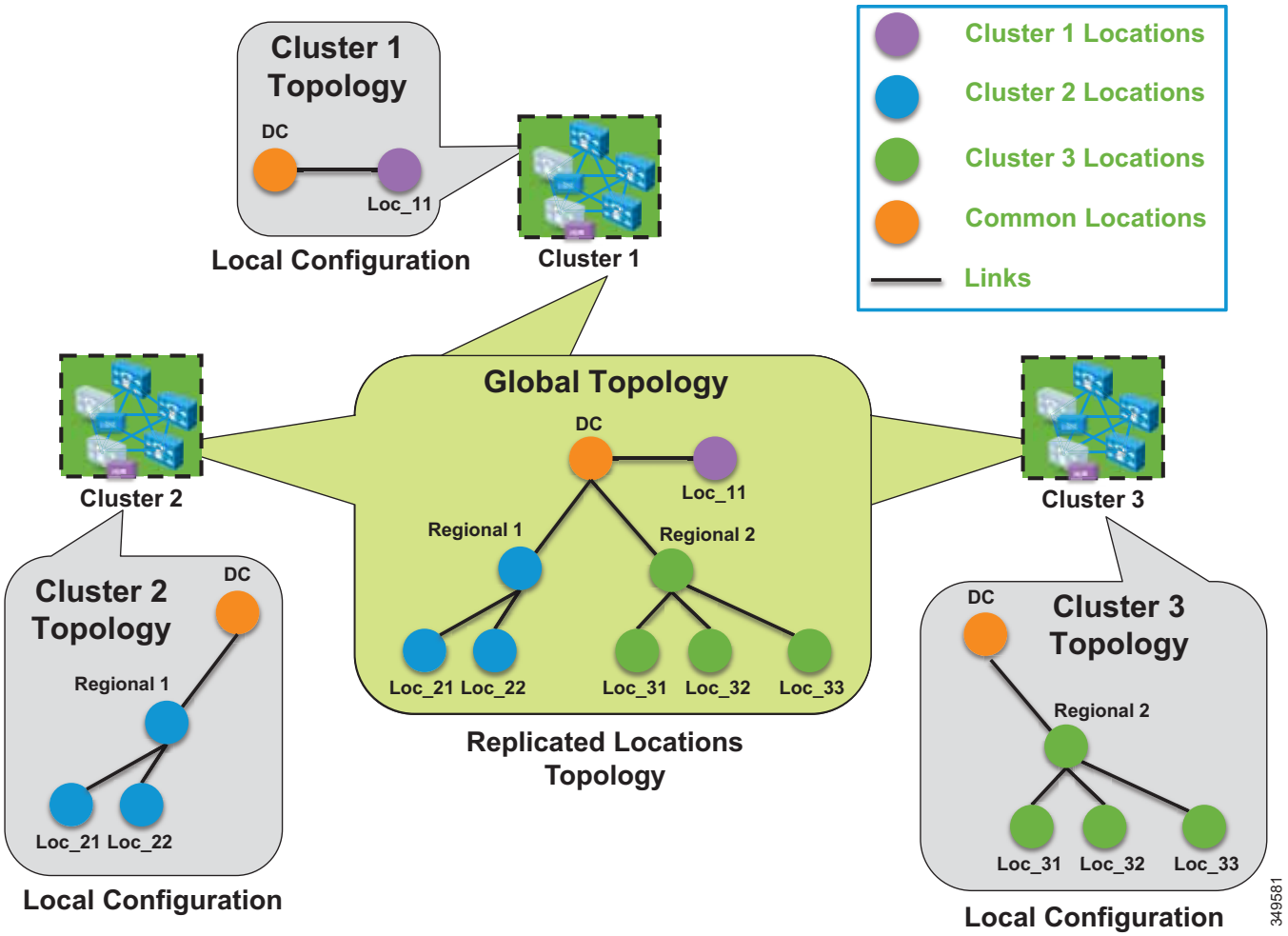
http://www.cisco.com/en/US/products/sw/voicesw/ps556/tsd_products_support_series_home.html

Common Locations (Shared Locations) and Links

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

Common locations and links can be configured across clusters for a number of different reasons. You might have a number of clusters that manage devices in the same physical site and use the same WAN up-links, and therefore the same location needs to be configured on each cluster in order to associate that location to the local devices on each cluster. You might also have clusters that manage their own topology, yet those topologies interconnect at specific locations and you will have to configure those locations as common locations across each cluster so that, when the global topology is being created, the clusters have the common interconnecting locations and links on each cluster to link each remote topology together effectively. Figure 8-18 illustrates linking topologies together and shows the common topology that each cluster shares.

Figure 8-18 Using Common Locations and Links to Create a Global Topology



In Figure 8-18, Cluster 2 has devices in locations Regional 1, Loc_21, and Loc_22, but it requires configuring DC and a link from Regional 1 to DC in order to link to the rest of the global topology. Cluster 3 is similar, with devices in Regional 2 and Loc_31, Loc_32, and Loc_33, and it requires configuring DC and a link from DC to Regional 2 to map into the global topology. Cluster 1 has devices in Loc_11 only, and it requires configuring DC and a link to DC from Loc_11 to map into Cluster 2 and Cluster 3 topologies.

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

Shadow Location

The shadow location is used to enable a SIP trunk to pass Enhanced Locations CAC information such as location name, among other things, required for Enhanced Locations CAC to function between clusters. In order to pass this location information across clusters, the SIP intercluster trunk (ICT) must be assigned to the shadow location. The shadow location cannot have a link to other locations, and therefore no bandwidth can be reserved between the shadow location and other locations. Any device other than a SIP ICT that is assigned to the shadow location will be treated as if it was associated to Hub_None.

Location and Link Management Cluster

In order to avoid configuration overhead and duplicated configuration across clusters that share a large number of locations, a Location and Link Management Cluster can be configured to manage all locations and links in the global topology. All other clusters uniquely configure the locations that they require for location-to-device association and do not configure links or any bandwidth values other than unlimited. The Location and Link Management Cluster is a design concept and is simply any cluster that is configured with the entire global topology of locations and links, while all other clusters in the LBM replication network are configured only with locations set to unlimited bandwidth values and without configured links. When intercluster Enhanced Locations CAC is enabled and the LBM replication network is configured, all clusters replicate their view of the network. The designated Location and Link Management Cluster has the entire global topology with locations, links, and bandwidth values; and once those values are replicated, all clusters use those values because they are the most restrictive. This design alleviates configuration overhead in deployments where a large number of common locations are required across multiple clusters.

Recommendations

Location and Link Management Cluster:

- One cluster should be chosen as the management cluster (the cluster chosen to manage administratively locations and links).
- The management cluster should be configured with the following:
  - All locations within the enterprise will be configured in this cluster.
  - All bandwidth values and weights for all locations and links will be managed in this cluster.

All other clusters in the enterprise:

- All other clusters in the enterprise should configure only the locations required for association to devices but should not configure the links between locations. This link information will come from the management cluster when intercluster Enhanced Locations CAC is enabled. By default there is always a link configured between a newly added location and hub_none. This link should be removed if hub_none is either not used or is not correct in the topology being built.
- When intercluster Enhanced Locations CAC is enabled, all of the locations and links will be replicated from the management cluster.

LBM will always use the lowest, most restrictive bandwidth and lowest weight value after replication.

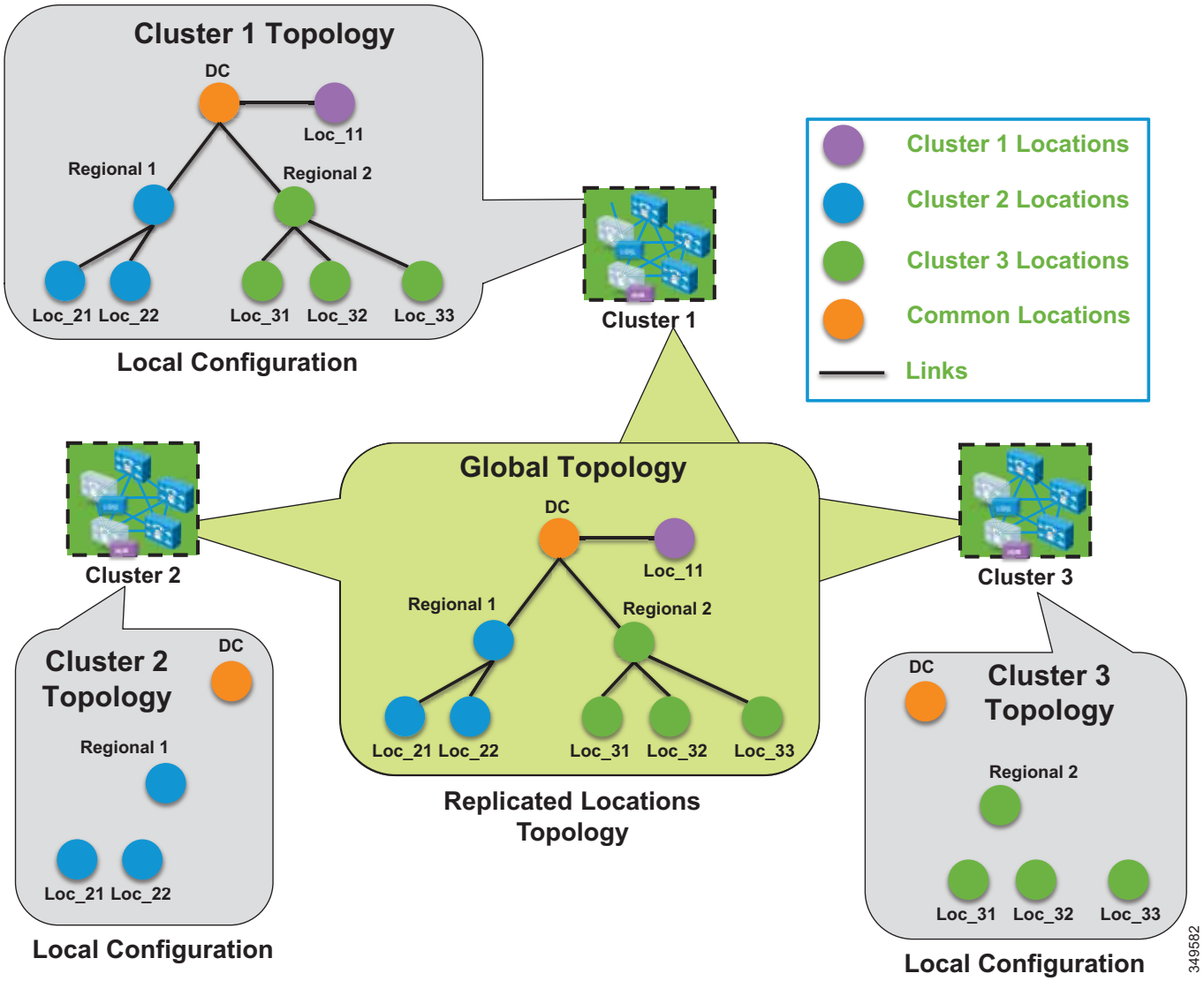
Benefits

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

Figure 8-19 illustrates a Location and Link Management Cluster for three clusters.

Note As mentioned, any cluster can act as the Location and Link Management Cluster. In Figure 8-19, Cluster 1 is the Location and Link Management Cluster.

Figure 8-19 Example of Cluster 1 as a Location and Link Management Cluster



In Figure 8-19 there are three clusters, each with devices in only a regional and remote locations. Cluster 1 has the entire global topology configured with locations and links, and intercluster LBM replication is enabled among all three clusters. None of the clusters in this example share locations, although all of the locations are common locations because Cluster 1 has configured the entire location and link topology. Note that Cluster 2 and Cluster 3 configure only the locations that they require to associate to devices and endpoints, while Cluster 1 has the entire global topology configured. After intercluster replication, all clusters will have the global topology with locations and links.

Design Considerations for Call Admission Control

This section describes how to apply the call admission control mechanisms to various IP WAN topologies. Unified CM Enhanced Locations CAC network modeling support, together with intercluster enhanced locations, can support most of the network topologies in any Unified CM deployment model. Enhanced Locations CAC is still a statically defined mechanism that does not query the network, and therefore the administrator still has to provision Unified CM accordingly whenever network changes affect admission control. This is where a network-aware mechanism such as RSVP can fill that gap and provide support for dynamic changes in the network, such as when network failures occur and media streams take different paths in the network. This is often the case in designs with load-balanced dual or multi-homed WAN up-links or unequally sized primary and backup WAN up-links.

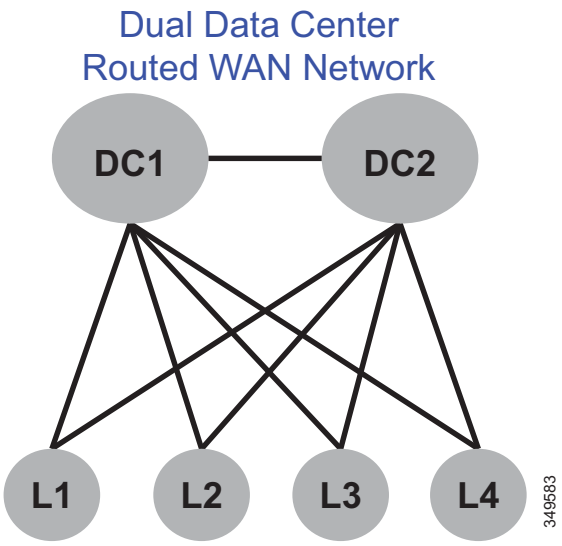
To learn how Enhanced Locations CAC functions, and for more design and deployment details of Enhanced Locations CAC, see the Enhanced Locations Call Admission Control information in the *Bandwidth Management* chapter of the Cisco Collaboration SRND.

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

Dual Data Center Design

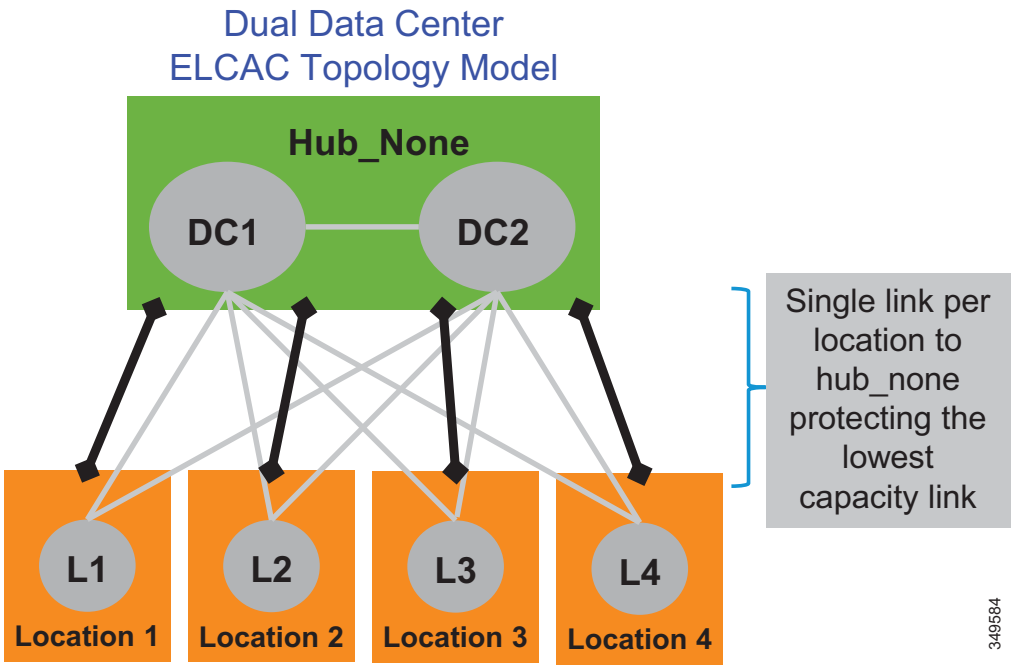
Figure 8-20 illustrates a simple dual data center WAN network design where each remote site has a single WAN up-link to each data center. The data centers are interconnected by a high-speed WAN connection that is over-provisioned for data traffic.

Figure 8-20 Dual Data Center WAN Network



Typically these WAN up-links from the remote sites to the data centers are load-balanced or in a primary/backup configuration, and there are limited ways for a static CAC mechanism to handle these scenarios. Although you could configure this multi-path topology in Enhanced Locations CAC, only one path would be calculated as the effective path and would remain statically so until the weight metric was changed. A better way to support this type of network topology is to configure the two data centers as one data center or hub location in Enhanced Locations CAC and configure a single link to each remote site location. Figure 8-21 illustrates an Enhanced Locations CAC locations and links overlay.

Figure 8-21 Enhanced Locations CAC Topology Model for Dual Data Centers



Design Recommendations

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

- A single location (Hub_None) represents both data centers.
- A single link between the remote locations and Hub_None protects the remote site up-links from over-subscription during normal conditions or failure of the highest bandwidth capacity links.
- The capacity of link bandwidth allocation between the remote site and Hub_None should be equal to the lowest bandwidth capacity for the applicable Unified Communications media for a single link. For example, if each WAN up-link can support 2 Mbps of audio traffic marked EF, then the link audio bandwidth value should be no more than 2 Mbps to support a failure condition or equal-cost path routing.

MPLS Clouds

When designing for Multiprotocol Label Switching (MPLS) any-to-any connectivity type clouds in the Enhanced Locations CAC network model, a single location can serve as the MPLS cloud. This location will not have any devices associated to it, but all of the locations that have up-links to this cloud will have links configured to the location representing the cloud. In this way the MPLS cloud serves as a transit location for interconnecting multiple variable-sized bandwidth WAN up-links to other remote locations.

Design Recommendations

- The MPLS cloud should be configured as a location that does not contain any endpoints but is used as a hub to interconnect locations.
- The MPLS cloud serves as a transit location for interconnecting multiple variable-sized bandwidth WAN up-links to other remote locations.
- Remote sites with connectivity to dual MPLS clouds should treat those connections as a single link and size to the lowest capacity of the links in order to avoid over-subscription during network failure conditions.

Call Admission Control Design Recommendations for Video Deployments

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

QoS is critical to admission control because without it the network has no way of prioritizing the media to ensure that admitted traffic gets the network resources it requires above that of non-admitted or other traffic classifications. Unified CM's CallManager service parameters for QoS as well as the SIP Profile settings provide five main QoS settings that are applicable to endpoint media classification. Table 8-8 shows the five main DSCP parameters along with their default and recommended values and Per Hop Behavior (PHB) equivalents.

Table 8-8 QoS Settings for Endpoint Media Classification

Cisco CallManager Service Parameters Clusterwide Parameters (System - QoS)	Default Values		Recommended Values	
	DSCP	PHB	DSCP	PHB
DSCP for Audio Calls	46	EF	46	EF
DSCP for Video Calls	34	AF41	34	AF41
DSCP for Audio Portion of Video Calls	34	AF41	46	EF
DSCP for TelePresence Calls	32	CS4	34	AF41
DSCP for Audio Portion of TelePresence Calls	32	CS4	46	EF

The **DSCP for Audio Calls** setting is used for any device that makes an audio-only call. The **DSCP for Video Calls** setting is used for the audio and video traffic of any device that is classified as "desktop." **DSCP for TelePresence Calls** is used for the audio and video traffic of any device that is classified as "immersive." The **DSCP for Audio Portion of Video Calls** and **DSCP for Audio Portion of TelePresence Calls** differentiate only the audio portion of video calls, dependent on the classified video call.

Enhanced Locations CAC Design Considerations and Recommendations

The following design recommendation applies to video solutions that employ Enhanced Locations CAC:

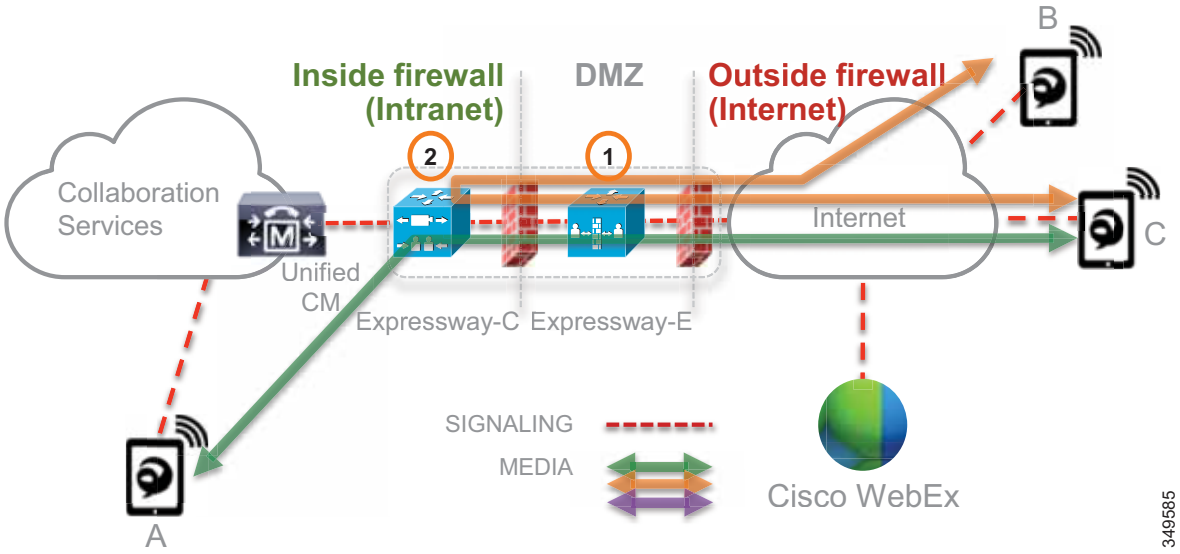
- Intercluster SIP trunks should be associated with the shadow location.

Design Recommendations for Cisco Expressway Deployments with Enhanced Locations CAC

In the Cisco Expressway mobile and remote access (MRA) solution, endpoints supporting the feature can register to Unified CM through a Cisco Expressway deployment without the use of a VPN. Cisco Expressway-C and Expressway-E servers are deployed, each with redundancy for high availability. Expressway-E is placed in the DMZ between the firewall to the Internet (outside) and the firewall to the enterprise (inside), while Expressway-C is placed inside the enterprise. Figure 8-22 illustrates this deployment. It also illustrates the following media flows:

- For Internet-based endpoints calling one another, the media is routed through Cisco Expressway E and Expressway C back out to the Internet, as is illustrated between endpoints B and C in Figure 8-22.
- For Internet-based endpoints calling internal endpoints, the media flows through Expressway-E and Expressway-C, as is illustrated between endpoints A and C in Figure 8-22.

Figure 8-22 Deployment of Cisco Expressway Mobile and Remote Access (MRA)



Enhanced Locations CAC for Cisco Expressway deployments requires the use of a feature in Unified CM called Device Mobility. Enabling Device Mobility on the endpoints allows Unified CM to know when the device is registered through Cisco Expressway or when it is registered from within the enterprise. Device Mobility also enables Unified CM to provide admission control for the device as it roams between the enterprise and the Internet. Device Mobility is able to do this by knowing that, when the endpoints register to Unified CM with the IP address of Expressway-C, Unified CM will associate the applicable Internet location. However, when the endpoint is registered with any other IP address, Unified CM will use the enterprise location that is configured directly on the device (or from the device pool directly configured on the device). It is important to note that Device Mobility does not have to be deployed across the entire enterprise for this function to work. Configuration of Device Mobility in Unified CM is required only for the Expressway IP addresses, and the feature is enabled only on the devices that require the function (that is to say, those devices registering through the Internet).

Bandwidth Management Deployment

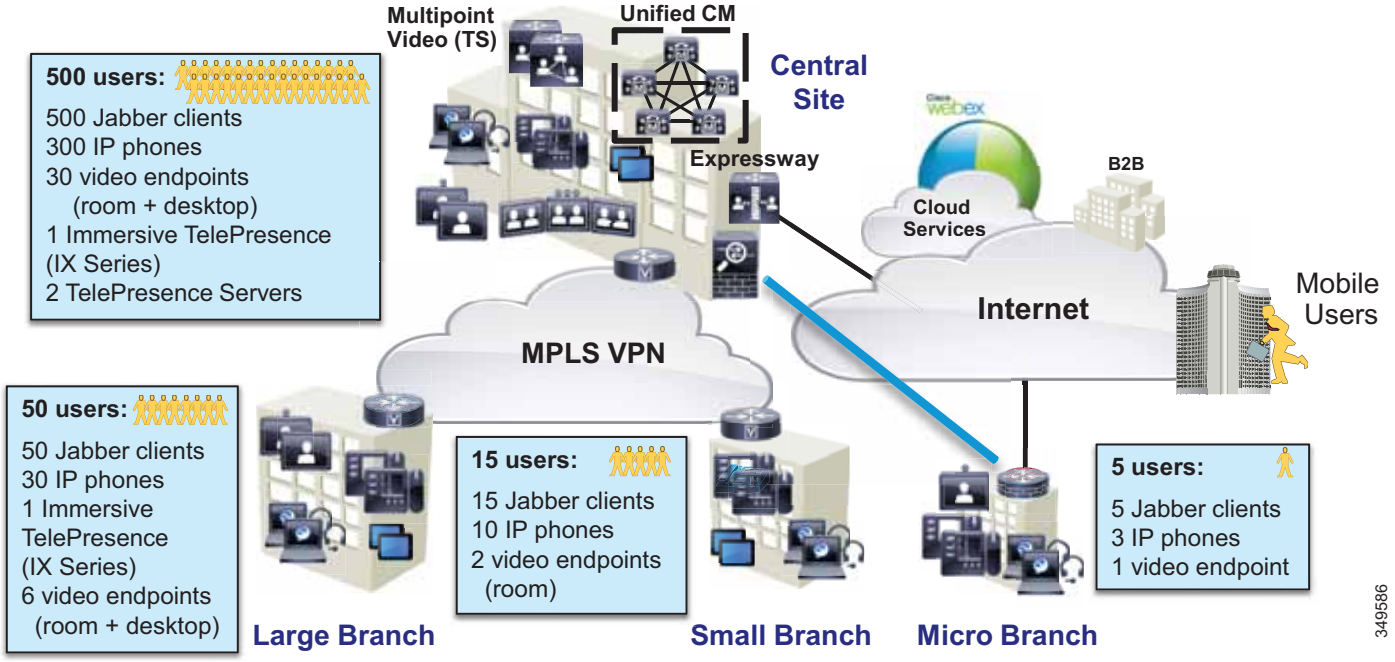
This section describes how to deploy bandwidth management for the PA. It explores all aspects discussed previously in this chapter, including identification and classification, WAN queuing and scheduling, provisioning, resource control, and bandwidth allocation guidelines for each site type.

Deployment Overview

The Preferred Architecture example in this section is a large enterprise with users across a large geographic area and with a headquarters site where the data center sits as well as multiple large, small, and micro sized branches with roughly 500, 50, 15, and 5 users in each branch type, respectively. To simplify the illustration of the network, these categories of sites (headquarters, large, small, and micro) are used as a template to size bandwidth considerations for each site that has a similar size user base and endpoint density. Figure 8-23 illustrates this with numbers of users and endpoints at each type of site. The enterprise in this example has deployed Jabber with video to ensure that users have access to a video terminal for conferencing. The video conferencing resources are located in the data center at the Headquarters Site. IP phones are for voice-only communications. Video endpoints are Jabber clients, collaboration desktop endpoints (DX Series), and room-based endpoints (MX Series and SX Series). The Headquarters Site has an immersive unit such as the IX Series.

The IT department is tasked with determining the bandwidth requirements for the WAN edge for each type of site. Each section below lists the requirements and illustrates a methodology for applying QoS, determining bandwidth and queuing requirements, and determining admission control requirements.

Figure 8-23 Preferred Architecture for Enterprise Collaboration



Deployment of bandwidth management for the Enterprise Collaboration Preferred Architecture involves the following major tasks:

- Identification and Classification
  - Access Layer Endpoint Identification and Classification
  - Application Server QoS
  - WAN Edge Identification and Classification
  - WAN Edge Queuing and Scheduling
- Provisioning and Admission Control
  - Enhanced Locations CAC
  - Deploy Device Mobility for Mobile and Remote Access (MRA)
  - Bandwidth Allocation Guidelines

Identification and Classification

In this phase the QoS requirements are established across the enterprise. The topics covered in this section include:

- Access Layer Endpoint Identification and Classification
  - Endpoints: Jabber
  - Endpoints: Desktop and TelePresence
- Application Server QoS
- WAN Edge Identification and Classification
- WAN Edge Queuing and Scheduling

This phase of the deployment involves the following high-level steps:

1. Configure endpoints in Unified CM with QoS for Jabber clients and desktop and telepresence endpoints.
2. Deploy an access layer policy for endpoint identification and classification for untrusted endpoints.
3. Configure application server QoS for media and SIP signaling.
4. Deploy a WAN Edge ingress marking policy for collaboration media and SIP signaling.
5. Deploy a WAN Edge egress queuing policy for collaboration media and SIP signaling.

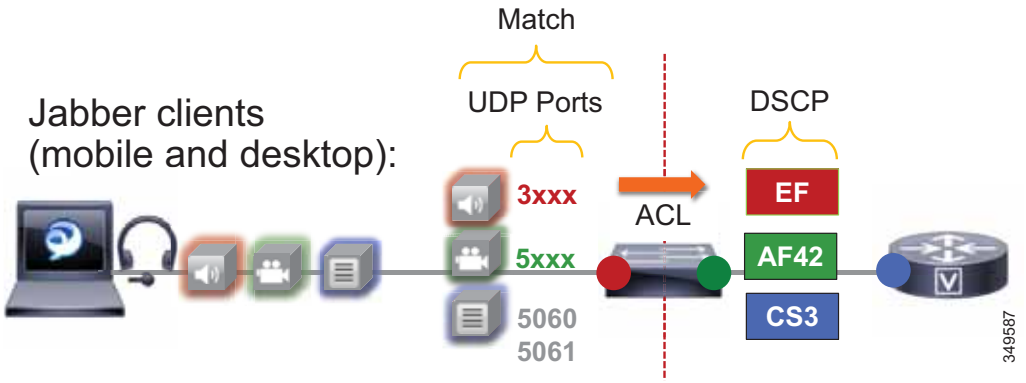
Access Layer Endpoint Identification and Classification

In this section endpoint QoS and media port ranges are configured in the network and in Unified CM.

Endpoints: Jabber

Jabber endpoints are untrusted and typically sit in the data VLAN. Specific UDP port ranges are used to re-mark signaling and media at the access layer switch. In this case Unified CM is configured with a SIP Profile specifically for all Jabber clients to use the **Separate Media and Signaling Port Range** value of 3000 to 3999 for audio and 5000 to 5999 for video. The SIP signaling port of 5060 is used for SIP signaling and 5061 for secure SIP signaling. The SIP signaling port is configured in the SIP Security Profile in Unified CM. This is illustrated in Figure 8-24.

Figure 8-24 Jabber Endpoint QoS



The administrator creates an ACL for the access switches for the data VLAN to re-mark UDP ports to the following DSCP values:

- Audio: UDP ports 3000 to 3999 marked as EF
- Video: UDP ports 5000 to 5999 marked as AF42
- Signaling: TCP ports 5060 to 5061 marked as CS3

Jabber classification summary:

- Audio streams of all Jabber calls (voice-only and video) are marked as EF
- Video streams of Jabber video calls are marked as AF42

For the Jabber endpoints, we also recommend changing the default QoS values in the Jabber SIP Profile. This is to ensure that, if for any reason the QoS is "trusted" via a wireless router or any other network component, then the correct "trusted" values are the same as they would be for the re-marked value. Therefore, the QoS parameters in the SIP Profile should be set as listed in Table 8-9, and the UDP port ranges should be set as listed in Table 8-10.

Table 8-9 QoS Parameter Settings in SIP Profile for Jabber Endpoints

QoS Service Parameter Name (SIP Profile)	Default Value	Changed Value
DSCP for Audio Calls	EF	No change
DSCP for Video Calls	AF41	AF42
DSCP for Audio Portion of Video Calls	AF41	EF
DSCP for TelePresence Calls	CS4	AF41
DSCP for Audio Portion of TelePresence Calls	CS4	EF

Table 8-10 UDP Port Settings for Jabber Endpoints

Media Port Ranges > Separate Port Range for Audio and Video	Value
Audio start port	3000
Audio stop port	3999
Video start port	5000
Video stop port	5999

The settings in Table 8-9 ensure that audio of Jabber clients is set to EF, and the video will be set to AF42 if for any reason the traffic goes through a trusted path and is not re-marked via UDP port range at the access switch. This is simply to ensure a consistent configuration across Jabber endpoints.

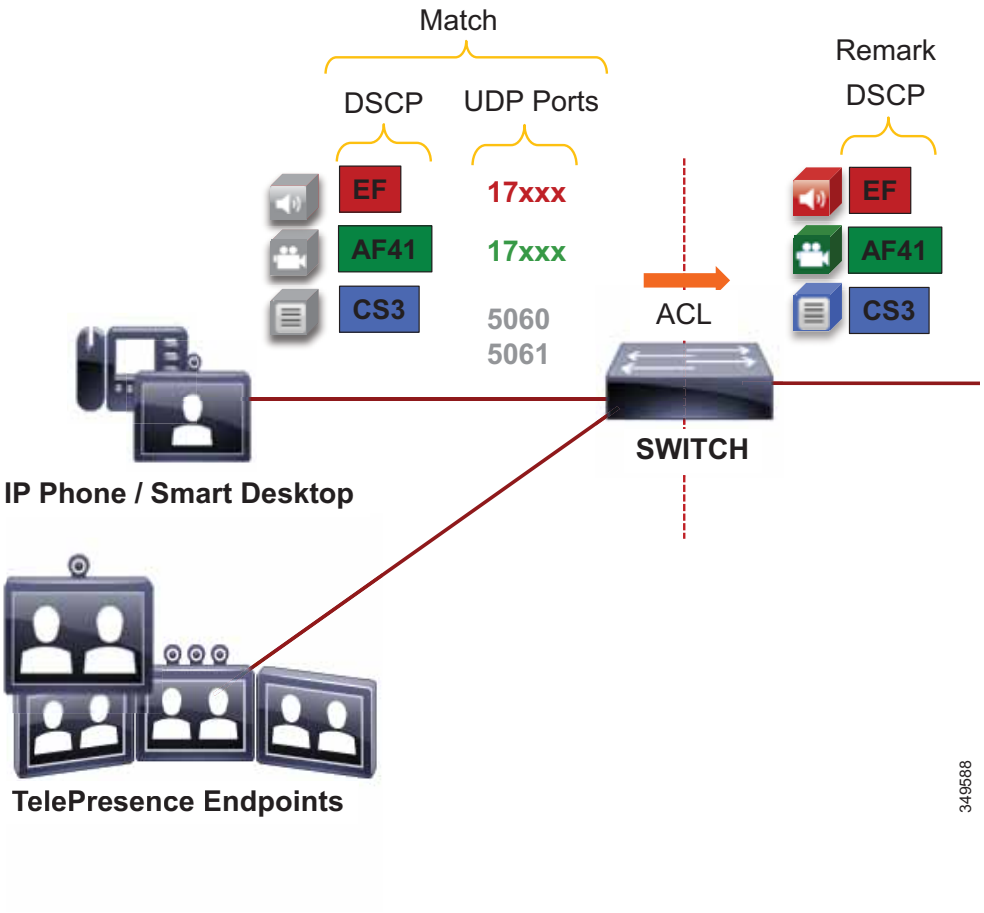
For Jabber on mobile devices, we recommend copying the **Standard SIP Profile for Mobile Device** when building a new SIP profile for these devices, because the default standard SIP profile for mobile devices includes recommended timer values for maintaining Jabber registration on Android and Apple iOS devices. These timers are required for any SIP profile assigned to dual-mode and tablet Jabber client devices.

Note Jabber for Mac, iPad, iPhone, and Android all natively mark DSCP by the OS. Jabber for Windows, however, requires Group Policy Objects to re-mark DSCP by the OS. Without Group Policy Objects, Jabber for Windows will mark all traffic with a DSCP of 0. This is why specific port ranges are used for Jabber and without matching on DSCP.

Endpoints: Desktop and TelePresence

IP phones, smart desktop, and TelePresence endpoints also rely on an access layer switch ACL to re-mark traffic. Specific UDP port ranges and DSCP are used to re-mark signaling and media at the access layer switch. In this case Unified CM is configured with a SIP Profile specifically for all IP phones, smart desktop, and TelePresence endpoints to use the common Media and Signaling Port Range value of 17000 to 17999 for audio and video. The SIP signaling port of 5060 is used for SIP signaling and 5061 for secure SIP signaling. The SIP signaling port is configured in the SIP Security Profile in Unified CM. This is illustrated in Figure 8-25.

Figure 8-25 Desktop and TelePresence Endpoint QoS



The administrator creates an ACL for the access switch ports to re-mark UDP ports to the following DSCP values:

- Audio: UDP ports 17000 to 17999 with DSCP of EF to be re-marked as EF
- Video: UDP ports 17000 to 17999 with DSCP of AF41 to be re-marked as AF41
- Signaling: TCP ports 5060 to 5061 marked as CS3

Desktop and TelePresence endpoint classification summary:

- Audio streams of all desktop and TelePresence endpoint calls (voice-only and video) are marked EF.
- Video streams of desktop and TelePresence endpoint video calls are marked AF41.

For the desktop and TelePresence endpoints, the default QoS values must be changed in the SIP Profile and set as shown in Table 8-11, and the UDP port ranges should be set as listed in Table 8-12.

Table 8-11 QoS Parameters in SIP Profile for Desktop and TelePresence Endpoints

QoS Service Parameter Name (SIP Profile)	Default Value	Changed Value
DSCP for Audio Calls	EF	No change
DSCP for Video Calls	AF41	No change
DSCP for Audio Portion of Video Calls	AF41	EF
DSCP for TelePresence Calls	CS4	AF41
DSCP for Audio Portion of TelePresence Calls	CS4	EF

Table 8-12 UDP Port Settings for Desktop and TelePresence Endpoints

Media Port Ranges > Common Port Range for Audio and Video	Value
Media start port	17000
Media stop port	17999

Example Switch ACL-Based QoS Policy for Endpoint Switch Ports

Desktop and TelePresence endpoints:

- Match UDP port range 17xxx with DSCP EF –> Re-mark to DSCP EF
- Match UDP port range 17xxx with DSCP AF41 –> Re-mark to DSCP AF41
- Match TCP ports 5060 to 5061 –> Re-mark to DSCP CS3

Jabber clients

- Match UDP port range 3xxx –> Re-mark to DSCP EF
- Match UDP port range 5xxx –> Re-mark to DSCP AF42
- Match TCP ports 5060 to 5061 –> Re-mark to DSCP CS3

Generic matching

- Matches the rest of the traffic and sets DSCP to 0 (Best Effort or BE) using a default class-map

Note The following is an example access control list based on the Cisco Common Classification Policy Language (C3PL).

```
! This section configures the ACLs to match the UDP port ranges and DSCP.
ip access-list extended QOS_VOICE
   permit udp any range 17000 17999 any dscp ef
   permit udp any range 3000 3999  any
ip access-list extended QOS_PRIORITIZED_VIDEO
   permit udp any range 17000 17999 any dscp af41
ip access-list extended QOS_JABBER_VIDEO
   permit udp any range 5000 5999 any
ip access-list extended QOS_SIGNALING
   permit tcp any any range 5060 5061
   permit tcp any range 5060 5061 any
```

```
! This section configures the classes that match on the ACLs above.
class-map match-any VOICE
  match access-group name QOS_VOICE
class-map match-any PRIORITIZED_VIDEO
  match access-group name QOS_PRIORITIZED_VIDEO
class-map match-any JABBER_VIDEO
  match access-group name QOS_JABBER_VIDEO
class-map match-any SIGNALING
  match access-group name QOS_SIGNALING

! This section configures the policy-map matching the classes configured above and sets
DSCP for voice, video, and SIP signaling on ingress. Note that the class-default sets
everything that does not match the above to a DSCP of 0 (BE).
policy-map INGRESS_MARKING
 class VOICE
   set dscp ef
 class PRIORITIZED_VIDEO
   set dscp af41
 class JABBER_VIDEO
   set dscp af42
 class SIGNALING
   set dscp cs3
class class-default
   set dscp 0

! This section applies the policy-map to the interface.
  Switch (config-if)# service-policy input INGRESS-MARKING
```

As mentioned, endpoints send and receive other data and signaling such as ICMP, DHCP, TFTP, BFCP, LDAP, XMPP, FECC, CTI, and so forth. The QoS values for this traffic should follow the enterprise's best practices for each type of traffic. Without doing this step, all other traffic apart from media and SIP signaling will be set to a DSCP of BE (DSCP 0) by the class-default in this configuration. We recommend either passing through the traffic marking by matching on DSCP and then re-marking the DSCP to the same value, or else using the TCP and UDP ports for each protocol that the endpoints use for communications.

The following example illustrates this. A class-map is created to match on a DSCP of AF21 which is transactional data, and the policy sets that data to AF21, effectively re-marking the DSCP to the same value. This is simply an example of matching on a DSCP to re-mark to the same DSCP:

```
class-map match-any TRANSACTIONAL-DATA
  match dscp af21

policy-map INGRESS_MARKING
….
 class TRANSACTIONAL-DATA
   set dscp af21
```

TCP and UDP port ranges can also be used. For more information on the TCP and UDP ports used for communication between the endpoints and Unified CM, see the *Cisco Unified Communications Manager TCP and UDP Port Usage* information in the *System Configuration Guide for Cisco Unified Communications Manager*, available at

http://www.cisco.com/c/en/us/support/unified-communications/unified-communications-manager-callmanager/products-installation-and-configuration-guides-list.html

Also see the endpoints administration guides or the Jabber planning guide to determine the various protocols and ports used for other endpoint traffic. Some examples of these documents include:

- *Cisco DX Series Administration Guide*, available at

  http://www.cisco.com/c/en/us/support/collaboration-endpoints/desktop-collaboration-experience-dx600-series/products-maintenance-guides-list.html

- *Cisco Jabber Planning Guide*, available at

  http://www.cisco.com/c/en/us/support/unified-communications/jabber-android/products-installation-guides-list.html

Application Server QoS

Configure QoS on all media originating and terminating applications and MCUs across the solution. This section covers non-default configuration on all application servers in the PA. It is also equally important to ensure that the switch ports to which the application servers are connected trust the QoS set by the servers. Some switches such as the Cisco Catalyst 3850 Series trust the QoS by default, so verify the switch configuration to ensure that the switch port is trusted by default or enable QoS trust.

QoS settings for the various application servers:

- Cisco Unified CM (endpoint)

  System > **Service Parameters** > **Select Publisher** > **Select Cisco CallManager Service** > **Clusterwide Parameters (System - QOS)** > Change the QoS values from their defaults and set them as indicated in Table 8-13.

Table 8-13 QoS Parameter Settings for Unified CM Endpoints

QoS Service Parameter Name (SIP Profile)	Default Value	Changed Value
DSCP for Audio Calls	EF	No change
DSCP for Video Calls	AF41	No change
DSCP for Audio Portion of Video Calls	AF41	EF
DSCP for TelePresence Calls	CS4	AF41
DSCP for Audio Portion of TelePresence Calls	CS4	EF

- Cisco Unity Connection

  System settings > **Advanced** > **Telephony**

  – Default = Audio (46 / EF), Video (46 / EF), Signaling (24 / CS3)

  – Change Video to 34 / AF41

- Cisco Meeting Server

  – Cisco Meeting Server DSCP settings are configured through the command line interface (CLI). The DSCP settings should be configured after Cisco Meeting Server has been configured as indicated in the chapter on Conferencing, page 3-1. The default for all values is DSCP 0, therefore all DSCP values need to be configured. *These changes require a server restart.*

  – Command line values:

```
dscp 4 signaling 24
dscp 4 voice 46
dscp 4 multimedia 34
dscp 4 oa&m 24
```

- Cisco Expressway

  The Expressway DSCP value for video is set to opportunistic video using a value of AF42 (DSCP 36). All other values (Audio, Signaling, and XMPP) are set to default values.

  System > **Quality of Service**

  – DSCP Signaling value 24 (default)

  – DSCP Audio value 46 (default)

  – DSCP Video value 36

  – DSCP XMPP value 24 (default)

WAN Edge Identification and Classification

At the WAN edge on ingress from the enterprise to the service provider, it is expected that the packets that arrive with a specific DSCP value because the collaboration traffic have been re-marked at the access layer switch. On ingress it is important to re-mark any traffic at the WAN edge that could not be re-marked at the access layer, as a failsafe in case any traffic from the access switches was trusted through the LAN. While QoS is important in the LAN, it is paramount in the WAN; and as routers assume a trust on ingress traffic, it is important to configure the correct QoS policy that aligns with the business requirements and user experience. The WAN edge re-marking is always done on the ingress interface into the router, while the queuing and scheduling is done on the egress interface. The following example walks through the WAN ingress QoS policy as well as the egress queuing policy. Figure 8-26 through Figure 8-31 illustrate the configuration and the re-marking process.

In Figure 8-26 the packets from endpoints are identified and classified with the appropriate DSCP marking via a trusted port or via an ACL. Because there are typically areas of the switched access network that either cannot be configured with the correct QoS policies or re-mark collaboration traffic to Best Effort DSCP (BE), the WAN ingress policy is a good place for a catch-all policy to readdress what the access layer might have missed before the traffic heads into the WAN.

Figure 8-26 Example Router Ingress QoS Policy Process – 1

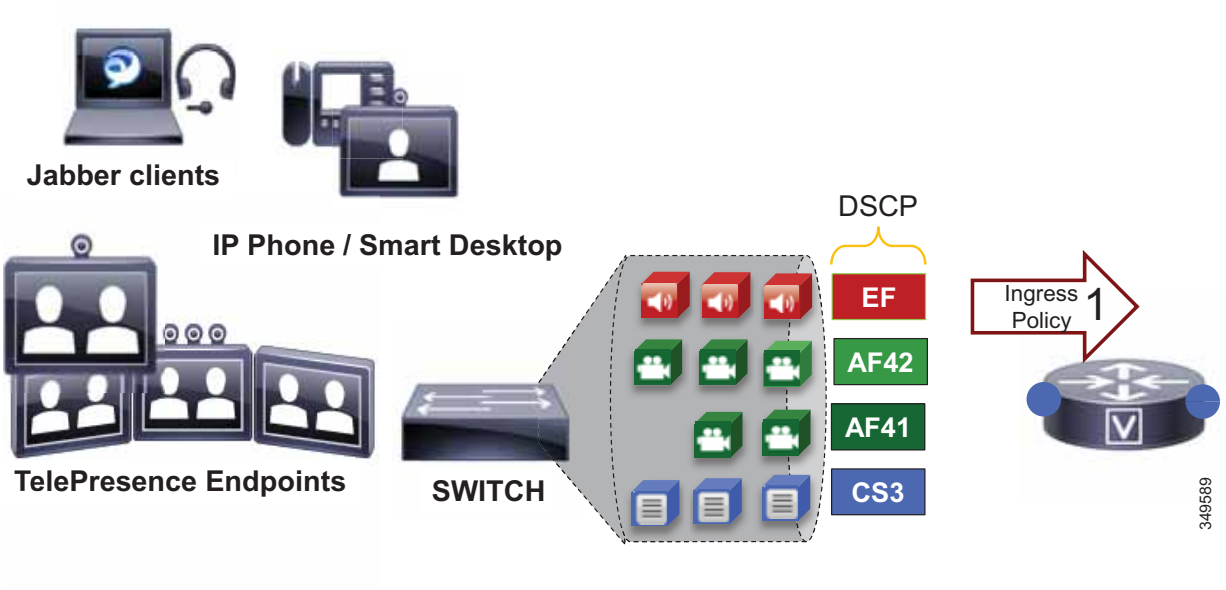
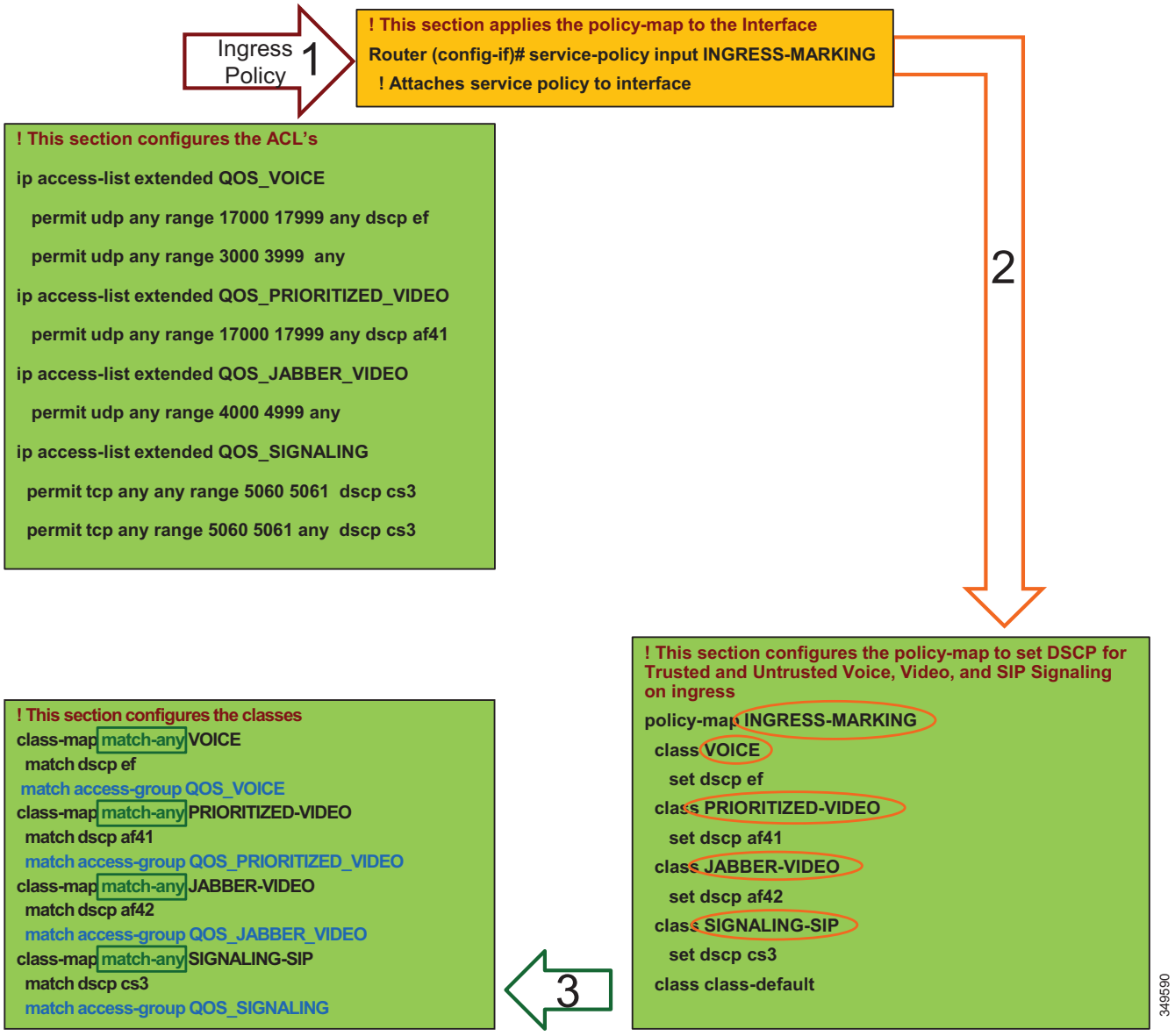


Figure 8-27 through Figure 8-29 illustrate the policy matching criteria and DSCP re-marking. The illustrations show the following steps:

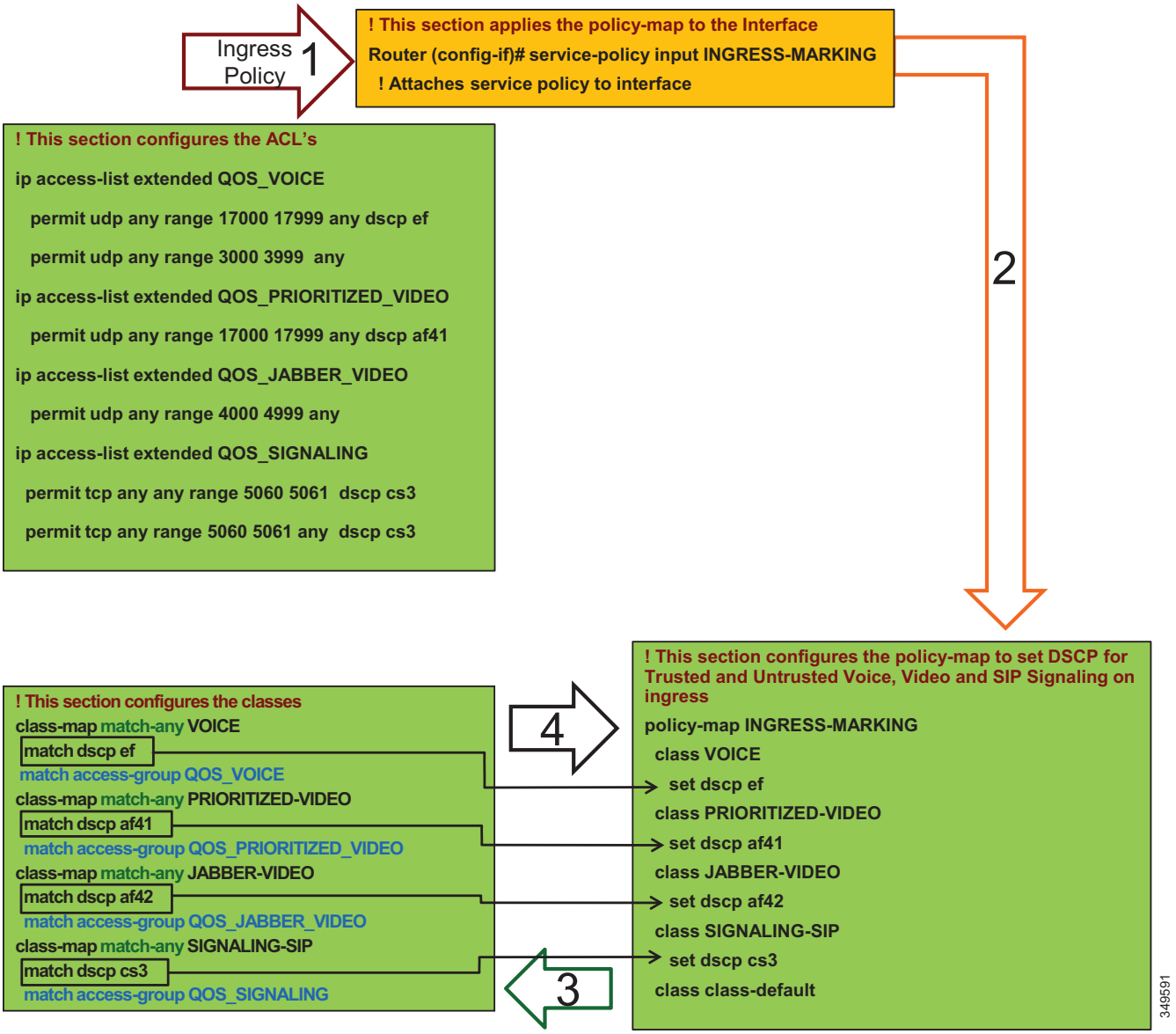
1. In step 1, packets arrive at the router ingress interface, which is configured with an input service policy.
2. In step 2, the policy-map is configured with four classes of traffic to set the appropriate DSCP (voice = EF; prioritized-video = AF41; Jabber-video = AF42; signaling = CS3).
3. In step 3, each one of these classes matches a class-map of the same name configured with match-any criterion. This match-any criterion means that the process will start top-down, and the first matching criterion will be executed to set the DSCP according to each class in the policy-map statements.

Figure 8-27 Example Router Ingress QoS Policy Process – 2



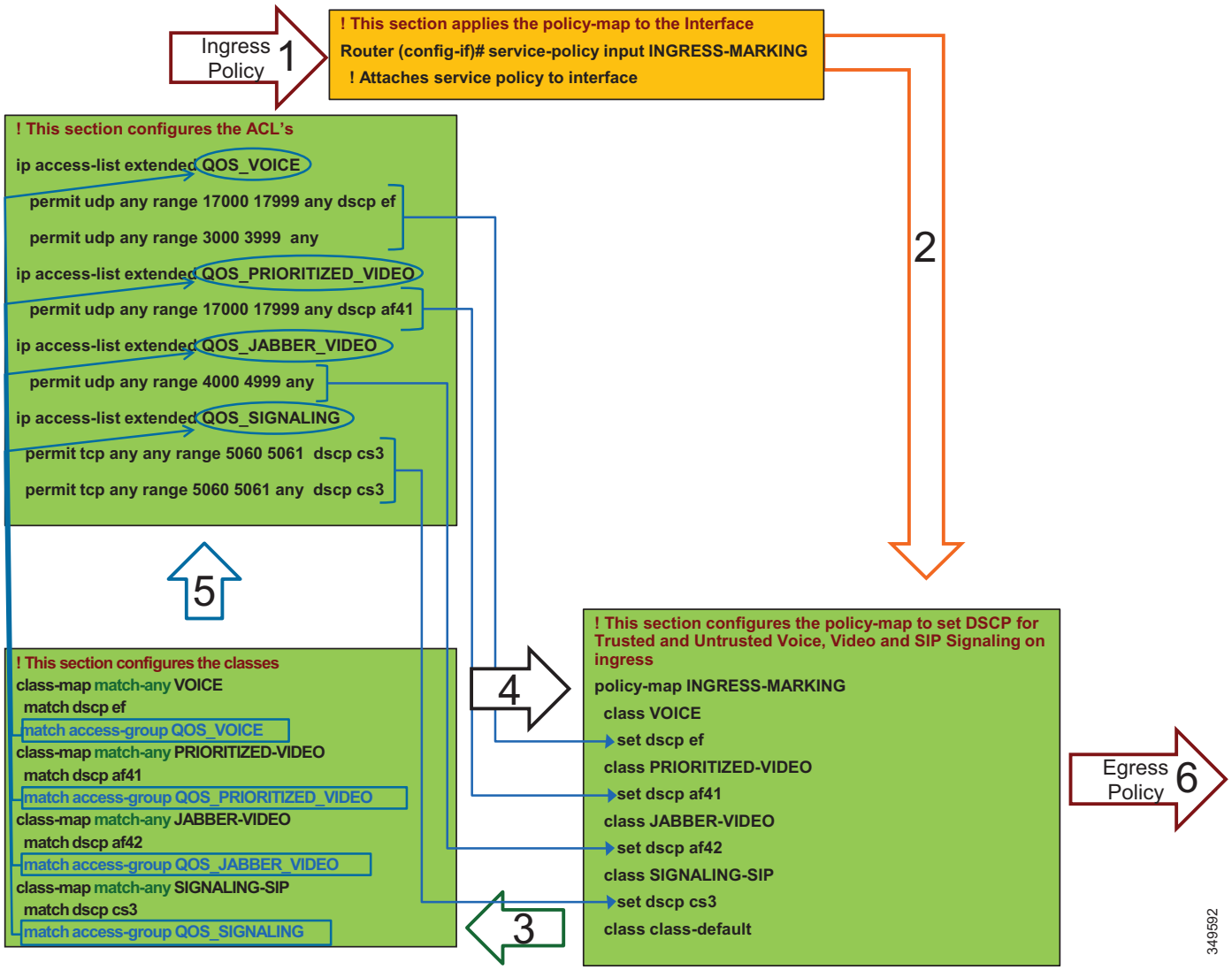
4. In step 4, the first match statement in the class-map is **match dscp**. If the traffic matches the DSCP, then DSCP is set again to the same value that was matched and as is configured in the policy-map statements. In this case the router is simply matching on DSCP and resetting the DSCP to the same value.

Figure 8-28 Example Router Ingress QoS Policy Process – 3



5. In step 5, if DSCP was not matched, then the next line in the class-map statement is parsed, which is the ACL that matches the UDP ports set in Unified CM for the Jabber clients in the Identification and Classification section. When the ACL criteria are met (protocol, port range, and/or DSCP), then the traffic is set as is configured in the corresponding policy-map statements.

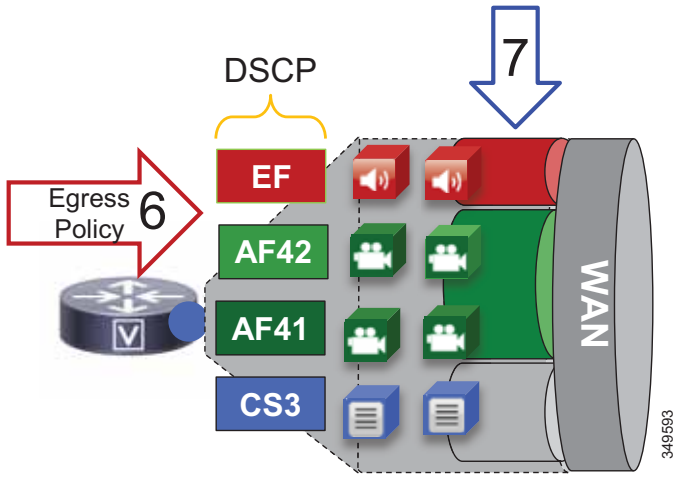
Figure 8-29 Example Router Ingress QoS Policy Process – 4



Note This is an example QoS ingress marking policy based on the Cisco Common Classification Policy Language (C3PL). Refer to your specific router configuration guide for any updated C3PL commands and for information on how to configure a similar policy on a Cisco router supporting C3PL.

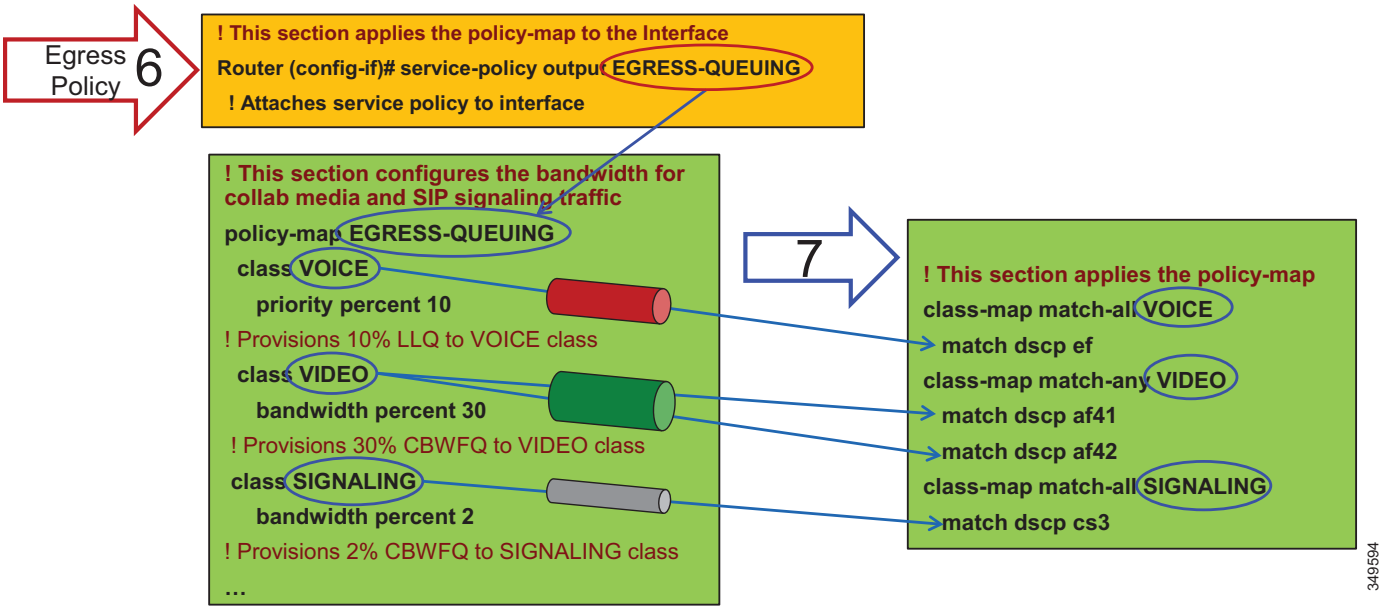
6. In step 6, the traffic goes to an outbound interface to be queued and scheduled by an output service policy that has three queues created: a Priority Queue called VOICE, a CBWFQ called VIDEO, and another CBWFQ called SIGNALING. This is illustrated in Figure 8-30 through Figure 8-31. This highlights the fact that the egress queuing policy is based only on DSCP as network marking occurring at the access switch and/or on ingress into the WAN router ingress interface. This is an example simply to illustrate the matching criteria and queues, and it does not contain the WRED functionality. For information on WRED, see the WAN Edge Queuing and Scheduling section.

Figure 8-30 Example Router Egress Queuing Policy Process – 1



7. In step 7, the traffic is matched against the class-map match statements. All traffic marked EF goes to the VOICE PQ, AF41 and AF42 traffic goes to the VIDEO CBWFQ, and CS3 traffic goes to the SIGNALING CBWFQ.

Figure 8-31 Example Router Egress Queuing Policy Process – 2



Note This is an example egress queuing policy based on the Cisco Common Classification Policy Language (C3PL). Refer to your specific router configuration guide for any updated C3PL commands and for information on how to configure a similar policy on a Cisco router supporting C3PL.

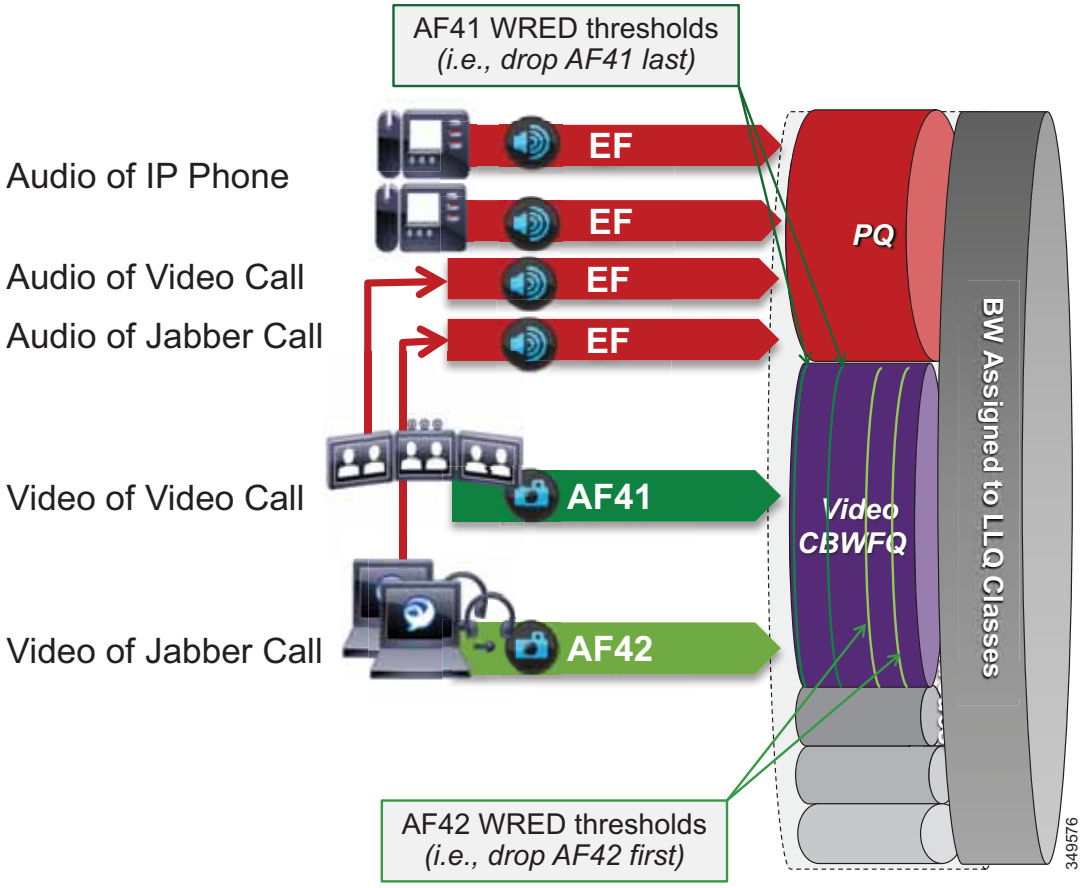
Example Configuration of Egress Queuing

```
! This section applies the policy-map classes to match media and signaling QoS.
class-map match-any VIDEO
 match dscp af41
 match dscp af42
class-map match-any VOICE
 match dscp ef
class-map match-any SIGNALING
 match dscp cs3

! This section configures the bandwidth for Collaboration media and SIP signaling traffic.
policy-map EGRESS-QUEUING
 class VOICE
  priority percent 10
 class VIDEO
  bandwidth percent 30
  fair-queue
 class SIGNALING
  bandwidth percent 2
…

! This section applies the policy-map to the interface.
Router (config-if)# service-policy output EGRESS-QUEUING
  ! Attaches service policy to interface
```

WAN Edge Queuing and Scheduling

This section covers the interface queuing. Figure 8-32 illustrates the voice PQ, video CBWFQ, and WRED thresholds used for the CBWFQ.

Figure 8-32 Queuing and Scheduling Collaboration Media

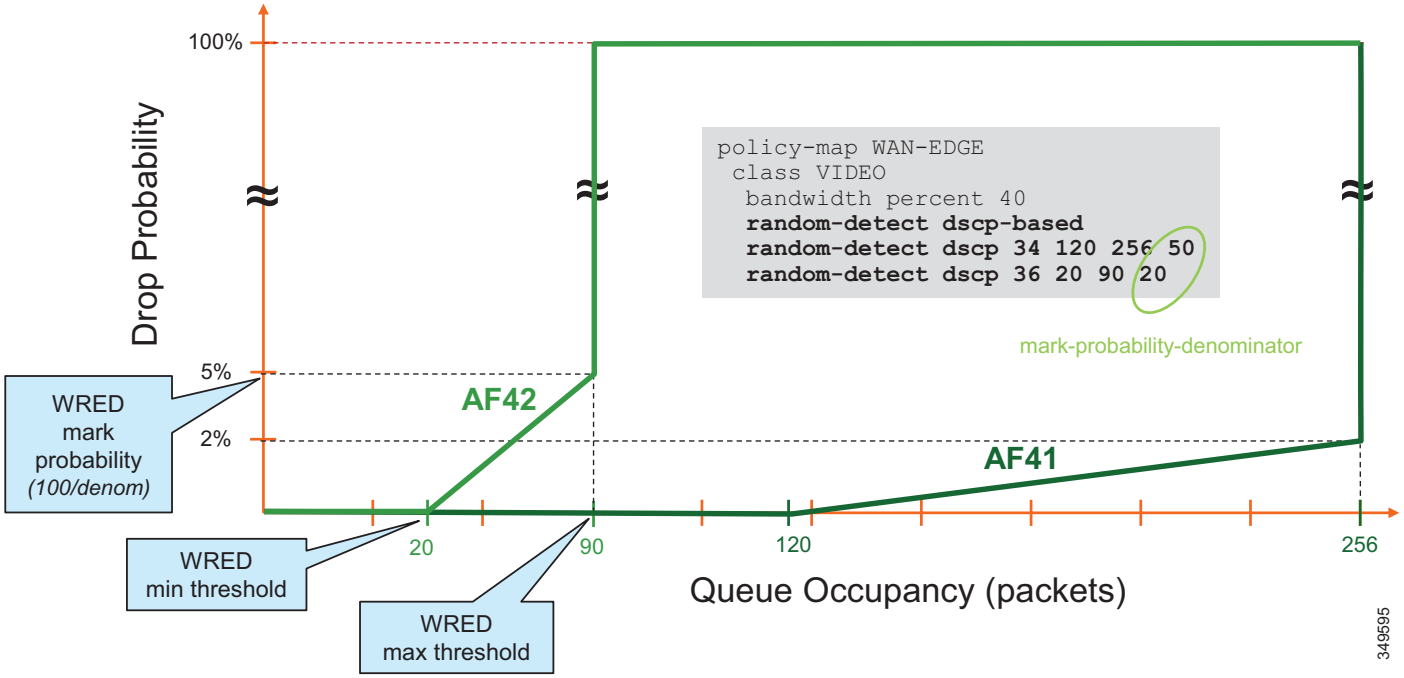


- All audio from all endpoints marked EF is mapped to the PQ.
- Video calls and Jabber share the same CBWFQ.
  - EF for audio streams of video calls from endpoints
  - AF41 for video streams of video calls from endpoints
  - EF for audio streams of all calls from Jabber clients
  - AF42 for video streams of video calls from Jabber clients
- WRED is configured on the video queue.
  - Minimum to maximum thresholds for AF42: approximately 10% to 30% of queue limit
  - Minimum to maximum thresholds for AF41: approximately 45% to 100% of queue limit

Weighted Random Early Detection (WRED) threshold minimum and maximum values are configured in the Video CBWFQ. To illustrate how the WRED thresholds are configured, assume that the interface had been configured with a queue depth of 256 packets. Then following the guidelines above, the WRED

minimum and maximum thresholds for AF42 and AF41 would be configured as illustrated in Figure 8-33.

Figure 8-33 Threshold Example for Video CBWFQ with WRED



Recommended WRED Thresholds

Figure 8-34 lists the WRED thresholds for each traffic class (AF41 and AF42) and the recommended mark probability denominators that have been tested for various link speeds. These are just examples, and testing and customization are expected based on the amount of traffic in each traffic class and the aggressiveness required in the WRED drop probability during the busy hour.

Figure 8-34 Recommended WRED Thresholds by Link Speed

WAN Link Speed / WRED Values		622 Mbps (OC12)	155 Mbps (OC3)	34-44 Mbps (E3/DS3)	10 Mbps	5 Mbps
AF41	min-threshold	240	180	120	60	60
	max-threshold	512	384	256	128	128
	mark-probability-denominator	50	50	50	50	50
AF42	min-threshold	40	30	20	15	15
	max-threshold	180	135	90	40	40
	mark-probability-denominator	20	20	20	20	20
Video queue bandwidth %		43	53	55	40	30

349596

The following example configuration is for WRED in the video Class-Based Weighted Fair Queue (CBWFQ) of a DS3 link (44 Mbps).

```
policy-map EGRESS-QUEUING
 class VOICE
  priority percent 10
 class VIDEO
  bandwidth percent 30
  random-detect dscp-based
  random-detect dscp 34 120 256 50
  random-detect dscp 36 20 90 20
fair-queue
 class SIGNALING
  bandwidth percent 2
```

Note The WRED values might have o be customized for the specific environment. For example, if there is a much larger amount of AF42 traffic than AF41 traffic, then it makes sense to adjust the WRED threshold variables to suit those cases. Tweaking the variables and monitoring levels of drop is always the best way to achieve the desired results.

Provisioning and Admission Control

This section addresses admission control and provisioning bandwidth to the queues for each site type. It covers the following topics:

- Enhanced Locations CAC
  - Region Configuration
  - Deploy Enhanced Locations Call Admission Control
- Deploy Device Mobility for Mobile and Remote Access (MRA)
- Bandwidth Allocation Guidelines

This phase of the deployment involves the following high-level steps:

1. Configure Enhanced Locations CAC.
2. Configure a region matrix for maximum video bit rate groups.
3. Deploy Device Mobility for mobile and remote access (MRA) endpoints.
4. Follow bandwidth allocation guidelines.

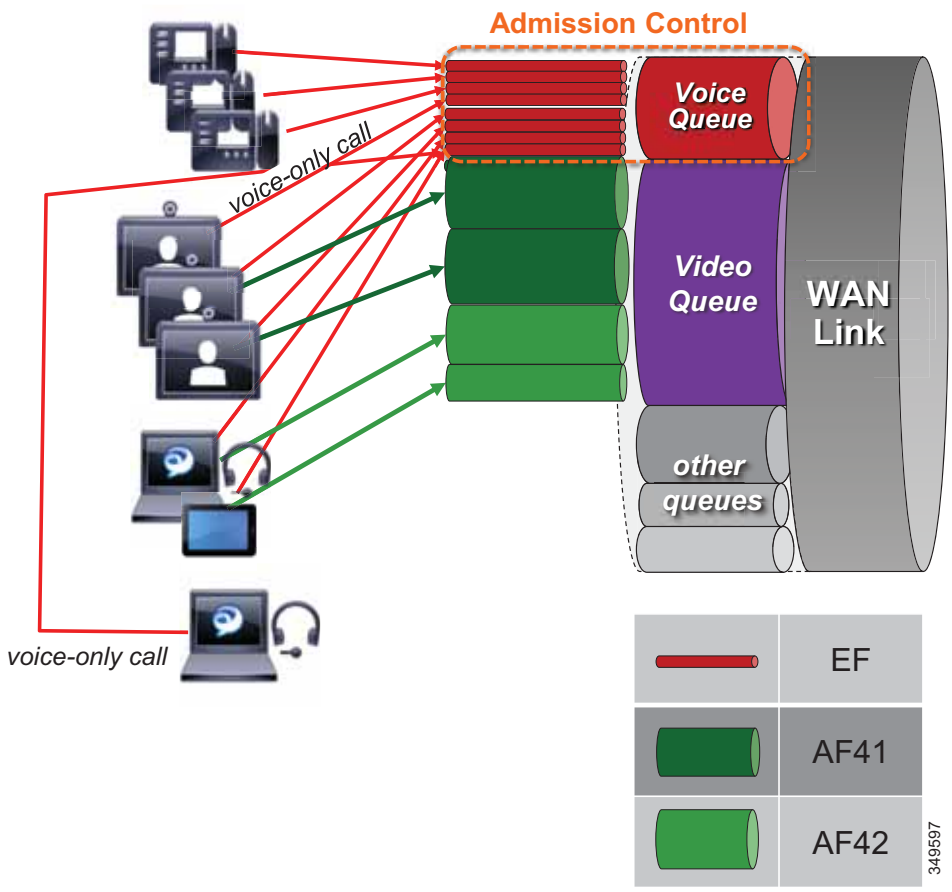
Enhanced Locations CAC

Admission control is not used in this case to manage the video bandwidth but instead to manage the audio traffic to ensure that the Priority Queue (PQ) is not over-subscribed. In this specific example the Voice pool in Enhanced Locations CAC admits the audio for both the voice-only calls and the video calls.

In Unified CM this feature is enabled by setting the service parameter **Deduct Audio Bandwidth from Audio Pool for Video Call** to **True** under the Call Admission Control section of the CallManager service. **False** is the default setting, and by default Unified CM deducts both audio and video streams of video calls from the video pool. This parameter changes that behavior and is key to the QoS alterations in the Preferred Architecture.

Figure 8-35 illustrates the various call flows, their corresponding audio and video streams, and queues to which queue they are directed.

Figure 8-35 Provisioning and Admission Control



The following conditions apply to the example in Figure 8-35:

- The Priority Queue is provisioned for all calls from endpoints and is protected by admission control (*E-LCAC voice BW pool)*.
- The Video queue is over-provisioned for room-based video systems:
  - Ratios are applied to desktop video endpoint usage.
  - Jabber video calls can use any bandwidth unused by video room systems.
  - During congestion, video streams of Jabber calls are subject to WRED drops and dynamically reduce video bit rate.

Region Configuration

Group video endpoints into classes of maximum video bit rate to limit bandwidth consumption based on endpoint type and usage within the solution. Three regions are required in total (see Table 8-14), and three device pools are required per site. This applies to a configuration where a single audio codec of G.722 is used across the entire organization, both LAN and WAN. Otherwise three regions per site are also required. See the considerations for regions in the Architecture section.

Table 8-14 Example Region Matrix for Three Groups

Endpoint Groupings	Video_1.5MB	Video_2.5MB	Video_20MB
Video_1.5MB	1,500 kbps	1,500 kbps	1,500 kbps
Video_2.5MB	1,500 kbps	2,500 kbps	2,500 kbps
Video_20MB	1,500 kbps	2,500 kbps	20,000 kbps

Deploy Enhanced Locations Call Admission Control

Limit video calling based only in areas of the network where bandwidth resources are restricted beyond AF41 marked traffic; otherwise, video bandwidth in the Location links should be unlimited.

- Enable LBM services on every node where the Cisco CallManager service is enabled.
- Configure locations.
  - On the locations and links management cluster, configure all locations and links in the organization.
  - On all other clusters (subordinate to a locations and links management cluster), configure only locations and remove any links to/from the locations.
- Add locations to each device pool. The devices that must be configured in a location either directly or via a device pool include:
  - IP phones (via device pool)
  - Conference bridges (via device pool)
  - Gateways (via device pool)
  - SIP trunks (via device pool)
  - Music on hold (MoH) servers (directly)
  - Annunciator (via device pool)

Intercluster Configuration

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

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

    An LBM is a spoke when it is not assigned to an LBM hub group.
  - Name: Cluster1_LBM_Hub_1
  - Bootstrap Servers: <*names or IP addresses of bootstrap servers*> (see the LBM Hub Replication Network section)
  - Select up to two LBMs in the cluster to serve as hubs.
- Recommendations for location configuration when intercluster ELCAC is implemented:
  - A cluster requires the location to be configured locally for location-to-device pool association.
  - Each cluster should have locations configured with the immediately neighboring locations of other clusters, so that each cluster's topology can inter-connect and create a single global topology. This does not apply to Location and Link Management Cluster deployments.
  - Discrepancies of bandwidth limits and weights on common locations and links are resolved by using the lowest bandwidth and weight values.
  - Naming locations consistently across clusters is critical. Follow the practice: "Same location, same name; different location, different name."
  - The Hub_none location should be renamed to be unique in each cluster. If Hub_none is left as default on all clusters, then it will be treated as the same location, which may or may not be desired, depending on the locations design being configured (see the Enhanced Locations Call Admission Control section).
  - Cluster-ID should be configured and must be unique on each cluster for serviceability reports to be usable.

Deploy Device Mobility for Mobile and Remote Access (MRA)

Configure Device Mobility

Figure 8-36 illustrates an overview of the device mobility configuration. Although this is a minimum configuration requirement for Device Mobility for ELCAC to function for Internet-based devices, Device Mobility can be configured to support mobility for these same endpoints within the enterprise. See the Cisco Collaboration SRND for more information on Device Mobility for devices within the enterprise.

Figure 8-36 Device Mobility Configuration and Location Association

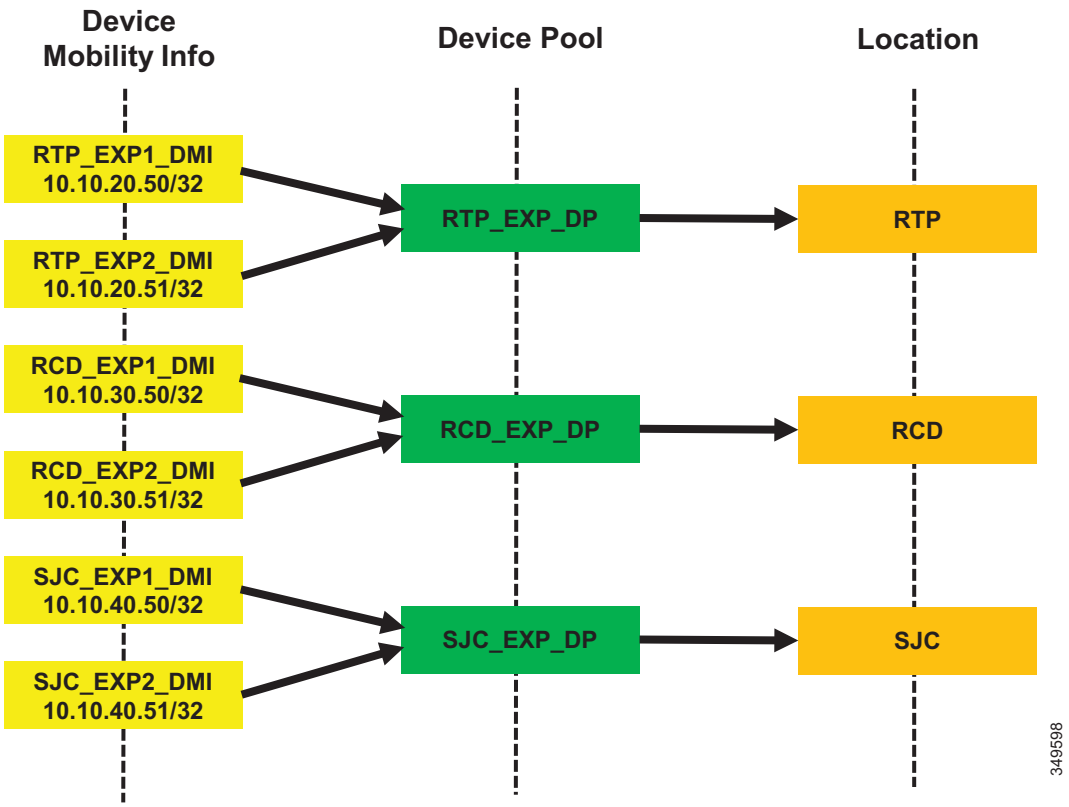
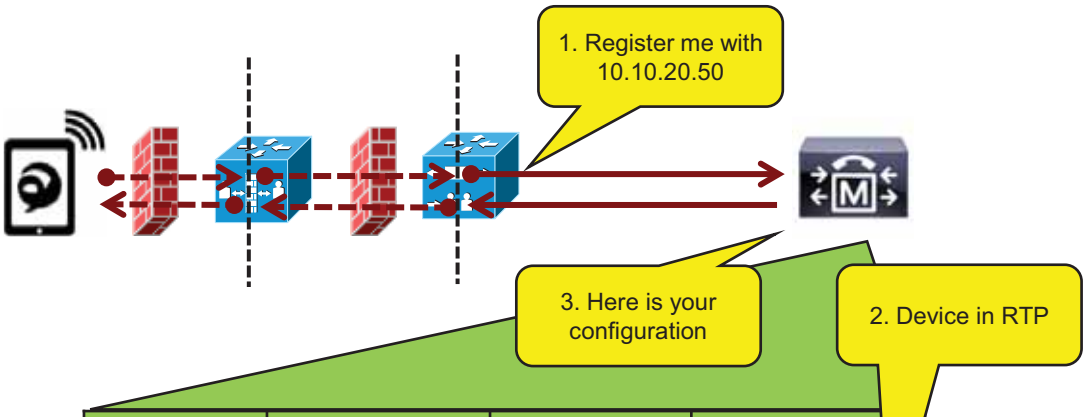


Figure 8-36 shows a simplified version of device mobility for the example deployment of ELCAC. The IP addresses of the Expressway-C servers are configured in the device mobility information. In this example there is a redundant pair of Expressway-C servers for each of the three sites: RTP, BLD, and SJC. RTP_EXP1_DMI and RTP_EXP2_DMI are configured respectively with the server IP addresses of the RTP Expressway-C servers. These two are associated to a new device pool called RTP_EXP_DP, which has the location RTP configured on it. Each site is configured similarly. With this configuration, when any device enabled for device mobility registers to Unified CM with the IP address that corresponds to the device mobility information in RTP_EXP1_DMI or RTP_EXP2_DMI, it will be associated with the RTP_EXP_DP device pool and thus with the RTP location.

With the above configuration, when an Internet-based device registers through the Expressway to Unified CM, it will register with the IP address of Expressway-C. Unified CM then uses the IP address configured in the device mobility information and associates the device pool and thus the Internet location associated to this device pool. This process is illustrated in Figure 8-37.

Figure 8-37 Association of Device Pool and Location Based on Expressway IP Address



IP Subnet	Device Mobility Info	Device Pool	Location
10.10.20.50	RTP_EXP1_DMI	RTP_EXP_DP	RTP
10.10.30.50	RCD_EXP1_DMI	RCD_EXP_DP	RCD
10.10.40.50	SJC_EXP1_DMI	SJC_EXP_DP	SJC

349599

In Figure 8-37. the client registers with Unified CM through the Expressway in RTP. Because the signaling is translated at the Expressway-C in RTP, the device registers with the IP address of that Expressway-C. The device pool RTP_EXP_DP is associated to the device based on this IP address. The RTP_EXP_DP pool is configured with the RTP location, and therefore that location is associated to the device. Thus, when devices register to the Expressway, they get the correct location association through device mobility. When the endpoint relocates to the enterprise, it will return to its static location configuration. Also, if the endpoint relocates to another Expressway in SJC, for example, it will get the correct location association through device mobility.

Configure Device Mobility Information (DMI) for Expressway-Cs:

- Create two DMIs per Expressway-C group (two Expressway-C nodes in a pair).
- Add the IP address of the Expressway-C node in a subnet with a mask of 32 bits (this matches the IP address exactly).
- Add the site device pool to respective DMIs. This is the device pool of the site where the Expressway pairs are located, which should contain the correct region and location.

  Example for one DMI:

  Name: SJC_EXP1_DMI

  Subnet: 10.10.40.50

  Mask: 32

  Selected Device Pool: SJC_Video_1.5MB

Enable devices for device mobility. The bulk administration tool (BAT) can and should be used to facilitate this step.

Bandwidth Allocation Guidelines

The bandwidth allocations in Figure 8-38 are unique guidelines based on this example enterprise. They provide some guidance on percentages of available bandwidth for different common classes of Collaboration traffic.

Figure 8-38 Bandwidth Allocation Guidelines

WAN Link Speed / Class	622 Mbps (OC12)	155 Mbps (OC3)	34-44 Mbps (E3/DS3)	10 Mbps	5 Mbps	<2 Mbps (T1/E1)
Control (%)	1	1	1	1	2	10
Voice (%)	10	10	10	20	25	30
Video (%)	43	53	55	40	30	--
Signalling (%)	2	2	2	2	2	5
Scavenger (%)	1	1	1	1	1	1
Default (%)	43	33	31	36	40	54

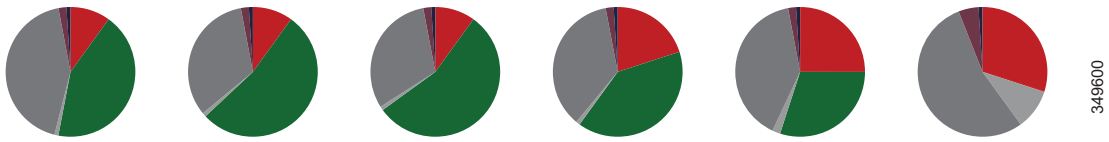
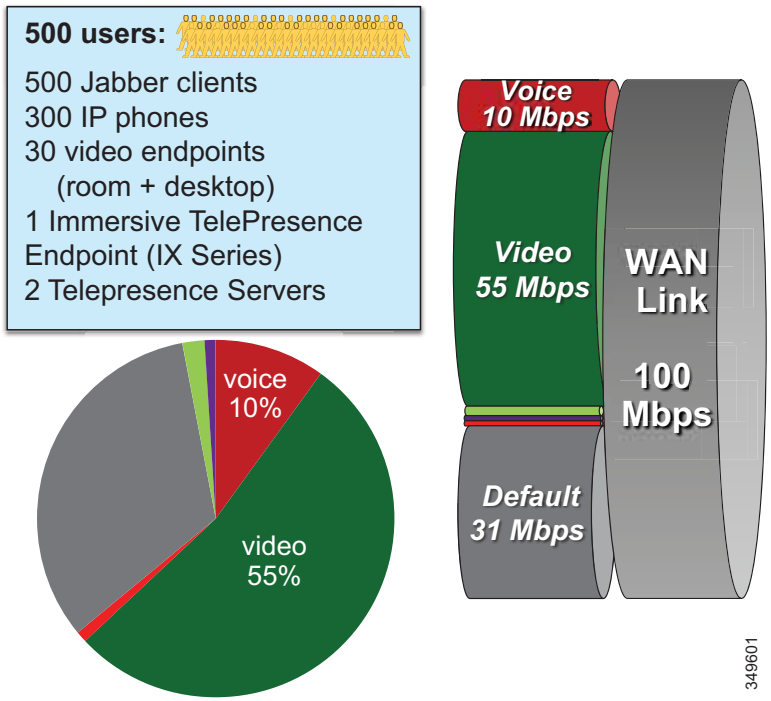


Figure 8-39 through Figure 8-42 illustrate each site (Central, Large Branch, Small Branch, Micro Branch) and the link bandwidth provisioned for each class based on the number of users and available bandwidth for each class. Keep in mind that these values are based on bandwidth calculated for Layer 3 and above. Therefore, the values do not include the Layer 2 overhead, which is dependent on the link type (Ethernet, Frame-relay, MPLS, and so forth). See the Network Infrastructure chapter of the Collaboration SRND for more information on L2 overhead. Also note that the audio portion of bandwidth for the video calls is deducted from the voice pool, so the voice queue is provisioned to include the audio bandwidth of both voice-only and video calls.

Note The calculations in the following examples use the maximum bandwidth for the number of endpoints and then multiply that value by a percentage to account for active calls. For example, for 30 video endpoints (30 calls possible) at 20% active video call rate, the calculation would be:
1.2 Mbps * 30 calls * 0.2 = 7.2 Mbps.

Figure 8-39 Central Site



Central Site Link (100 Mbps) Bandwidth Calculation

As illustrated in Figure 8-39, the Central Site has the following bandwidth requirements:

- Voice queue (PQ): 10 Mbps (L3 bandwidth)

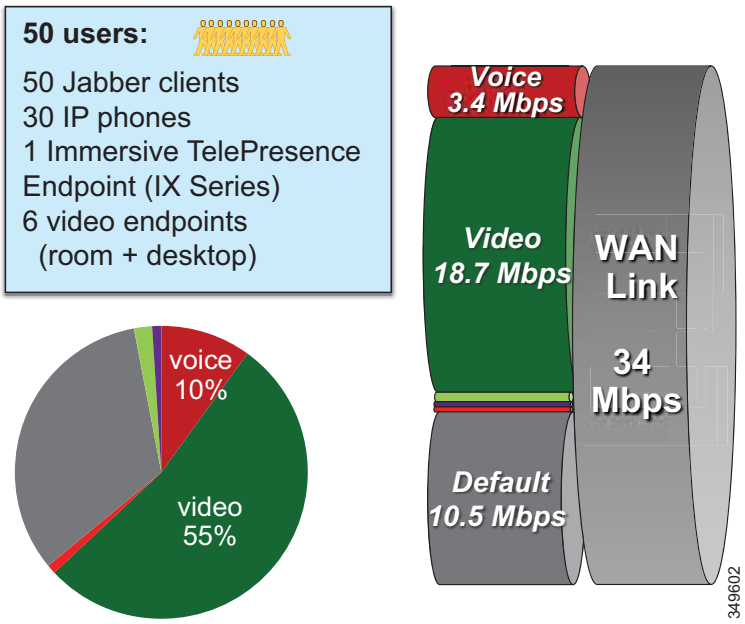
  125 calls @ G.711 or G.722
- Unified CM Location link bandwidth for the voice pool:

  125 ∗ 80 kbps = 10 Mbps
- Video queue: 55 Mbps (L3 bandwidth)
  - Immersive video endpoint (IX Series): 3 Mbps ∗ 1 call = 3 Mbps
  - Video endpoints: 1.2 Mbps ∗ 30 calls ∗ 0.2 = 7.2 Mbps
  - TelePresence Servers: 1.5 Mbps ∗ 40 calls ∗ 0.5 = 30 Mbps
  - 55 Mbps – (3 Mbps + 7.2 Mbps + 30 Mbps) = 14.8 Mbps for Jabber media

    11 Jabber video calls @ 720p, or 18 @ 576p, or 50 @ 288p

    (Plus any leftover bandwidth)

Figure 8-40 Large Branch



Large Branch Link (34 Mbps) Bandwidth Calculation

As illustrated in Figure 8-40, the Large Branch site has the following bandwidth requirements:

- Voice queue (PQ): 3.4 Mbps (L3 bandwidth)

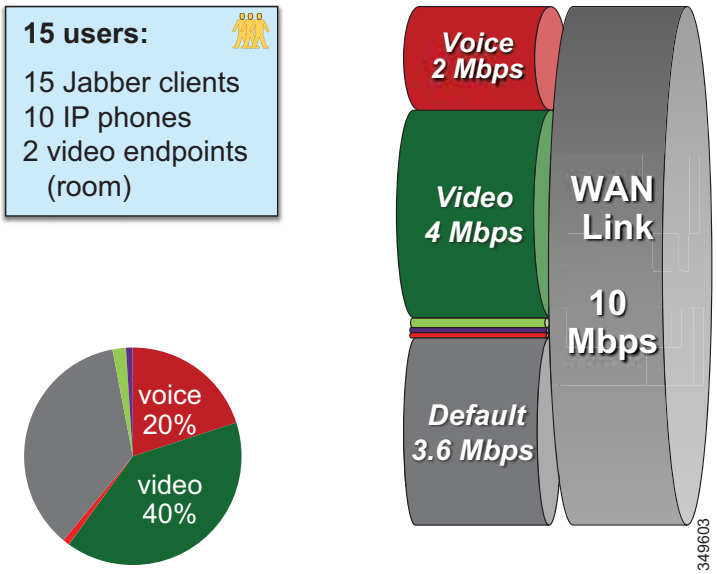
  42 calls @ G.711 or G.722
- Unified CM Location link bandwidth for the voice pool:

  42 ∗ 80 kbps = 3.360 Mbps
- Video queue: 18.7 Mbps (L3 bandwidth)
  - Immersive video endpoint (IX Series): 3 Mbps ∗ 1 call = 3 Mbps
  - Video endpoints: 1.2 Mbps ∗ 6 calls = 7.2 Mbps
  - 18.7 Mbps – (3 Mbps + 7.2 Mbps) = 8.5 Mbps for Jabber media

    6 Jabber video calls @ 720p, or 10 @ 576p, or 36 @ 288p

    (Plus any leftover bandwidth)

Figure 8-41 Small Branch



Small Branch Link (10 Mbps) Bandwidth Calculation

As illustrated in Figure 8-41, the Small Branch site has the following bandwidth requirements:

- Voice queue (PQ): 2 Mbps (L3 bandwidth)

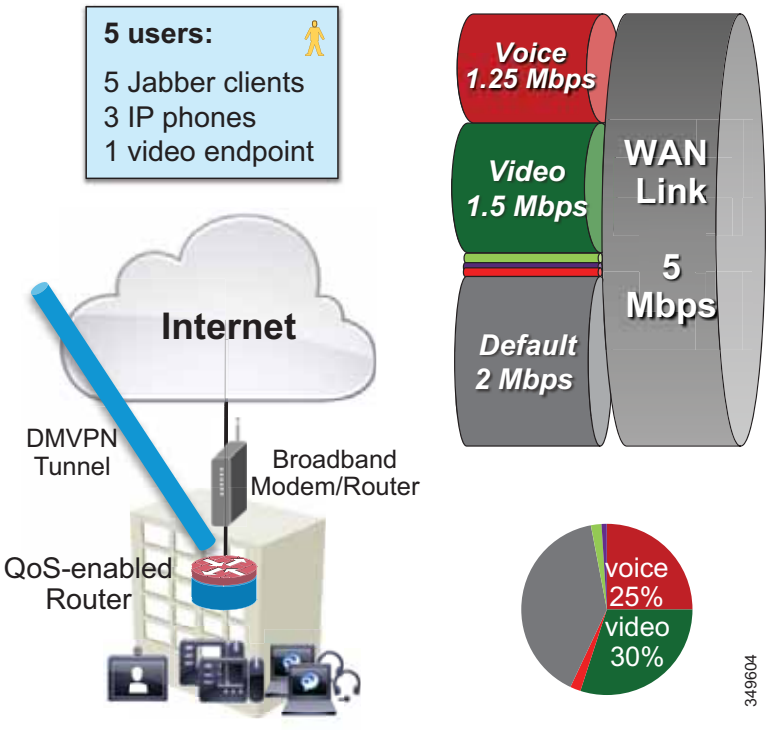
  25 calls @ G.711 or G.722
- Unified CM Location link bandwidth for the voice pool:

  25 ∗ 80 kbps = 2 Mbps
- Video queue: 4 Mbps (L3 bandwidth)
  - Video endpoints: 1.2 Mbps * 2 calls = 2.4 Mbps
  - 4 Mbps – 2.4 Mbps = 1.6 Mbps for Jabber media

    1 Jabber video call @ 720p, or 2 @ 576p, or 5 @ 288p

    (Plus any leftover bandwidth)

Figure 8-42 Micro Branch



Micro Branch Broadband Internet Connectivity (5 Mbps) Bandwidth Calculation

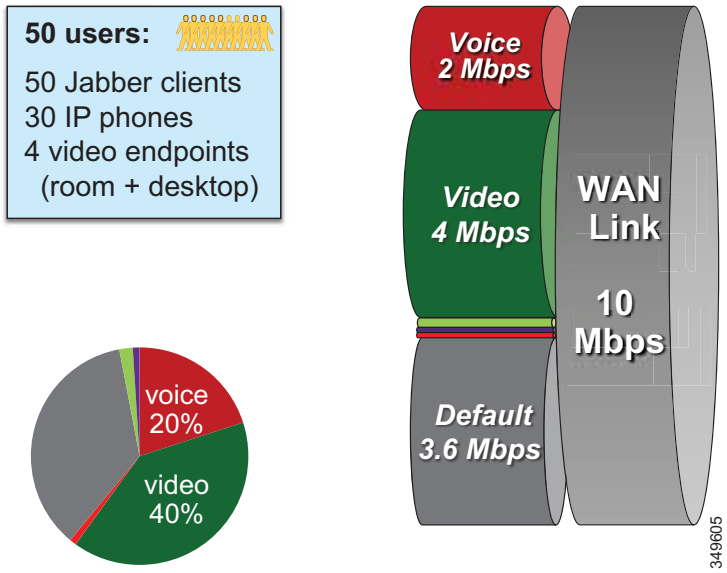
As illustrated in Figure 8-42, the Micro Branch site has the following bandwidth requirements:

- Broadband Internet connectivity + DMVPN to central site
- Configure interface of VPN router to match broadband up-link speed
- Enable QoS on VPN router to prevent bufferbloat from TCP flows
- Asymmetric download/upload broadband: consider limiting transmit bit rate on video endpoint
- Bandwidth calculation will ultimately depend on the broadband bandwidth available and should follow the same recommendations as in the Small Branch site link for provisioning.

Large Branch with Constrained WAN Link (Enhanced Locations CAC Enabled for Video)

In specific branch sites with lower-speed WAN links, over-provisioning the video queue is not feasible. ELCAC can be applied to these location links for video to ensure that video calls do not over-subscribe the link bandwidth. This template requires using site-specific region configuration to limit maximum bandwidth used by video endpoints and Jabber clients. Also keep in mind that device mobility is required if Jabber users roam across sites.

Figure 8-43 Large Branch with Constrained WAN Link (Enhanced Locations CAC Enabled for Video)



As illustrated in Figure 8-43, a Large Branch site with a constrained WAN link (10 Mbps) has the following bandwidth requirements:

- Voice queue (PQ): 2 Mbps (L3 bandwidth)

  25 calls @ G.711 or G.722

- Unified CM Location link bandwidth for the voice pool:

  25 ∗ 80 kbps = 2 Mbps

- Video queue: 4 Mbps (L3 bandwidth)
  - Possible usage:

    1 call @ 720p (1,220 kbps) + 3 calls @ 576p (810 kbps) = 3,650 kbps

    Or 2 calls @ 576p (768 kbps) + 5 calls @ 288p (320 kbps) = 3136 kbps
  - Unified CM Location link bandwidth for video calls: 3.7 Mbps (L3 bandwidth)
  - Leaves room for L2 overhead