



## CHAPTER 4

# Quality of Service Design for TelePresence

---

## Overview

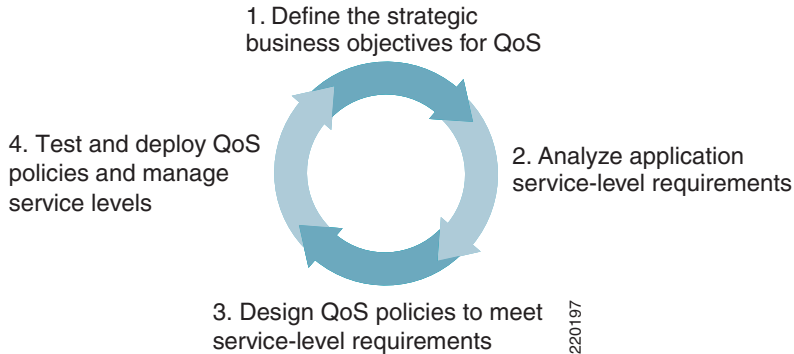
A major benefit of Cisco's TelePresence solution over competitive offerings is that the realtime, high-definition video and audio are transported over a converged IP network rather than a dedicated network (although dedicated networks are also supported). The key enabling technology to accomplish this convergence is Quality of Service (QoS).

QoS technologies refer to the set of tools and techniques to manage network resources, such as bandwidth, latency, jitter, and loss. QoS technologies allow different types of traffic to intelligently contend for network resources. For example, voice and realtime video—such as TelePresence—may be granted strict priority service, while some critical data applications may receive (non-priority) preferential services and some undesired applications may be assigned deferential levels of service. Therefore, QoS is a critical, intrinsic element for the successful network convergence of voice, video, and data.

There are four principal phases to a successful QoS deployment:

- Clearly define the strategic business objectives of the QoS deployment.
- Analyze application service-level requirements.
- Design (and test) QoS policies to accommodate service level requirements.
- Roll out the QoS policies and monitor service levels.

These phases are sequential and the success of each subsequent phase directly depends on how well the previous phase has been addressed. Furthermore, the entire process is generally cyclical, as business applications and objectives evolve over time and their related QoS policies periodically need to be adjusted to accommodate (see [Figure 4-1](#)).

**Figure 4-1 The Four Phases of Successful QoS Deployments**

The following sections examine how each of these phases relate to a successful deployment of QoS for TelePresence.

## Defining the Strategic Business Objective for QoS for TelePresence

QoS technologies are the enablers for business/organizational objectives. Therefore, the way to begin a QoS deployment is not to activate QoS features simply because they exist, but to start by clearly defining the QoS-related business objectives of the organization.

For example, among the first questions that arise during a QoS deployment are: How many traffic classes should be provisioned for? And what should they be? To help answer these fundamental questions, QoS-related organizational objectives need to be defined, such as:

- Is the business objective to enable TelePresence only? Or is VoIP also required to run over the converged network?
- Are there any non-realtime applications that are considered critical to the core business objectives? If so, what are they?
- Are there applications which should be squelched (i.e., deferential treatment)? If so, what are they?

The answers to these questions define the applications that require QoS policies, either preferential QoS or deferential QoS. Each application that has a unique service level requirement—whether preferential or deferential—requires a dedicated service class to deliver and guarantee the requisite service levels.

Additionally, Cisco offers a non-technical recommendation for this first phase of a successful QoS deployment, namely to always seek executive endorsement of the QoS business objectives prior to design and deployment. This is because QoS is a system of managed application preference and as such often includes political and organizational repercussions when implemented. To minimize the effects of these non-technical obstacles to deployment, it is recommended to address these political and organizational issues as early as possible, garnishing executive endorsement whenever possible.

# Analyzing the Service Level Requirements of TelePresence

Once the applications requiring QoS have been defined by the organization business objectives, then the network administrators must carefully analyze the specifics of the service levels required by each application to be able to define the QoS policies to meet them. The service level requirements of realtime applications, such as TelePresence, are defined by the following four parameters:

- Bandwidth
- Latency (delay)
- Jitter (variations in delay)
- Packet loss

## TelePresence Bandwidth Requirements

Cisco TelePresence systems are currently available in one screen (CTS-1000) and three screen (CTS-3000) configurations. A CTS-3000 obviously has greater bandwidth requirements than a CTS-1000, but not necessarily by a full-factor of three, as will be shown. Furthermore, the resolution of each CTS-1000 or CTS-3000 system can be set to 720p or 1080p (full HDTV); the resolution setting also significantly impacts the bandwidth requirements of the deployed TelePresence solution.

As discussed in [Chapter 1, “Cisco TelePresence Solution Overview,”](#) Cisco TelePresence has even more levels of granularity in overall image quality within a given resolution setting, as the motion handling quality can also be selected. Therefore, TelePresence supports three levels of motion handling quality within a given resolution, specifically 720p-Good, 720p-Better, and 720p-Best, as well as 1080p-Good, 1080p-Better, and 1080p-Best. Each of these levels of resolution and motion handling quality results in slightly different bandwidth requirements, as detailed in [Table 4-1](#).

To keep the following sections and examples simple to understand, only two cases will be broken down for detailed analysis: 720p-Good and 1080p-Best.

Let’s break down the bandwidth requirements of the maximum bandwidth required by a CTS-1000 system running at 720p-Good, with an auxiliary video stream (a 5 frame-per-second video channel for sharing Microsoft PowerPoint or other collateral via the data-projector) and an auxiliary audio stream (for at least one additional person conferenced in by an audio-only bridge). The bandwidth requirements by component are:

1 primary video streams @ 1 Mbps:	1,000 Mbps (1 Mbps)
1 primary audio streams @ 64 Kbps:	64 Kbps
1 auxiliary video stream (5 fps):	500 Kbps
1 auxiliary audio stream:	<u>64 Kbps</u>
Total audio and video bandwidth (not including burst and network overhead):	1,628 Kbps (1.628 Mbps)

The total bandwidth requirements—without network overhead—of such a scenario would be 1.628 Mbps. However a 10% burst factor on the video channel, along with the IP/UDP/RTP overhead (which combined amounts to 40 bytes per packet) must also be taken into account and provisioned for, as must media-specific Layer 2 overhead. In general, video—unlike voice—does not have clean formulas for calculating network overhead because video packet sizes and rates vary proportionally to the degree of

motion within the video image itself. From a network administrator's point of view, bandwidth is always provisioned at Layer 2, but the variability in the packet sizes and the variety of Layer 2 mediums the packets may traverse from end-to-end make it difficult to calculate the real bandwidth that should be provisioned at Layer 2. Cisco TelePresence video packets average 1,100 bytes per packet. However, the conservative rule of thumb that has been thoroughly tested and widely deployed is to overprovision video bandwidth by 20%. This accommodates the 10% burst and the Layer 2-Layer 4 network overhead.

With this 20% overprovisioning rule applied, the requisite bandwidth for a CTS-1000 running at 720p-Good becomes 2 Mbps (rounded).

Now, let's break down the maximum bandwidth required by a CTS-3000 system running at full 1080p-Best, with an auxiliary video stream and an auxiliary audio stream. The detailed bandwidth requirements are:

3 primary video streams @ 4 Mbps each:	12,000 Kbps (12 Mbps)
3 primary audio streams @ 64 Kbps each:	192 Kbps
1 auxiliary video stream:	500 Kbps
1 auxiliary audio stream:	<u>64 Kbps</u>
Total audio and video bandwidth (not including burst and network overhead):	12,756 Kbps (12.756 Mbps)

With the 20% overprovisioning rule applied, the requisite bandwidth for a CTS-3000 running at 1080p-Best becomes approximately 15 Mbps (with a bit of rounding applied). This value of 15 Mbps for a CTS-3000 at 1080p-Best is used in most of the examples throughout this design guide.

It is important to note that as the Cisco TelePresence software continues to evolve and add new feature support, the bandwidth requirements for TelePresence will correspondingly evolve and expand. For example, [Table 4-1](#) shows the bandwidth requirements for CTS software version 1.2, with and without network overhead, of CTS-1000 and CTS-3000 systems running at 720p and 1080p with all grades of motion handling quality (Good, Better, and Best).

**Table 4-1 Cisco TelePresence Software Version 1.2 Bandwidth Requirements**

Resolution	1080p	1080p	1080p	720p	720p	720p
Motion Handling	Best	Better	Good	Best	Better	Good
Video per Screen (kbps)	4000	3500	3000	2250	1500	1000
Audio per Microphone (kbps)	64	64	64	64	64	64
(5 fps) Auto Collaborate Video Channel (kbps)	500	500	500	500	500	500
Auto Collaborate Audio Channel (kbps)	64	64	64	64	64	64
CTS-500/1000 Total Audio and Video (kbps)	4,628 <sup>1</sup>	4,128 <sup>1</sup>	3,628 <sup>1</sup>	2,878 <sup>1</sup>	2,128 <sup>1</sup>	1,628 <sup>1</sup>
CTS-3000/3200 Total Audio and Video (kbps)	12,756	11,256	9,756	7,506	5,256	3,756
CTS-500/1000 total bandwidth (Including Layer 2-Layer 4 overhead)	5.5 Mbps <sup>1</sup>	5.0 Mbps <sup>1</sup>	4.3 Mbps <sup>1</sup>	3.4 Mbps <sup>1</sup>	2.5 Mbps <sup>1</sup>	2 Mbps <sup>1</sup>
CTS-3000/3200 total bandwidth (Including Layer 2-Layer 4 overhead)	15.3 Mbps	13.5 Mbps	11.7 Mbps	9.0 Mbps	6.3 Mbps	4.5 Mbps

1. The CTS-1000 transmits up to 128kbps of audio, but can receive up to 256kbps when participating in a meeting with a CTS-3000.

**Note**

These bandwidth numbers represent the worst-case scenarios (i.e., peak bandwidth transmitted during periods of maximum motion within the encoded video). Normal use (i.e., average bandwidth), with users sitting and talking and gesturing naturally, typically generates only about 60-80% of these maximum bandwidth rates. This means that a CTS-3000 running at 1080-Best averages only 10-12 Mbps and a CTS-1000 running at 720-Good averages only 1.2-1.6 Mbps.

Release 1.3 of CTS software introduced support for an interoperability feature, which allows for TelePresence systems to interoperate with H.323-based video-conferencing systems. From a bandwidth perspective, the only change to existing TelePresence flows is that there is an additional video channel transmitted (768 Kbps) and an additional audio channel transmitted (64 Kbps) by the TelePresence endpoints (these values are exclusive of network overhead). These additions have been highlighted in bold in [Table 4-2](#). When the 20% network-overhead overprovisioning rule is applied, the additional bandwidth required to support this interoperability feature becomes about 1 Mbps for any TelePresence system, regardless of the number of segments, resolution-levels, and motion-handling capabilities configured for TelePresence primary video.

**Table 4-2 Cisco TelePresence Software Version 1.3 Bandwidth Requirements (Including the Interoperability Feature)**

Resolution	1080p	1080p	1080p	720p	720p	720p
Motion Handling	Best	Better	Good	Best	Better	Good
Video per Screen (kbps)	4000	3500	3000	2250	1500	1000
Audio per Microphone (kbps)	64	64	64	64	64	64
(5 fps) Auto Collaborate Video Channel (kbps)	500	500	500	500	500	500
Auto Collaborate Audio Channel (kbps)	64	64	64	64	64	64
<b>Interoperability Video Channel (kbps)</b>	<b>768</b>	<b>768</b>	<b>768</b>	<b>768</b>	<b>768</b>	<b>768</b>
<b>Interoperability Audio Channel (kbps)</b>	<b>64</b>	<b>64</b>	<b>64</b>	<b>64</b>	<b>64</b>	<b>64</b>
CTS-500/1000 Total Audio and Video (kbps)	5,460	4,960	4,460	3,710	2,960	2,460
CTS-3000/3200 Total Audio and Video (kbps)	13,588	12,088	10,588	8,338	6,088	4,588
CTS-500/1000 total bandwidth (Including Layer 2-Layer 4 overhead)	6.5 Mbps	6.0 Mbps	5.3 Mbps	4.4 Mbps	3.5 Mbps	3 Mbps
CTS-3000/3200 total bandwidth (Including Layer 2-Layer 4 overhead)	16.3 Mbps	14.5 Mbps	12.7 Mbps	10 Mbps	7.3 Mbps	5.5 Mbps

Subsequently, with the release of CTS software version 1.4, the auxiliary video/Auto Collaborate video channel was expanded to transmit 30 frame-per-second of video (from the previous level of 5 frames-per-second) by leveraging an optional, dedicated presentation codec. The 30 fps Auto Collaborate video channel requires 4 Mbps of video bandwidth (along with 64 kbps of audio bandwidth). In comparison, the 5 fps Auto Collaboration feature required only 500 kbps of video bandwidth (along with 64 kbps of audio bandwidth). This change has been highlighted in bold in [Table 4-3](#). Therefore, the net increase in bandwidth required to support this 30 fps Auto Collaborate feature is 3.5 Mbps (exclusive of network overhead). When the 20% network-overhead overprovisioning rule is applied, the additional bandwidth required to support this 30 fps Auto Collaborate feature becomes about 4.2 Mbps for any TelePresence system, regardless of the number of segments, resolution-levels, and motion-handling capabilities configured for TelePresence primary video.

**Table 4-3 Cisco TelePresence Software Version 1.4 Bandwidth Requirements (Including the 30 fps Auto Collaborate Feature)**

Resolution	1080p	1080p	1080p	720p	720p	720p
Motion Handling	Best	Better	Good	Best	Better	Good
Video per Screen (kbps)	4000	3500	3000	2250	1500	1000
Audio per Microphone (kbps)	64	64	64	64	64	64
<b>(30 fps) Auto Collaborate Video Channel (kbps)</b>	<b>4000</b>	<b>4000</b>	<b>4000</b>	<b>4000</b>	<b>4000</b>	<b>4000</b>
Audio Add-In channel (kbps)	64	64	64	64	64	64
Interoperability Video Channel (kbps)	768	768	768	768	768	768
Interoperability Audio Channel (kbps)	64	64	64	64	64	64
CTS-500/1000 Total Audio and Video (kbps)	8,960	8,460	7,960	7,210	6,460	5,960
CTS-3000/3200 Total Audio and Video (kbps)	17,088	15,588	14,088	11,838	9,588	8,088
CTS-500/1000 total bandwidth (Including Layer 2-Layer 4 overhead)	10.9 Mbps	10.2 Mbps	9.5 Mbps	8.6 Mbps	7.7 Mbps	7.2 Mbps
CTS-3000/3200 total bandwidth (Including Layer 2-Layer 4 overhead)	20.5 Mbps	18.7 Mbps	16.9 Mbps	14.2 Mbps	11.5 Mbps	9.7 Mbps

In conclusion, it bears repeating that as Cisco TelePresence software continues to evolve and add new feature support, the bandwidth requirements for TelePresence will correspondingly evolve and expand.

## Burst Requirements

So far, we have discussed bandwidth in terms of bits per second (i.e., how much traffic is sent over a one second interval). However, when provisioning bandwidth and configuring queuing, shaping, and policing commands on routers and switches, burst must also be taken into account. Burst is defined as the amount of traffic (generally measured in bytes) transmitted per millisecond which exceeds the per-second average. For example, a CTS-3000 running at 1080p-Best at approximately 15 Mbps divides evenly into approximately 1,966 bytes per millisecond ( $15 \text{ Mbps} \div 1,000 \text{ milliseconds}$ ).

Cisco TelePresence operates at 30 frames per second. This means that every 33ms a video frame is transmitted; we refer to this as a frame interval. Each frame consists of several thousand bytes of video payload, and therefore each frame interval consists of several dozen packets, with an average packet size of 1,100 bytes per packet. However, because video is variable in size (due to the variability of motion in the encoded video), the packets transmitted by the codec are not spaced evenly over each 33ms frame interval, but rather are transmitted in bursts measured in shorter intervals. Therefore, while the overall bandwidth (maximum) averages out to 15 Mbps over one second, when measured on a per millisecond basis the packet transmission rate is highly variable, and the number of bytes transmitted per millisecond for a 15 Mbps per second call bursts well above the 1,966 bytes per millisecond average. Therefore, adequate burst tolerance must be accommodated by all switch and router interfaces in the path (platform-specific recommendations are detailed in the subsequent design chapters).

## TelePresence Latency Requirements

Cisco TelePresence has a network latency target of 150 ms; this target does not include codec processing time, but purely network flight time.

There may be scenarios, however, where this latency target may not always be possible to achieve, simply due to the laws of physics and the geographical distances involved. Therefore, TelePresence codecs have been designed to sustain high levels of call quality even up to 200 ms of latency. Beyond this threshold (which we refer to as ‘Latency Threshold 1’) a warning message appears on the screen indicating that network conditions may be affecting call quality. Nonetheless, the call continues. If network latency exceeds 400 ms (which we refer to as ‘Latency Threshold 2’) another warning message appears on the screen and the call quality steadily degrades as latency increases. Visually, the call quality is the same, but aurally the lagtime between one party speaking and the other party responding becomes unnaturally excessive. In the original release of the TelePresence codec, calls were self-terminated by the codec if network latency increased beyond 400 ms. However, due to some unique customer requirements, such as some customers looking at provisioning TelePresence calls over satellite circuits, this behavior changed for release 1.1 of the codec, in which the calls were no longer terminated if Latency Threshold 2 was exceeded. Nonetheless, should customers choose to provision TelePresence over such circuits, user expectations need to be adjusted accordingly.

Network latency time can be broken down further into fixed and variable components:

- Serialization (fixed)
- Propagation (fixed)
- Queuing (variable)

Serialization refers to the time it takes to convert a Layer 2 frame into Layer 1 electrical or optical pulses onto the transmission media. Therefore, serialization delay is fixed and is a function of the line rate (i.e., the clock speed of the link). For example, a 45 Mbps DS3 circuit would require 266  $\mu$ s to serialize a 1500 byte Ethernet frame onto the wire. At the circuit speeds required for TelePresence (generally speaking DS3 or higher), serialization delay is not a significant factor in the overall latency budget.

The most significant network factor in meeting the latency targets for TelePresence is propagation delay, which can account for over 90% of the network latency time budget. Propagation delay is also a fixed component and is a function of the physical distance that the signals have to travel between the originating endpoint and the receiving endpoint. The gating factor for propagation delay is the speed of light: 300,000 km/s or 186,000 miles per second. Roughly speaking, the speed of light in an optical fiber is slightly less than one third the speed of light in a vacuum. Thus, the propagation delay works out to be approximately 6.3  $\mu$ s per km or 8.2  $\mu$ s per mile.

Another point to keep in mind when calculating propagation delay is that optical fibers are not always physically placed over the shortest path between two geographic points, especially over transoceanic links. Due to installation convenience, circuits may be hundreds or thousands of kilometers longer than theoretically necessary.

Nonetheless, the network flight-time budget of 150 ms allows for nearly 24,000 km or 15,000 miles worth of propagation delay (which is approximately 60% of the earth’s circumference); the theoretical worst-case scenario (exactly half of the earth’s circumference) would require only 126 ms. Therefore, this latency target should be achievable for virtually any two locations on the planet, given relatively direct transmission paths. However, for some of the more extreme scenarios, user expectations may have to be set accordingly, as there is little a network administrator can do about increasing the speed of light.

Given the end-to-end latency targets and thresholds for TelePresence, the network administrator also must know how much of this budget is to be allocated to the service provider and how much to the enterprise. The general recommendation for this split is 80:20, with 80% of the latency budget allocated to the service provider (demarc-to-demarc) and 20% to the enterprise (codec-to-demarc on one side and

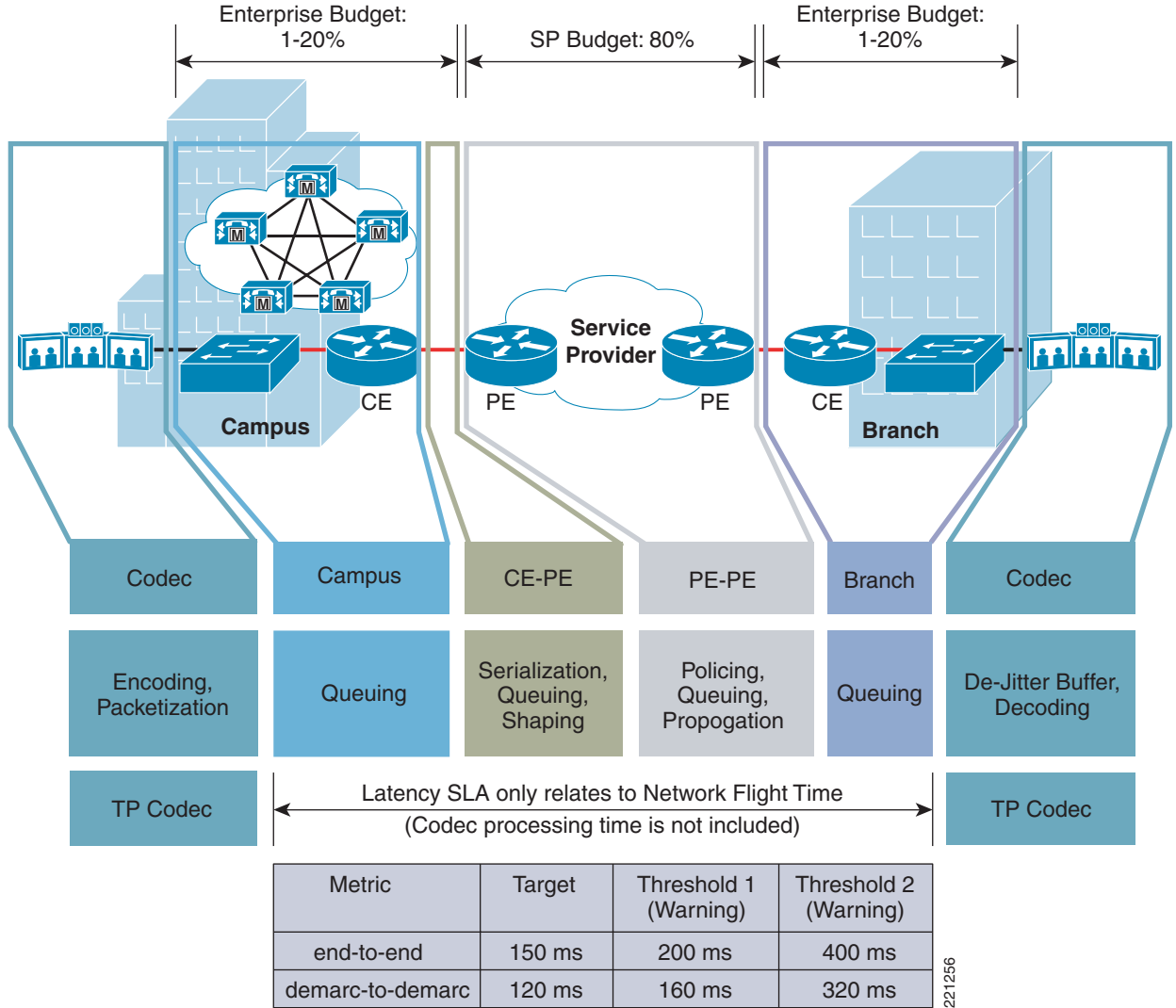
demarc-to-codec on the other). However, some enterprise networks may not require a full 20% of the latency budget and thus may reallocate their allowance to a 90:10 service provider-to-enterprise split, or whatever the case may be. The main point is that a fixed budget needs to be clearly apportioned to both the service provider and to the enterprise, such that the network administrators can design their networks accordingly. Given the target (150ms), threshold1 (200ms), and the service provider-enterprise split of 80:20 or 90:10, it is recommended that SPs engineer their network to meet the target, but base their SLA on threshold1. Threshold1 provides global coverage between any two sites on the planet and allows the SP to offer a 100% guarantee that their network (demarc-to-demarc) will never exceed 160ms (80% of threshold1).

Another point to bear in mind here is the additional latency introduced by multipoint resources. Latency is always measured from end-to-end (i.e., from codec1 to codec2). However, in a multipoint call the media between the two codecs traverses a Multipoint Switch. The multipoint switch itself introduces approximately 20ms of latency, and the path from codec1 to the MS and from the MS to codec2 may be greater than the path between codec1 and codec2 directly, depending on the physical location of the MS. Therefore, when engineering the network with respect to latency, one must calculate both scenarios for every TelePresence System deployed: one for the path between each system and every other system for point-to-point call, and a second for the path between each system, through the MS, to every other system.

The final TelePresence latency component to be considered is queuing delay, which is variable. Queuing delay is a function of whether a network node is congested and what the scheduling QoS policies are to resolve congestion events. Given that the latency target for TelePresence is very tight and, as has been shown, the majority of factors contributing to the latency budget are fixed, careful attention has to be given to queuing delay, as this is the only latency factor that is directly under the network administrator's control via QoS policies.

The latency targets, thresholds and service provider-to-enterprise splits are illustrated in [Figure 4-2](#).

Figure 4-2 Network Latency Target and Thresholds for Cisco TelePresence



## TelePresence Jitter Requirements

Cisco TelePresence has a peak-to-peak jitter target of 10 ms. Jitter is defined as the variance in network latency. Thus, if the average latency is 100 ms and packets are arriving between 95 ms and 105 ms, the peak-to-peak jitter is defined as 10 ms. Measurements within the Cisco TelePresence codecs use peak-to-peak jitter.

Similar to the latency service level requirement, Cisco TelePresence codecs have built in thresholds for jitter to ensure a high quality user experience. Specifically, if peak-to-peak jitter exceeds 20 ms (which we call Jitter Threshold 1) for several seconds, then two things occur:

- A warning message appears at the bottom of the 65" plasma display indicating that the network is experiencing congestion and that call quality may be affected.
- The TelePresence codecs downgrade to a lower level of motion handling quality within the given resolution.

As previously mentioned, Cisco TelePresence codecs have three levels of motion handling quality within a given resolution, specifically 720p-Good, 720p-Better, and 720p-Best and 1080p-Good, 1080p-Better, and 1080p-Best. Therefore, for example, if a call at 1080p-Best would exceed Jitter Threshold 1 (20 ms) for several seconds, the codec would display the warning message in and would downgrade the motion handling quality to 1080p-Good. Similarly a call at 720p-Best would downgrade to 720-Good. Incidentally, downgraded calls do not automatically upgrade should network conditions improve, because this could cause a “flapping” effect where the call upgrades and then downgrades again, over and over.

A second jitter threshold (Jitter Threshold 2) is also programmed into the TelePresence codecs, such that if peak-to-peak jitter exceeds 40 ms for several seconds, then two things occur. The TelePresence codecs:

- Self-terminate the call.
- Display an error message on the 7975G IP Phone indicating that the call was terminated due to excessive network congestion.

Finally, as with latency, the jitter budget is proportioned between the service provider and enterprise networks. Unfortunately, unlike latency or packet loss, peak-to-peak jitter is not necessarily cumulative. Nonetheless, simply for the sake of setting a jitter target for each party, the recommended peak-to-peak jitter split is 50/50 between the service provider and enterprise, such that each group of network administrators can design their networks to a clear set of jitter targets and thresholds. Also like latency, this split may be negotiated differently between the service provider and enterprise to meet certain unique scenarios, such as satellite connections. Again, the main point is that a fixed jitter budget needs to be clearly apportioned to both the service provider and to the enterprise, such that the end-to-end target and thresholds are not exceeded.

It is recommended that SPs engineer their network to meet the target, but base their SLA on threshold1. Threshold1 provides global coverage between any two sites on the planet and allows the SP to offer a 100% guarantee that their network (demarc-to-demarc) will never exceed 10ms of jitter (50% of threshold1).

The TelePresence Jitter targets and thresholds are summarized in [Table 4-4](#).

**Table 4-4** *TelePresence Jitter Targets, Thresholds, and Service Provide/Enterpriser Splits*

Metric	Target	Threshold 1 (Warning and Downgrade)	Threshold 2 (Call Drop)
End-to-end	10 ms	20 ms	40 ms
Service Provider	5 ms	10 ms <sup>1</sup>	20 ms

1. SP SLA should be based on Threshold 1.

## TelePresence Loss Requirements

Cisco TelePresence is highly sensitive to packet loss, and as such has an end-to-end packet loss target of 0.05%.

It may be helpful to review a bit of background information to better understand why TelePresence is so extremely sensitive to packet loss. Specifically, let’s review how much information is actually needed to transmit a 1080p30 HD video image, which is the highest video transmission format used by Cisco TelePresence codecs. The first parameter (1080) refers to 1080 lines of horizontal resolution, which are matrixed with 1920 lines of vertical resolution (as per the 16:9 Widescreen Aspect Ratio used in High Definition video formatting), resulting in 2,073,600 pixels per screen. The second parameter, p, indicates a progressive scan, which means that every line of resolution is refreshed with each frame (as opposed to an interlaced scan, which would be indicated with an i and would mean that every other line is

refreshed with each frame). The third parameter 30 refers to the transmission rate of 30 frames per second. While video sampling techniques may vary, each pixel has approximately 3 Bytes of color and/or luminance information. When all of this information is factored together (2,073,600 pixels x 3 Bytes x 8 bits per Byte x 30 frames per second), it results in approximately 1.5 Gbps of information. This is illustrated in Figure 4-3.

**Figure 4-3 1080p30 Information Breakdown**



As shown earlier in this chapter, Cisco TelePresence codecs transmit at approximately 5 Mbps (max) per 1080p display, which translates to over 99% compression. Therefore, the overall effect of packet loss is proportionally magnified and dropping even one packet in 2000 (0.05% packet loss) becomes readily noticeable to end users.

Similar to the latency and jitter service level requirement, Cisco TelePresence codecs have built in thresholds for packet loss to ensure a high-quality user experience. Specifically, if packet loss exceeds 0.10% (or 1 in 1000 packets, which we call Loss Threshold 1) for several seconds, then two things occur:

- A warning message appears at the bottom of the on the 65" plasma display indicating that the network is experiencing congestion and that call quality may be affected.
- The TelePresence codecs downgrade to a lower level of motion handling quality within the given resolution.

As previously mentioned, Cisco TelePresence codecs have three levels of motion handling quality within a given resolution, specifically 720p-Good, 720p-Better, and 720p-Best and 1080p-Good, 1080p-Better, and 1080p-Best. Therefore, for example, if a call at 1080p-Best would exceed Loss Threshold 1 (0.10%) for several seconds, the codec would display the warning message and would downgrade the motion handling quality to 1080p-Good. Similarly a call at 720p-Best would downgrade to 720p-Good in the same scenario. Incidentally, downgraded calls do not automatically upgrade should network conditions improve, because this could cause a "flapping" effect where the call upgrades and then downgrades again, over and over.

A second packet loss threshold (Loss Threshold 2) is also programmed into the TelePresence codecs, such that if packet loss exceeds 0.20% (or 1 in 500 packets) for several seconds, then two things occur. The TelePresence codecs:

- Self-terminate the call.
- Display an error message on the 7975G IP Phone indicating that the call was terminated due to excessive network congestion.

Finally, as with previously defined service level requirements, the loss budget is proportioned between the service provider and enterprise networks. The recommend split is 50/50 between the service provider and enterprise, such that each group of network administrators can design their networks to a clear set of packet loss targets and thresholds. Of course, This split may be negotiated differently between the service provider and enterprise to meet certain unique scenarios, such as satellite connections. Again, the main point is that a fixed packet loss budget needs to be clearly apportioned to both the service provider and to the enterprise, such that the end-to-end target and thresholds are not exceeded.

It is recommended that SPs engineer their network to meet the target, but base their SLA on threshold1. Threshold1 provides global coverage between any two sites on the planet and allows the SP to offer a 100% guarantee that their network (demarc-to-demarc) will never exceed .05% loss (50% of threshold1).

The TelePresence packet loss targets and thresholds are summarized in [Table 4-5](#).

**Table 4-5** *TelePresence Jitter Targets, Thresholds, and Service Provider/Enterprise Splits*

Metric	Target	Threshold 1 Warning and Downgrade)	Threshold 2 (Call Drop)
End-to-end	0.05% (1 in 2000)	0.10% (1 in 1000)	0.20 (1 in 500)
Service Provider	.025%	.05% <sup>1</sup>	.10%

1. SP SLA should be based on Threshold 1.

## Tactical QoS Design Best Practices for TelePresence

Once the service level requirements of TelePresence are defined, then the network administrator can proceed to the next step of the QoS deployment cycle (illustrated in [Figure 4-1](#)) of designing the actual policies.

A couple of tactical QoS best practices design principles bear mentioning at this point, as these serve to improve the efficiency and scope of your QoS designs. The first principle is to always deploy QoS in hardware, rather than software, whenever a choice exists. Cisco Catalyst switches perform QoS operations in hardware Application Specific Integrated Circuits (ASICs) and as such have zero CPU impact; Cisco IOS routers, on the other hand, perform QoS operations in software, resulting in a marginal CPU impact, the degree of which depends on the platform, the policies, the link speeds, and the traffic flows involved. So, whenever supported, QoS policies like classification, marking/remarking, and/or policing can all be performed at line rates with zero CPU impact in Catalyst switches (as opposed to IOS routers), which makes the overall QoS design more efficient. A practical example of how this principle is applied is as follows: while all nodes in the network path must implement queuing policies, classification policies should be implemented in Cisco Catalyst hardware as close to the source of the traffic as possible (e.g., on the access edge switch to which the TelePresence System is attached), rather than waiting until the traffic hits the WAN router to be classified.

Another best practice principle to keep in mind is to follow industry standards whenever possible, as this extends the effectiveness of your QoS policies beyond your direct administrative control. For example, if you mark a realtime application, such as VoIP, to the industry standard recommendation as defined in RFC 3246 (An Expedited Forwarding Per-Hop Behavior), then you will no doubt provision it with strict priority servicing at every node within your enterprise network. Additionally, if you handoff to a service provider following this same industry standard, they will similarly provision traffic marked Expedited Forwarding (EF - or DSCP 46) in a strict priority manner. Therefore, even though you do not have direct administrative control of the QoS policies within the service provider's cloud, you have extended the influence of your QoS design to include your service provider's cloud, simply by following the industry

standard recommendations. Therefore, in line with this principle, it would be beneficial to briefly consider some of the relevant industry standards to QoS design, particularly as these relate to TelePresence.

## Relevant Industry Standards and Recommendations

Let's briefly review some of the relevant DiffServ standards and recommendations and see how these relate to TelePresence QoS design.



### Note

---

Although Cisco TelePresence requires Cisco CallManager (CCM) 5.1 (or higher) for call processing, and CCM 5.x supports Resource Reservation Protocol (RSVP) for Call Admission Control, the initial phase of the TelePresence solution does not require leveraging RSVP functionality (RSVP remains optional during this phase); therefore, the discussion in this paper focuses on DiffServ QoS designs and standards for Cisco TelePresence (not IntServ/RSVP).

---

### RFC 2474 Class Selector Code Points

This standard defines the use of 6 bits in the IPv4 and IPv4 Type of Service (ToS) byte, termed Differentiated Services Code Points (DSCP). Additionally, this standard introduces Class Selector codepoints to provide backwards compatibility for legacy (RFC 791) IP Precedence bits.

### RFC 2597 Assured Forwarding Per-Hop Behavior Group

This standard defines the Per-Hop Behavior of the Assured Forwarding (AF) classes. Four AF classes are defined: AF1, AF2, AF3, and AF4. Additionally, each class has three states of increasing Drop Preference assigned within it, corresponding to three traffic states: conforming (analogous to a green traffic light signal), exceeding (analogous to a yellow traffic light signal), and violating (analogous to a red traffic light signal). For example, conforming AF1 traffic would be marked to AF11 (the second 1 representing the lowest Drop Preference setting), exceeding traffic would have its Drop Preference increased to AF12, and violating traffic would have its Drop Preference increased further to AF13. When such traffic enters a node experiencing congestion, AF13 traffic is more aggressively dropped than AF12 traffic, which in turn is more aggressively dropped than AF11 traffic.

### RFC 3246 An Expedited Forwarding Per-Hop Behavior

This standard defines an Expedited Forwarding (EF) Per-Hop Behavior for realtime applications. When traffic marked EF enters a node experiencing congestion, it receives strict priority behavior.

### RFC 3662 A Lower Effort Per-Domain Behavior for Differentiated Services

This informational RFC defines a less than Best Effort service for undesired applications and specifies that such applications should be marked to Class Selector 1 (CS1).

### Cisco's QoS Baseline

While the IETF RFC standards provided a consistent set of per-hop behaviors for applications marked to specific DSCP values, they never specified which application should be marked to which DiffServ Codepoint value. Much confusion and disagreements over matching applications with standards-defined

codepoints led Cisco in 2002 to put forward a standards-based marking recommendation in their strategic architectural QoS Baseline document. Eleven different application classes that could exist within the enterprise were examined and extensively profiled, and then matched to their optimal RFC-defined Per-Hop Behaviors (PHBs). The application-specific marking recommendations from Cisco's QoS Baseline of 2002 are summarized in [Figure 4-4](#).

**Figure 4-4 Cisco's QoS Baseline Marking Recommendations**

Application	L3 Classification		IETF
	PHB	DSCP	RFC
Routing	CS6	48	RFC 2474
Voice	EF	46	RFC 3246
Interactive Video	AF41	34	RFC 2597
Streaming Video	CS4	32	RFC 2474
Mission-Critical Data	AF31	26	RFC 2597
Call Signaling	CS3	24	RFC 2474
Transactional Data	AF21	18	RFC 2597
Network Management	CS2	16	RFC 2474
Bulk Data	AF11	10	RFC 2597
Best Effort	0	0	RFC 2474
Scavenger	CS1	8	RFC 2474

220199

The adoption of Cisco's QoS Baseline was a great step forward in QoS consistency, not only within Cisco, but also within the industry in general.

## RFC 4594 Configuration Guidelines for DiffServ Classes

More than four years after Cisco put forward its QoS Baseline document, RFC 4594 was formally accepted as an informational RFC (in August 2006).

Before getting into the specifics of RFC 4594, it is important to comment on the difference between the IETF RFC categories of informational and standard. An informational RFC is an industry recommended best practice, while a standard RFC is an industry requirement. Therefore RFC 4594 is a set of formal DiffServ QoS configuration best practices, not a requisite standard.

RFC 4594 puts forward twelve application classes and matches these to RFC-defined Per-Hop Behaviors (PHBs). These application classes and recommended PHBs are summarized in [Figure 4-5](#).

**Figure 4-5 RFC 4594 Marking Recommendations**

Application	L3 Classification		IETF
	PHB	DSCP	RFC
Network Control	CS6	48	RFC 2474
VoIP Telephony	EF	46	RFC 3246
Call Signaling	CS5	40	RFC 2474
Multimedia Conferencing	AF41	34	RFC 2597
Real-Time Interactive	CS4	32	RFC 2474
Multimedia Streaming	AF31	26	RFC 2597
Broadcast Video	CS3	24	RFC 2474
Low-Latency Data	AF21	18	RFC 2597
OAM	CS2	16	RFC 2474
High-Throughput Data	AF11	10	RFC 2597
Best Effort	DF	0	RFC 2474
Low-Priority Data	CS1	8	RFC 3662

220200

It is fairly obvious that there are more than a few similarities between Cisco's QoS Baseline and RFC 4594, as there should be, since RFC 4594 is essentially an industry-accepted evolution of Cisco's QoS Baseline. However, there are some differences that merit attention.

The first set of differences are minor, as they involve mainly nomenclature. Some of the application classes from the QoS Baseline have had their names changed in RFC 4594. These changes in nomenclature are summarized in [Table 4-6](#).

**Table 4-6 Nomenclature Changes from Cisco QoS Baseline to RFC 4594**

Cisco QoS Baseline Class Names	RFC 4594 Class Names
Routing	Network Control
Voice	VoIP Telephony
Interactive Video	Multimedia Conferencing
Streaming Video	Multimedia Streaming
Transactional Data	Low-Latency Data
Network Management	Operations/Administration/Management (OAM)
Bulk Data	High-Throughput Data
Scavenger	Low-Priority Data

The remaining changes are more significant. These include one application class deletion, two marking changes, and two new application class additions. Specifically:

- The QoS Baseline Locally-Defined Mission-Critical Data class has been deleted from RFC 4594.
- The QoS Baseline marking recommendation of CS4 for Streaming Video has been changed in RFC 4594 to mark Multimedia Streaming to AF31.

- The QoS Baseline marking recommendation of CS3 for Call Signaling has been changed in RFC 4594 to mark Call Signaling to CS5.
- A new video class has been added to RFC 4594: Real-Time Interactive, which is to be marked CS4. This was done to differentiate between lower-grade desktop video telephony (referred to as Multimedia Conferencing) and higher-grade videoconferencing and TelePresence. Multimedia Conferencing uses the AF4 class and is subject to markdown policies, while TelePresence uses the CS4 class and is not subject to markdown.
- A second new video class has been added to RFC 4594: Broadcast video, which is to be marked CS3. This was done to differentiate between lower-grade desktop video streaming (referred to as Multimedia Streaming) and higher-grade Broadcast Video applications. Multimedia Streaming uses the AF3 class and is subject to markdown policies, while Broadcast Video uses the CS3 class and is not subject to markdown.

The most significant of the differences between Cisco's QoS Baseline and RFC 4594 is the RFC 4594 recommendation to mark Call Signaling to CS5. Cisco has just completed a lengthy and expensive marking migration for Call Signaling from AF31 to CS3 (as per the original QoS Baseline of 2002), and as such, there are no plans to embark on another marking migration in the near future. It is important to remember that RFC 4594 is an informational RFC (i.e., an industry best-practice) and not a standard. Therefore, lacking a compelling business case at the time of writing, Cisco plans to continue marking Call Signaling as CS3 until future business requirements arise that necessitate another marking migration.

Therefore, for the remainder of this document, RFC 4594 marking values are used throughout, with the one exception of swapping Call-Signaling marking (to CS3) and Broadcast Video (to CS5). These marking values are summarized in [Figure 4-6](#).

**Figure 4-6 Cisco-Modified RFC4594 Marking Values (Call-Signaling is Swapped with Broadcast Video)**

Application	L3 Classification		IETF
	PHB	DSCP	RFC
Network Control	CS6	48	RFC 2474
VoIP Telephony	EF	46	RFC 3246
Broadcast Video	CS5	40	RFC 2474
Multimedia Conferencing	AF41	34	RFC 2597
Real-Time Interactive	CS4	32	RFC 2474
Multimedia Streaming	AF31	26	RFC 2597
Call Signaling	CS3	24	RFC 2474
Low-Latency Data	AF21	18	RFC 2597
OAM	CS2	16	RFC 2474
High-Troughput Data	AF11	10	RFC 2597
Best Effort	DF	0	RFC 2474
Low-Priority Data	CS1	8	RFC 3662

221258

## Classifying TelePresence

One of the first questions to be answered relating to TelePresence QoS design is: should TelePresence be assigned to a dedicated class or should it be assigned to the same class as existing Videoconferencing/Video Telephony? The answer to this question directly relates to whether TelePresence has the same service-level requirements as these other two interactive video applications or whether it has unique service level requirements. [Table 4-7](#) summarizes the service level requirements of both generic Videoconferencing applications and TelePresence.

**Table 4-7 Service Level Requirements of Generic Video-Conferencing and TelePresence**

Service Level Parameter (Target Values)	(Generic) Videoconferencing/Video Telephony	Cisco TelePresence
<b>Bandwidth</b>	384 kbps or 768 kbps + network overhead	1.5 Mbps to 12.6 Mbps + network overhead
<b>Latency</b>	400-450 ms latency	150 ms latency
<b>Jitter</b>	30-50 ms peak-to-peak jitter	10 ms peak-to-peak jitter
<b>Loss</b>	1% random packet loss	0.05% random packet loss

From [Table 4-7](#) it becomes apparent that TelePresence has unique (and higher/tighter) service level requirements than do generic Videoconferencing/Video Telephony applications; therefore, TelePresence requires a dedicated class along with a dedicated classification marking value.

Videoconferencing/Video Telephony applications have traditionally been marked to (RFC 2597) Assured Forwarding Class 4, which is the recommendation from both the Cisco QoS Baseline as well as RFC 4594. However, the Assured Forwarding (AF) Per-Hop Behavior (PHB) includes policing (to conforming, exceeding, and violating traffic rates), as well as correspondingly increasing the Drop Preferences (to Drop Preference 1, 2, and 3 respectively), and ultimately dropping traffic according to the Drop Preference markings. TelePresence traffic has a very low tolerance to drops (0.05%) and therefore would not be appropriately serviced by an AF PHB.

Because of the low-latency and jitter service-level requirements of TelePresence, it may seem attractive to assign it an (RFC 3246) Expedite Forwarding (EF) Per-Hop Behavior; after all, there is nothing in RFC 3246 that dictates that only VoIP can be assigned to this PHB. However, it is important to recognize that VoIP behaves considerably differently than video. As previously mentioned, VoIP has constant packet sizes and packet rates, whereas video packet sizes vary and video packet rates also vary in a random and bursty manner. Thus, if both video and voice were assigned to the same marking value and class, (bursty) video could easily interfere with (well-behaved) voice. Therefore, for both operational and capacity planning purposes, it is recommended not to mark both voice and video to EF. This recommendation is reflected in both the Cisco QoS Baseline as well as RFC 4594.

What then should TelePresence be marked to? The best formal guidance is provided in RFC 4594, where a distinction is made between a Multimedia Conferencing (i.e., generic Videoconferencing/Video Telephony) service class and a Real-Time Interactive service class. The Real-Time Interactive service class is intended for inelastic video flows, such as TelePresence. The recommended marking for this Real-Time Interactive service class, and thus **the recommended marking for TelePresence is Class Selector 4 (CS4)**.

## Policing TelePresence

**In general, policing TelePresence traffic should be avoided whenever possible, although some exceptions exist.**

As previously mentioned, TelePresence is highly sensitive to drops (with a 0.05% packet loss target); therefore policing TelePresence traffic rates with either a Single Rate Three Color Marker (as defined in RFC 2697) or a Two Rate Three Color Marker (as defined in RFC 2698) could be extremely detrimental to TelePresence flows and ultimately ruin the high-level of user experience that this application is intended to deliver.

**However, there are three places where TelePresence traffic may be legitimately policed over the network.**

**The first automatically occurs if TelePresence is assigned to a Low-Latency Queue (LLQ) within Cisco IOS routers at the WAN or VPN edge.** This is because any traffic assigned to a LLQ is automatically policed by an implicit policer set to the exact value as the LLQ rate. For example, if TelePresence is assigned a LLQ of 15 Mbps, it is also implicitly policed by the LLQ algorithm to exactly 15 Mbps; any excess TelePresence traffic is dropped.



### Note

The implicit policer within the LLQ feature is only active when LLQ is active. In other words, since queuing only engages when there is congestion, LLQ never engages unless the link is physically congested or a (hierarchical QoS) shaper forces LLQ to engage prior to physical link congestion. Similarly, the implicit policer of LLQ never engages unless there is physical congestion on the link or a (hierarchical QoS) shaper forces it to engage prior to physical link congestion. Put another way, when the physical link is un-congested and/or a hierarchical QoS shaper is inactive, neither LLQ nor the implicit policer of LLQ is active.

**The second most common place that TelePresence is likely to be policed in the network is at the service provider's provider edge (PE) routers, in the ingress direction.** Service providers need to police traffic classes, especially realtime traffic classes, to enforce service contracts and prevent possible oversubscription on their networks and thus ensure service level agreements.

**The third place (and optional) place, where policing TelePresence may prove beneficial in the network is at the campus access edge.** Administrators can deploy access-edge policers for security purposes to mitigate the damage caused by the potential abuse of trusted switch ports. Since TelePresence endpoints can mark TelePresence flows to the recommended 802.1Q/p CoS value (CoS 4) and DSCP codepoint value (CS4), the network administrator may choose to trust the CoS or DSCP values received from these ports. However, if a disgruntled employee gains physical access to the TelePresence switch ports, they may send whatever traffic they choose to over these ports and their flows are trusted over the network. Such rogue traffic flows may hijack voice or video queues and easily ruin call or video quality over the QoS-provisioned network infrastructure. Therefore, the administrator may choose to limit the scope of damage that such network abuse may present by configuring access-edge policers on TelePresence switch ports to remark (to Scavenger: DSCP CS1) or drop out-of-profile traffic originating on these ports (e.g., CS4 traffic exceeding 15 Mbps). Supporting this approach, RFC 4594 recommends edge policing the Real-Time Interactive service class via a single-rate policer.

## Queuing TelePresence

To achieve the high-levels of service required by the Cisco TelePresence Experience, queuing must be enabled on every node along the path to provide service guarantees, regardless of how infrequently congestion may occur on certain nodes (i.e., congestion can and does occur even on very high-bandwidth mediums). If queuing is not properly configured on every node, the Cisco TelePresence eXperience (CTX) cannot be guaranteed.

RFC 4594 specifies the minimum queuing requirement of the Real-Time Interactive service class to be a rate-based queue (i.e., a queue that has a guaranteed minimum bandwidth rate). However, RFC 4594 makes an allowance that while **the PHB for Real-Time Interactive service class** should be configured to provide high bandwidth assurance, it **may be configured as a second EF PHB** that uses relaxed performance parameters, a rate scheduler, and a CS4 DSCP value.

This means that, for example, TelePresence, which has been assigned to this Real-Time Interactive service class, can be queued with either a guaranteed rate non-priority queue (such as a Cisco IOS Class-Based Weighted Fair Queue-CBWFQ) or a guaranteed-rate strict priority queue (such as a Cisco IOS Low-Latency Queue-LLQ); in either case, TelePresence is to be marked as Class Selector 4 (and not EF).

Therefore, since RFC 4594 allows for the Real-Time Interactive service-class to be given a second EF PHB and because of the low latency, low jitter, and low loss requirements of TelePresence, **it is recommended to place TelePresence in a strict-priority queue**, such as a Cisco IOS LLQ or a Cisco Catalyst hardware priority queue whenever possible.

However, an additional provisioning consideration must be taken into account when provisioning TelePresence with a second EF PHB, which relates to the amount of bandwidth of a given link that should be assigned for strict priority queuing. The well-established and widely-deployed Cisco best-practice recommendation is to limit the amount of strict priority queuing configured on an interface to no more than one-third of the link's capacity. This has commonly been referred to as the 33% LLQ Rule.

The rationale behind this rule is that if you assign too much traffic for strict priority queuing, then the overall effect is a dampening of QoS functionality for non-realtime applications. Remember, the goal of convergence is to enable voice, video, and data to transparently co-exist on a single network. When realtime applications such as voice and/or TelePresence dominate a link (especially a WAN/VPN link), then data applications fluctuate significantly in their response times when TelePresence calls are present versus when they are absent, thus destroying the transparency of the converged network.

For example, consider a (45 Mbps) DS3 link configured to support 2 separate CTS-3000 calls, both configured to transmit at full 1080p-Best resolution. Each such call requires 15 Mbps of realtime traffic. Prior to TelePresence calls being placed, data applications have access to 100% of the bandwidth (to simplify the example, we are assuming there are no other realtime applications, such as VoIP, on this link). However, once these TelePresence calls are established, all data applications would suddenly be contending for less than 33% of the link. TCP windowing would take effect and many data applications will hang, time-out, or become stuck in a non-responsive state, which usually translates into users calling the IT help desk complaining about the network (which happens to be functioning properly, albeit in a poorly-configured manner).

To obviate such scenarios, Cisco Technical Marketing has done extensive testing and has found that a significant decrease in data application response times occurs when realtime traffic exceeds one-third of link bandwidth capacity. Extensive testing and customer deployments have shown that a general best queuing practice is to limit the amount of strict priority queuing to 33% of link bandwidth capacity. This strict priority queuing rule is a conservative and safe design ratio for merging realtime applications with data applications.

**Note**

As Cisco IOS software allows the abstraction (and thus configuration) of multiple strict priority LLQs, in such a multiple LLQ context, this design principle would apply to the sum of all LLQs to be within one-third of link capacity.

It is vitally important, however, to understand that this strict priority queuing rule is simply a best practice design recommendation and is not a mandate. There may be cases where specific business objectives cannot be met while holding to this recommendation. In such cases, enterprises must provision according to their detailed requirements and constraints. However, it is important to recognize the tradeoffs involved with over-provisioning strict priority traffic and its negative performance impact on non-realtime-application response times. It is also worth noting that the 33% rule only applies for converged networks. In cases where customers choose to deploy dedicated WAN circuits for their TelePresence traffic, the 33% rule does not apply since TelePresence (and perhaps some nominal amount of management and signaling traffic) is the only traffic on the circuit. In these cases, customers are free to use up to 98% of the link capacity for TelePresence (reserving 2% for routing protocols, network management traffic such as SSH and SNMP, and signaling).

## Shaping TelePresence?

**It is recommended to avoid shaping TelePresence flows unless absolutely necessary.** This is because of the QoS objective of shapers themselves. Specifically, the role of shapers is to delay traffic bursts above a certain rate and to smooth out flows to fall within contracted rates. Sometimes this is done to ensure traffic rates are within a carrier's Committed Information Rate (CIR); other times shaping is performed to protect other data classes from a bursty class.

Shapers temporarily buffer traffic bursts above a given rate and as such introduce variable delay (jitter) as well as absolute delay. Since TelePresence is so sensitive to delay (150 ms) and especially jitter (10 ms), it is recommended not to shape TelePresence flows.

If the objective of the shaper was to meet a carrier's CIRs, this can be achieved by properly provisioning the adequate bandwidth and burst allowances on the circuit.

If the objective of the shaper was to protect other traffic classes from TelePresence bursts, then a better approach would be to explicitly protect each class with a guaranteed minimum bandwidth rate (such as a Cisco IOS CBWFQ).

In either case, a shaper would be a sub-optimal tool to meet the desired objective and would cause quality issues on the TelePresence flows and therefore would not be recommended.

The TelePresence traffic queue (whether you choose to place it in a CBWFQ or a second strict priority LLQ) must be provisioned with the proper mean rate (bits per second) and burst allowance (burst bytes exceeding the mean).

## Compressed RTP (cRTP) with TelePresence

**It is recommended to not enable cRTP for TelePresence.** This is because of the large CPU impact of IP RTP Header Compression and the negligible returns in bandwidth savings it entails at TelePresence circuit speeds.

TelePresence, like VoIP, is encapsulated by IP, UDP, and RTP headers and these headers, when combined, account for 40 bytes per packet (at Layer 3). To enhance bandwidth efficiency, compression tools, like IP RTP Header Compression (cRTP) can reduce this overhead from 40 bytes to 2-5 bytes per packet.

However, it is important to recognize that cRTP is the most computationally-intensive QoS operation in the Cisco IOS toolset. Furthermore, it is only recommended on slow-speed links, usually 768 kbps or less, as it is at these speeds that the bandwidth gain offsets the increased CPU cost of the operation and is only useful for RTP-based applications that have a small amount of payload per packet. On high-speeds links and applications like TelePresence in which the payload of each packet averages 1100 bytes, cRTP offers no benefit and only results in sending the routers CPU through the ceiling. **Therefore, it is recommended to not enable cRTP on links carrying TelePresence.**

## Link Fragmentation and Interleaving (LFI) with TelePresence

Like cRTP, LFI is only useful on slow-speed links (usually 768 kbps or less) and is used to fragment larger data packets into smaller chunks and interleave voice in between them to reduce the serialization and queuing delays for VoIP applications. Since TelePresence packets average 1100 bytes payload per packet, LFI would want to fragment them. This introduces unwanted jitter and out-of-order and late packets into the TelePresence stream. On high-speed links the serialization delay for large packets is inconsequential to VoIP and thus LFI offers no benefit and only results in sending the routers CPU through the ceiling. **Therefore, it is recommended to not implement LFI on links carrying TelePresence.**

## GRE/IPSec Tunnels with TelePresence

Tunneling TelePresence traffic over GRE/IPSec tunnels is supported. The Cisco TelePresence codecs are designed to limit their packets to a maximum of 1200 bytes to leave enough room for GRE/IPSec encapsulation overhead to avoid having the TelePresence traffic fragmented for exceeding the Maximum Transmission Unit (MTU) of any link in the path.

## Place in the Network TelePresence QoS Design

At this point, the strategic QoS business objectives for TelePresence have been defined, the service level-requirements of TelePresence have been specified, and the tactical QoS design approach has been sketched via the best practice principles and recommendations reviewed in the previous section. What remains is to flesh out these sketches into detailed Place-in-the-Network (PIN) platform-specific designs.

As the Cisco TelePresence solution evolves, it will become more complex and touch more Places-in-the-Network. The first deployment model to receive Cisco Verified Design (CVD) certification is the Intra-Enterprise, Point-to-Point Deployment Model (as described in [Chapter 3, “TelePresence Network Deployment Models”](#)). Such deployments will directly impact enterprise campus, branch, and WAN/MAN PINs, as well as service provider edge and core networks.

An addition to the Intra-Enterprise Deployment Model came with the release of the Cisco TelePresence Multipoint Solution, based on the Cisco TelePresence Multipoint Switch (CTMS) product offering. This addition may require an additional PIN, namely the enterprise and/or service provider data center, as these are often the locations where multipoint resources are hosted. However, note that while many customers are beginning to deploy multipoint resources, the addition of multipoint resources within the Intra-Enterprise Deployment Model has not yet received CVD certification.

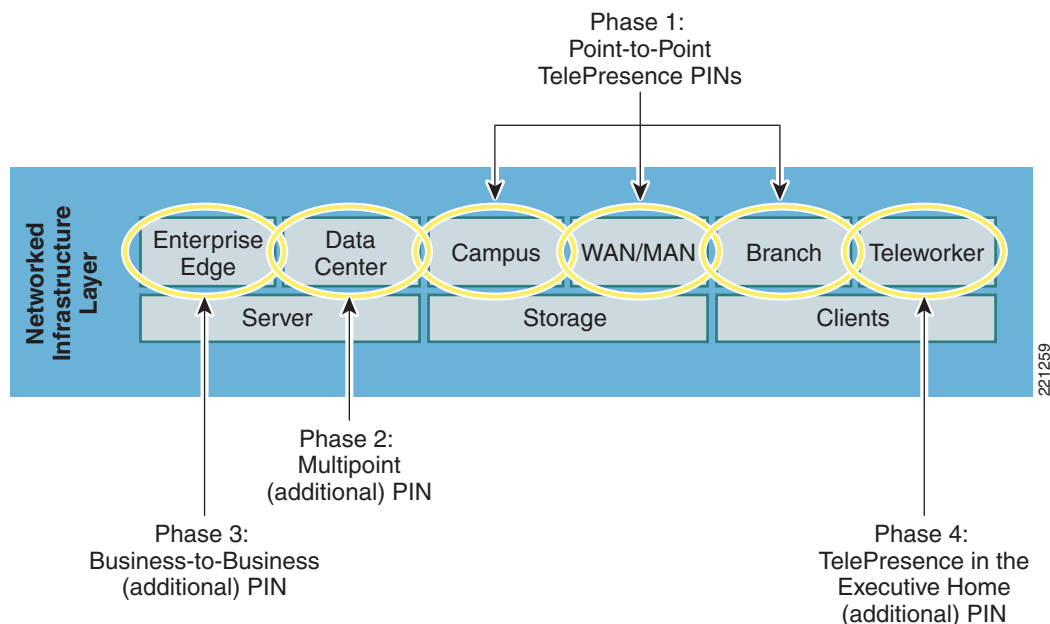
The next phase of TelePresence deployments will begin with the release of the Business-to-Business TelePresence solution, enabling enterprises to move to a Inter-Enterprise Deployment Model (as described in [Chapter 3, “TelePresence Network Deployment Models”](#)). These Inter-Enterprise deployments may be Point-to-Point or Multipoint. With this additional functionality, a new enterprise

PIN, the enterprise edge, will require design modifications. Additionally, service providers will need to develop shared services domains to provide the necessary connectivity, security, and QoS services required to enable this solution. Early Field Trials (EFT) of B2B services have begun. However, the Inter-Enterprise Deployment Model has not yet received CVD certification.

Finally, TelePresence systems are already emanating considerable executive-perk appeal, especially CTS-1000 systems that are designed for an executive's office. Already some executives are deploying TelePresence systems within their homes, taking advantage of very high-speed residential internet access options, like fiber optics to the home. Therefore, an inevitable fourth phase of TelePresence deployments will undoubtedly include the executive teleworker PIN. Early Field Trials (EFT) of TelePresence systems deployed in executive homes has begun. However, the Executive-Class Teleworker Deployment Model has not yet received CVD certification.

The relevant enterprise PINs for the above deployment models, based on the Service Oriented Network Architecture (SONA), specifically the Networked Infrastructure Layer, are illustrated in Figure 4-7.

**Figure 4-7** SONA Networked Infrastructure Layer—Places in the Network (PINs) for Phases 1-4 TelePresence Deployments



The following chapters discuss and detail QoS designs for deploying TelePresence in each of these enterprise PINs. Information is provided on components pending CVD certification to allow customers to plan their network designs and deployment strategies accordingly. However, where detailed CVD design guidance is not yet available, note that the information provided is subject to change pending CVD certification.