



## Quality of Service

---

This chapter addresses the Quality of Service (QoS) requirements for implementations of IP telephone solutions over an enterprise network. By applying the prerequisite tools, you can achieve excellent quality voice, video, and data transmissions over an IP infrastructure, irrespective of media and even at low data rates. For more detailed information on designing Quality of Service networks for AVVID deployments, please see the *Cisco AVVID QoS Design Guide* at

[http://www.cisco.com/univercd/cc/td/doc/product/voice/ip\\_tele/index.htm](http://www.cisco.com/univercd/cc/td/doc/product/voice/ip_tele/index.htm)

This chapter includes the following major sections:

- Campus QoS Model, page 8-1
- WAN QoS Model, page 8-4

### Campus QoS Model

Until recently, conventional wisdom stated that Quality of Service would never be an issue in the enterprise campus due to the bursty nature of data traffic and the capability to withstand buffer overflow and packet loss. When applications such as voice and video, which are sensitive to loss and delay, began to traverse the data network, network designers gradually came to understand that buffers and not bandwidth are the issue in the campus. Buffers can fill instantaneously. When this occurs, packets can be dropped when attempting to enter the interface buffer. For applications like voice, which are extremely drop intolerant, this results in voice quality degradation. QoS tools are required to manage these buffers to minimize loss, delay, and delay variation.

Campus QoS really involves two separate areas of configuration, which are discussed in the following sections:

- Traffic Classification
- Interface Queuing

## Traffic Classification

Classifying or marking traffic as close to the edge of the network as possible has always been an integral part of the Cisco network design architecture. Traffic classification is an entrance criterion for access into the various queuing schemes used within the campus switches and WAN interfaces. When connecting an IP phone using a single cable model, the phone becomes the edge of the managed network. As such, the IP phone can and should classify traffic flows. Table 8-1 lists the AVVID traffic classification guidelines.

**Table 8-1 Traffic Classification Guidelines for AVVID Networks**

Traffic Type	Layer 2 Class of Service (CoS)	Layer 3 IP Precedence	Layer 3 DSCP
Voice RTP	5	5	EF
Voice Control	3	3	AF31
Video	4	4	AF41
Data	0-2	0-2	0-AF23

## Interface Queuing

To guarantee voice quality, it is a design requirement to enable QoS within the campus infrastructure. By enabling QoS on campus switches, you can configure all voice traffic to use separate queues, thus virtually eliminating the possibility of dropped voice packets when an interface buffer fills instantaneously.

Although network management tools may show that the campus network is not congested, QoS tools are still required to guarantee voice quality. Today's network management tools show only the average congestion over a sample time span. While useful, this average does not show the congestion peaks on a campus interface. Transmit interface buffers within a campus tend to congest absolutely

in small, finite intervals as a result of the bursty nature of network traffic. When this occurs, any packets destined for that transmit interface are dropped. The only way to prevent dropped voice traffic is to configure multiple queues on campus switches. Table 8-2 lists the Cisco Ethernet switches that support enhanced queuing services.

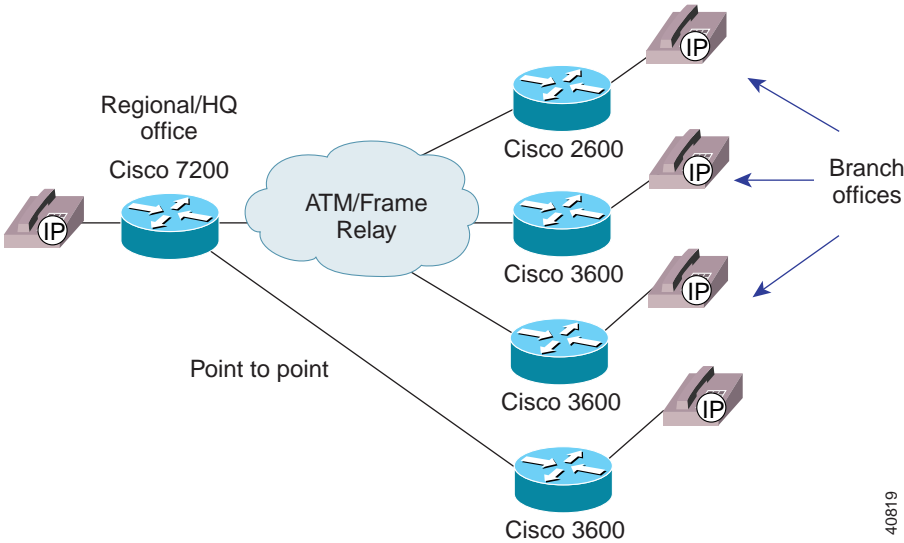
**Table 8-2 Queuing Services Supported by Cisco Switches**

Campus Switching Element	Queuing Scheme	Queue Scheduler	Queue Admission
Catalyst 6000	2Q2T and 1P2Q2T	WRR and PQ/WRR	Configurable
Catalyst 8500	4Q1T	WRR	Configurable in CoS pairs
Catalyst 4000	2Q1T	RR	Configurable in CoS pairs
Catalyst 3500	2Q1T	PQ	Not configurable. CoS 0-3 = Queue1 CoS 0-3 = Queue2
Catalyst 2900 XL (8 MB DRAM)	2Q1T	PQ	Not configurable. CoS 0-3 = Queue1 CoS 0-3 = Queue2
IP Phone	1P3Q1T	RR with a PQ timer	Not configurable. CoS 5 = Queue0 (PQ). All other CoS values = Queues 1-3.

# WAN QoS Model

The enterprise WAN model is shown in Figure 8-1.

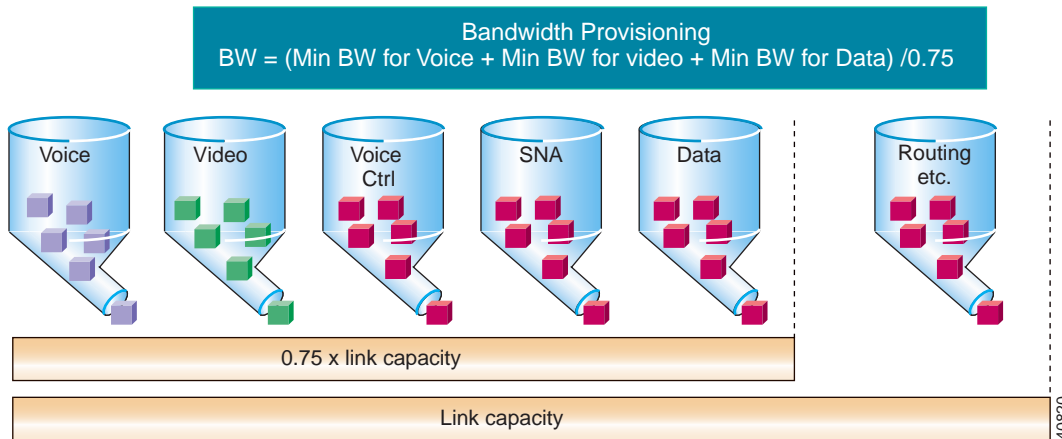
Figure 8-1 Typical Enterprise WAN



## WAN Provisioning

Before voice and video can be placed on a network, it is necessary to ensure that adequate bandwidth exists for all required applications. To begin, the minimum bandwidth requirements for each major application (for example, the voice media streams, video streams, voice control protocols, and all data traffic) should be summed. This sum represents the minimum bandwidth requirement for any given link, and it should consume no more than 75% of the total bandwidth available on that link. This 75% rule assumes that some bandwidth is required for overhead traffic such as routing and Layer 2 keepalives, as well as for additional applications such as e-mail, HTTP traffic, and other data traffic that is not so easily measured. See Figure 8-2.

Figure 8-2 Provisioning a Converged Network



## WAN QoS Tools

This section discusses the tools used to implement QoS for IP telephony applications over the enterprise WAN. These tools include traffic prioritization, link fragmentation and interleaving (LFI), and traffic shaping. This section concludes with a summary of best practices for each of the applicable data link protocols.

### Traffic Prioritization

In choosing from among the many available prioritization schemes, the major factors to consider include the type of traffic being put on the network and the wide area media to be traversed. For multiservice traffic over an IP WAN, Cisco recommends low-latency queuing for low-speed links. This allows up to 64 traffic classes with the ability to specify, for example, priority queuing behavior for voice and interactive video, a minimum bandwidth for Systems Network Architecture (SNA) data and market data feeds, and weighted fair queuing to other traffic types.

Figure 8-3 shows this prioritization scheme as follows:

- Voice is placed into a queue with priority queuing capabilities and is allocated a bandwidth of 48 kbps. The entrance criterion to this queue should be the differentiated services code point (DSCP) value of EF, or IP precedence value of 5. Traffic in excess of 48 kbps would be dropped if the interface becomes congested. Therefore, an admission control mechanism must be used to ensure that this value is not exceeded.
- Video conferencing traffic is placed into a queue with priority queuing capabilities and is allocated a bandwidth of 384 kbps. The entrance criterion to this queue should be a DSCP value of AF41, or IP precedence value of 4. Traffic in excess of 384 kbps would be dropped if the interface becomes congested. It is therefore imperative, as in the case of voice, to use an admission control mechanism to ensure that this value is not exceeded.



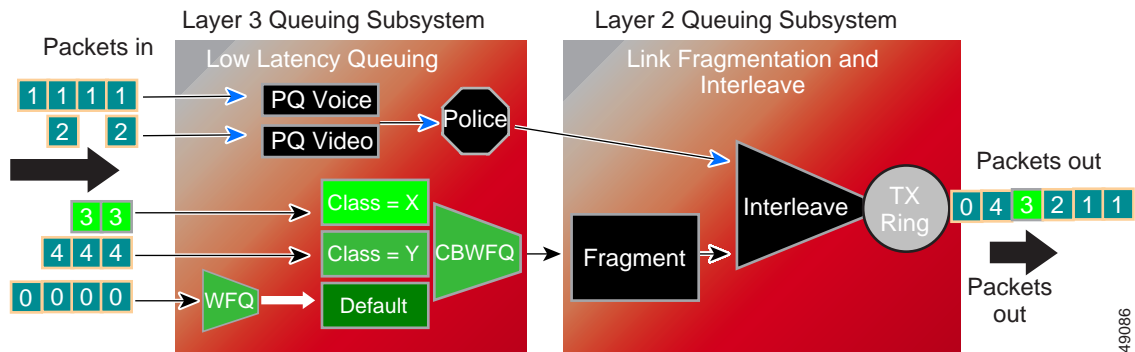
---

**Note** One-way video traffic, such as IP/TV, should use a class-based weighted fair queuing scheme because the delay tolerances are much higher.

---

- As the WAN links become congested, it is possible to completely starve the voice control signaling protocols, thereby eliminating the ability of the IP phones to complete calls across the IP WAN. Voice control protocol traffic, such as H.323 and the Skinny Client Control Protocol, requires its own class-based weighted fair queue with a minimum configurable bandwidth equal to a DSCP value of AF31, which correlates to an IP precedence value of 3.
- SNA traffic is placed into a queue that has a specified bandwidth of 56 kbps. Queuing operation within this class is first-in-first-out (FIFO) with a minimum allocated bandwidth of 56 kbps. Traffic in this class that exceeds 56 kbps is placed in the default queue. The entrance criterion to this queue could be TCP port numbers, Layer 3 address, IP precedence, or a DSCP.
- All remaining traffic can be placed in a default queue. If a bandwidth is specified, the queuing operation would be FIFO. Alternatively, by specifying the keyword **fair**, the operation would be weighted fair queuing (WFQ).

Figure 8-3 Optimized Queuing for VoIP over the WAN



The following points must be taken into account when configuring low-latency queuing (LLQ):

- The minimum system software for leased lines and Asynchronous Transfer Mode (ATM) is Cisco IOS Release 12.1(2)T.
- The minimum system software for Frame Relay is Cisco IOS Release 12.1(2)T.

Table 8-3 gives the minimum bandwidth requirements for voice, video, and data networks using Cisco CallManager Release 3.0(5). Note that these values are *minimum*, and any network should be engineered with adequate capacity.

Table 8-3 Minimum Bandwidth Requirements with Cisco CallManager 3.0(5)

Traffic Type	Leased Lines	Frame Relay	ATM	ATM/Frame Relay
Voice + data	64 kbps	64 kbps	128 kbps	128 kbps
Voice, video, and data	768 kbps	768 kbps	768 kbps	768 kbps

## Link Efficiency Techniques

Because wide-area bandwidth is often prohibitively expensive, only low-speed circuits may be available or cost effective when interconnecting remote sites. In these cases, it is important to achieve the maximum savings by transmitting as many voice calls as possible over the low-speed link. Many compression schemes,

such as G.729, can squeeze a 64-kbps call down to an 8-kbps payload. Cisco gateways and IP phones support a range of codecs that can enhance efficiency on these low-speed links.

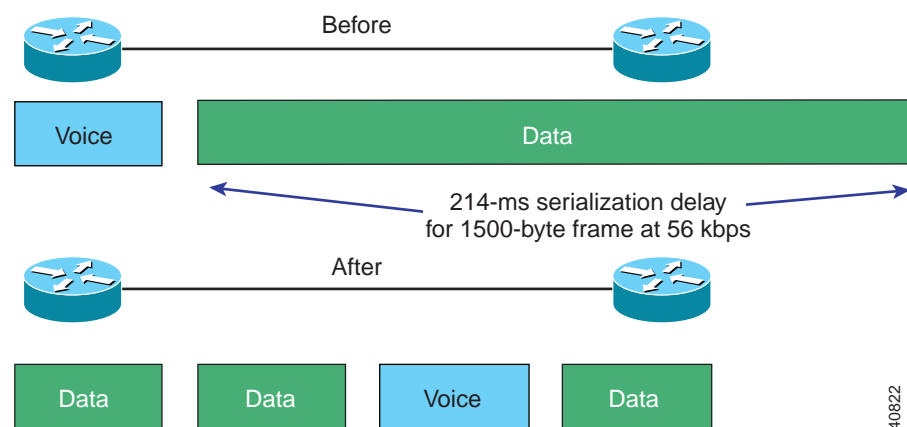
The link efficiency can be further increased by using compressed RTP (cRTP), which compresses a 40-byte IP + UDP + RTP header to approximately two to four bytes. In addition, voice activity detection (VAD) takes advantage of the fact that, in most conversations, only a single party is talking at a time. VAD recovers this empty time and allows data to use the bandwidth.

**Note**

cRTP is currently supported only for leased lines and Frame Relay media. Cisco IOS Release 12.1(2)T, which greatly enhances performance, is the recommended system software for cRTP.

For low-speed links (less than 768 kbps), it is necessary to use techniques that provide link fragmentation and interleaving (LFI). This places bounds on jitter by preventing voice traffic from being delayed behind large data frames. The three techniques that exist for this purpose are Multilink PPP (MLP) for point-to-point serial links, FRF.12 for Frame Relay, and MLP over ATM for ATM connections (available in Cisco IOS Release 12.1(5)T). Figure 8-4 depicts the general operation of LFI.

**Figure 8-4 Link Fragmentation and Interleaving (LFI) Operation**

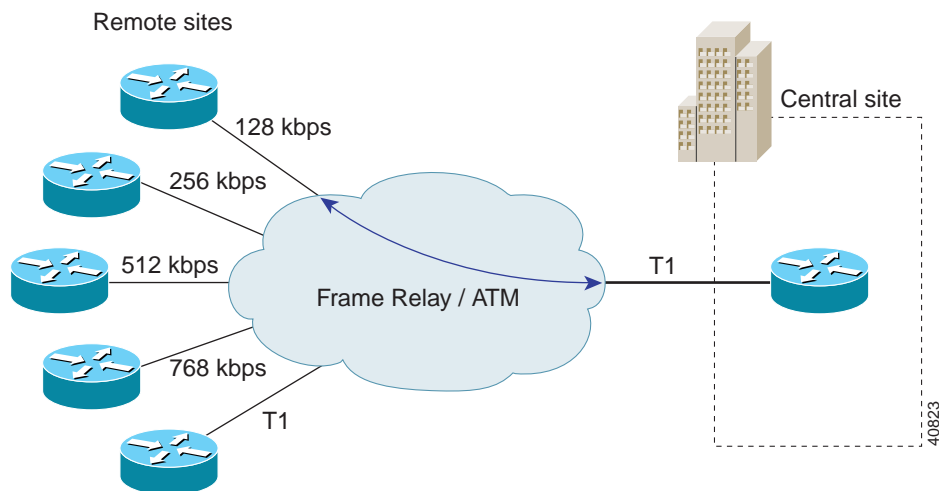


40822

## Traffic Shaping

Traffic shaping is required for multiple access, non-broadcast media such as ATM and Frame Relay, where the physical access speed varies between two endpoints. Traffic shaping technology accommodates mismatched access speeds. In the case of Frame Relay with FRF.12, traffic shaping also allows delay variation, or jitter, to be bounded appropriately. For ATM, data rates are such that fragmentation is typically not required. Figure 8-5 demonstrates traffic shaping with Frame Relay and ATM.

**Figure 8-5** Traffic Shaping with Frame Relay and ATM



## Best Practices

Table 8-4 shows the minimum recommended software release for enterprise voice implemented over the WAN and includes recommended parameters for QoS tools. The currently recommended Cisco IOS versions will change with future releases.

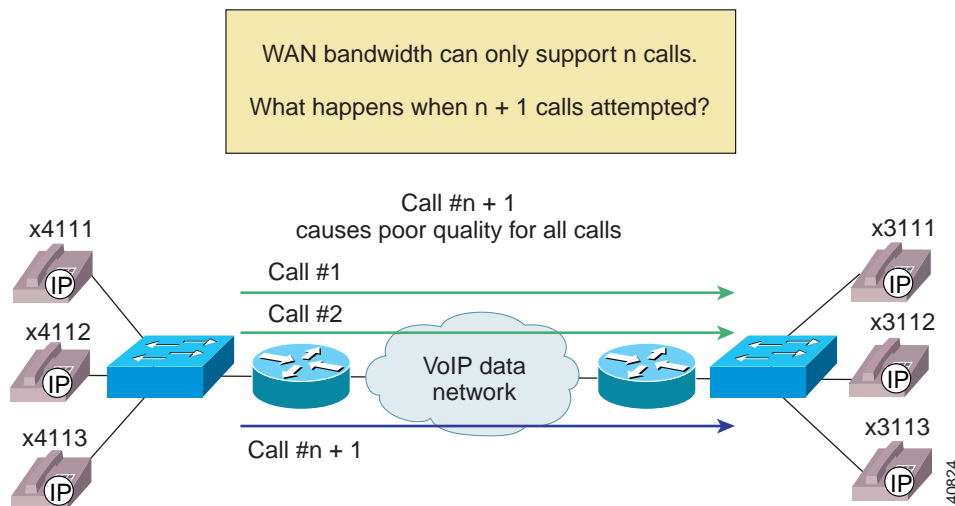
**Table 8-4 Recommended Cisco IOS and QoS Tools**

Data Link Type	Minimum Cisco IOS Release	Classification	Prioritization	LFI	Traffic Shaping
Serial Lines	12.0(7)T	DSCP = EF for voice; DSCP = AF41 for video; DSCP = AF31 for voice control traffic; other classes of traffic have a unique classification.	LLQ with CBWFQ	MLP	N/A
Frame Relay	12.1(2)T	DSCP = EF for voice; DSCP = AF41 for video; DSCP = AF31 for voice control traffic; other classes of traffic have a unique classification.	LLQ with CBWFQ	FRF.12	Shape traffic to committed information rate (CIR).
ATM	12.1(5)T	DSCP = EF for voice; DSCP = AF41 for video; DSCP = AF31 for voice control traffic; other classes of traffic have a unique classification.	LLQ with CBWFQ	MLP over ATM	Shape traffic to guaranteed portion of bandwidth.
ATM and Frame Relay	12.1(5)T	DSCP = EF for voice; DSCP = AF41 for video; DSCP = AF31 for voice control traffic; other classes of traffic have a unique classification.	LLQ with CBWFQ	MLP over ATM and Frame Relay	Shape traffic to guaranteed portion of bandwidth on slowest link.

## Call Admission Control

Call admission control is required to ensure that network resources are not oversubscribed. Calls that exceed the specified bandwidth are either rerouted using an alternative route such as the PSTN, or a busy tone is returned to the calling party. Figure 8-6 demonstrates that call admission control is needed regardless of whether the implementation model is toll bypass or IP telephony to the desktop.

**Figure 8-6 Call Admission Control Required to Protect WAN Bandwidth**



There are two schemes for providing call admission control for voice calls over the WAN:

- Gatekeeper call admission control—see the “Call Admission Control” section on page 6-3
- Locations call admission control—see the “Call Admission Control” section on page 7-3.

