Extended Reach: Implementing TelePresence over Cisco Virtual Office

Table of Contents

Overview

This document provides basic implementation guidelines for deploying a Cisco® TelePresence System 500 in the home office over high-speed broadband using a Cisco Virtual Office topology. The main aspects covered in this document are Cisco TelePresence traffic characteristics, spoke router recommendations, and hub recommendations.

A Cisco TelePresence System 500 provides business value in numerous applications:

- **Continuity when there are weather-related travel restrictions**: Employees may not be able to travel to the office due to winter storms, hurricanes, or other severe weather issues.
- **Healthcare**: Doctors can communicate with patients remotely or a hospital can connect with a doctor for various reasons.
- **Going "green"**: TelePresence helps to reduce carbon footprints, as well as reducing office real estate and IT costs.
- **Pandemic planning**: Travel outside the home may be restricted due to influenza or other biohazards. The fact is many events can keep employees from accessing the workplace—displacement events can be as ordinary as a transit strike or a bridge failure. Because these events are impossible to predict, the best strategy is to be prepared and to have an infrastructure in place, tested and ready for when a displacement event occurs.
- **Day extension**: It is becoming increasingly common for job responsibilities to include co-workers spread over multiple time zones. A manager based in the United States may have direct reports in China or India. Cultural differences and language barriers make face-to-face communications more effective than email or audio-only conferences.
- **Critical human resources**: Employee effectiveness can be enhanced by minimizing travel while maintaining the personal touch that Cisco TelePresence provides.

As the cost of business-class broadband decreases and available bandwidth increases, implementing a home office based on the Cisco TelePresence System is a viable way to extend the existing corporate TelePresence systems already in use.

Components

The relevant components to this deployment include the following:

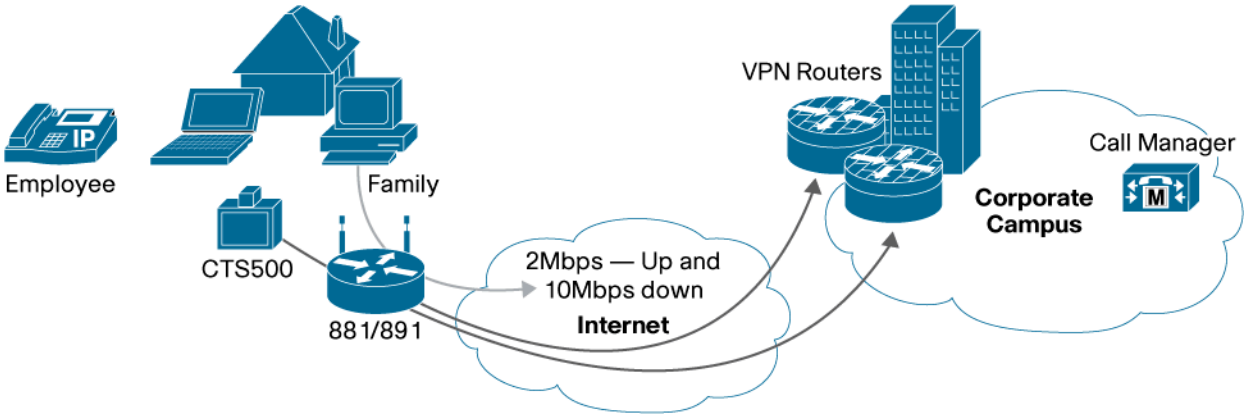
Cisco TelePresence System 500

Cisco Virtual Office router: a Cisco 881/891 Integrated Services Router (ISR) connected to headend aggregation VPN routers (this could be a Cisco 7200 Series Router or Cisco Aggregation Services Router [ASR]). DMVPN is used as the VPN technology for this deployment. Recommended Cisco IOS® Software Releases are Release 15(0)1M for the 891 ISR and Release 12.4(22)T3 for the 881 ISR. Business-class broadband and asymmetrical bandwidth, with a minimum of 2 Mbps uplink. Most service providers offer downlink 4 to 5 times the provisioned uplink bandwidth.

Network Topology

Figure 1 shows Extended Reach Setup with Cisco Virtual Office and CTS 500.

Figure 1. Network Topology for Cisco Virtual Office with Cisco Telepresence System



The Cisco 881W router is directly connected to the cable/DSL modem/router of the service provider. It is recommended that all devices on the home network be connected behind the Cisco Virtual Office router. This way, all traffic will be subject to the QoS policies configured on that router and there will be no negative impact on the video quality.

Cisco TelePresence Traffic Characteristics

This section documents the latency, jitter, and loss targets, thresholds, and threshold behavior for Cisco TelePresence. While the provisioned bandwidth of the business-class broadband connection is recommended to be at least 2 Mbps uplink, the various ISPs and the enterprise corporate network must also provide a service level sufficient for good video quality. Cisco TelePresence has stringent network requirements from all devices encountered on the network. In this type of scenario, a teleworker solution based on Cisco Virtual Office is an ideal choice.

Maximum Bandwidth Consumption per Second

The Cisco TelePresence System 1000 transmits two channels of audio, but can receive up to four channels when participating in a meeting with a Cisco TelePresence System 3000 or in a multipoint meeting. The actual Layer 2–4 overhead required varies by encapsulation method (e.g., Ethernet, HDLC, Packet over SONET, ATM, MPLS, etc.) and whether GRE and/or IPsec tunneling/encryption is used. Rather than leaving it up to the customer to determine how much overhead they should provision for per link, a conservative 20% overhead was estimated to account for all possible encapsulation and encryption.

Table 1 provides maximum bandwidth consumption per second for different options and TelePresence units.

Note: The calculations in Table 1 are based on 500 Kbps for 5 fps auto-collaboration. The current Cisco TelePresence auto-collaborate feature uses 30 fps at 4000 kbps. The adjusted difference is represented in the area of the table that lists optional add-on features.

Example calculation:

Cisco TelePresence System 1000 running at 1080p-Best. Total audio and video = 4756 kbps. Replace 500 kbps 5 fps auto- collaborate with 4000 kbps for 30 fps auto-collaborate = 8256 kbps. Multiply by 1.2 to add 20% for Layer 2–4 overhead = 9907 kbps or 9.9 Mbps.

For more information about the auto-collaborate feature, please refer to http://www.cisco.com/go/telepresence.

Table 1. Maximum Bandwidth Consumption per Second

Maximum Bandwidth Consumption Per Second								
Resolution		1080p	1080p	1080p	720p	720p	720p	720p
Motion Handling		Best	Better	Good	Best	Better	Good	Lite
Video per Screen (kbps)		4000	3500	3000	2250	1500	1000	936
Audio per Microphone (kbps)		64	64	64	64	64	64	64
Auto Collaborate Video channel		500	500	500	500	500	500	100
Auto Collaborate Audio channel (kbps)		64	64	64	64	64	64	64
CTS-1000/CTS-500 Total Audio and Video (kbps)	Tx	4,628	4,128	3,628	2,878	2,128	1,628	1,164
	Rx	4,756	4,256	3,756	3,006	2,256	1,756	1,292
CTS-3000/CTS-3200 Total Audio and Video (kbps)		12,756	11,256	9,756	7,506	5,256	3,756	
+ 20% for Layer 2–4 overhead								
CTS-1000/CTS-500 *max* bandwidth (kbps) *includes Layer 2–4 overhead*	Tx	5,554	4,954	4,354	3,454	2,554	1,954	1,397
	Rx	5,707	5,107	4,507	3,607	2,707	2,107	1,550
CTS-3000/CTS-3200 *max* bandwidth (kbps) *includes Layer 2–4 overhead*		15,307	13,507	11,707	9,007	6,307	4,507	

Optional Add-On Features			
30fps Auto Collaborate (kbps)	3,500	+ 20% for Layer 2–4 overhead	4,200
CTRS Recording in CIF (kbps)	704	+ 20% for Layer 2–4 overhead	845
SD Interoperability Video Channel (kbps)	704	+ 20% for Layer 2–4 overhead	922
SD Interoperability Audio Channel	64	+ 20% for Layer 2–4 overhead	

Not Applicable to 720p Lite

Note: An average Cisco TelePresence System 3000 call consumes roughly 11 Mbps, whereas the theoretical maximum consumption is roughly 15 Mbps. Both numbers are averages over 1 second, but could be easily confused.

Note: 720p lite is a new video quality on Cisco TelePresence System endpoints, introduced as a part of the extended reach program. In this mode, 720p main video is further compressed: including 100 kbps auto-collaborate at 1 fps, the total bandwidth requirements turn out to be approximately 1.5 Mbps. Hence, a 2-Mbps uplink broadband connection or a single T1 connection could support it. However, 720p lite will interoperate only with Cisco TelePresence devices running Release 1.6 or later.

Latency

Minimal network latency contributes to the usability of the TelePresence call. A TelePresence call is a combination of voice-over-IP (VoIP), video-over-IP (VoD), and the optional incorporation of an AUX channel for sharing the computer screen of the participants. High latency impacts the usability of the call, not the audio fidelity or video image clarity, by increasing the likelihood that two participants on the call will both speak at the same time. As latency increases, jitter usually also increases. Typically, jitter is a function of latency. Because of the asymmetrical nature of the provisioned bandwidth (15M downlink/2M uplink), it is expected (and observed) that the latency from campus to remote location is lower than from the home office to the campus.

At the human experience level, latency is defined and measured as the time it takes for the speech or gestures of one individual (the speaker) to reach the ears and eyes of another (the listener), and for the audible or visual reaction of that listener to come all the way back to speaker so they can hear and see the listener's reaction. Hence, the human experience is round-trip in nature. This is referred to as "conversational latency" or "experience-level latency." 250ms–350ms is the threshold at which the human mind will begin to perceive latency—and become annoyed by it.

At the technical level, however, the latency in Cisco TelePresence is defined and measured as the time it takes for an audio or video packet containing speech or motion to travel from the Ethernet network interface of the speaker's TelePresence system to the Ethernet network interface of the listener's TelePresence system, in one direction. The listener's TelePresence system processes the incoming packets and computes a running average of the latency based on timestamps within the Real-Time Protocol (RTP) packets. Therefore, latency is measured only at the network level from one TelePresence system to another, not at the experience-level. It is measured unidirectionally by each TelePresence system, not measured round-trip, and does not take into account the processing time (encoding/decoding) of the packets.

Latency Target

In order to maintain acceptable experience-level latency, Cisco recommends that customers engineer their networks with a target of no more than 150ms of network-level latency, in each direction, between any two TelePresence systems. Given the circumference of the earth, the speed of light, and the cabling paths that light travels on between cities, it is not always possible to achieve 150ms between any two points on the globe. Therefore, Cisco TelePresence alerts the network administrator, and the user, when network-level latency exceeds acceptable levels.

Understanding Latency Measurements in Multipoint Meetings

As audio and video packets traverse a Cisco TelePresence Multipoint Switch, the RTP header containing the original timestamp information is overwritten and a new timestamp value is applied by the Cisco TelePresence Multipoint Switch. Therefore, the latency measured by each participating TelePresence system is only a measurement of the latency from the multipoint switch to that endpoint. Because of this, it is possible for the end-to-end latency from one TelePresence system, through the multipoint switch, to another TelePresence system to exceed the 250ms latency threshold, but the TelePresence system will not realize it. For example, if the latency from a TelePresence system in Hong Kong to a Cisco TelePresence Multipoint Switch in London is 125ms, and the latency from the multipoint switch in London to the other TelePresence system in San Francisco is 125ms, the end-to-end latency from the Ethernet network interface of the Hong Kong system to the Ethernet network interface of the San Francisco system is 250ms, plus approximately 10ms added by the multipoint switch, for a total of 260ms. The TelePresence system in San Francisco will not realize this, and will think that the latency for that meeting is only 125ms. Therefore, care should be taken when designing the network and the location of the multipoint switch to reduce the probability of this occurrence. The Cisco TelePresence Multipoint Switch is the only device in the network that is aware of the end-to-end latency between any two TelePresence systems in a multipoint meeting. Network administrators can view the end-to-end statistics (the sum of any two legs in that meeting) via the Cisco TelePresence Multipoint Switch administration interface.

Jitter

Jitter is the variation in latency between voice and video packets as they traverse the network. When jitter is excessive, it may lead to packet loss by the receiving decoder if the packet arrives outside the jitter buffer and is discarded because of late arrival. The packet is received but discarded (RxDisc). Generally, paths with low latency also exhibit low jitter.

Simply put, jitter is variation in network latency. In Cisco TelePresence, jitter is measured by comparing the arrival time of the current video frame to the expected arrival time of that frame based on a running clock of fixed 33ms intervals. Unlike most other videoconferencing and TelePresence products on the market, which use variable frame rate codecs, Cisco TelePresence operates at a consistent 30 frames per second (30fps). Therefore, the sending codec generates a video frame every 33ms, and the receiving codec expects those video frames to arrive every 33ms. Note: Cisco TelePresence jitter refers to application-layer jitter, which is the difference in arrival time between two adjacent picture updates. It is not equivalent to the network packet jitter.

Video frames vary in size based on how much motion is represented by a given video frame. When there is a low amount of motion within the encoded video, the video frame is relatively small. When there is a large amount of motion within the encoded video, the video frame is large. Given a constant end-to-end network latency and relatively constant packet sizes (e.g., 0ms of packet-level jitter), there will still inevitably be variation in the arrival times of video frames simply due to the variation in their size. This variation is primarily a function of the serialization rate (speed) of the network interfaces the packets constituting those video frames traverse. However, it can also be affected by queuing and shaping algorithms within the network routers along the path that may need to queue (buffer) the packets in order to prioritize them relative to other traffic, shape them prior to transmission, and then transmit (serialize) them on their outgoing interface.

Cisco TelePresence systems implement jitter buffers to manage these variations in video frame arrival times. Upon receipt at the destination, the packets are buffered until an adequate portion of the video frame has arrived, and then the packets are removed from the buffer and decoded. The size (depth) of the jitter buffer dictates how much jitter can be managed before it begins to be noticeable to the user. Packets exceeding the jitter buffer are dropped by the receiving codec because they arrived too late to be decoded.

The depth of the jitter buffer has an important consequence to the experience-level latency—every millisecond spent waiting in the jitter buffer increases the end-to-end latency between the users—so jitter buffers must be kept as small as reasonably possible to accommodate network-level jitter without adding an unacceptable amount of experience-level latency. When you start a call, the jitter buffer starts at 85ms. The jitter buffer dynamically adjusts itself—based on your network connection, if any time during the meeting jitter exceeds 125ms, the jitter buffer goes to 125 ms, 165ms, and 245ms. Also as a part of Release 1.6, a smaller maximum frame is used, which reduces the difference in serialization delay jitter. With Cisco TelePresence Release 1.6 and the Cisco Virtual Office TelePresence Extended Reach solution, a TelePresence system can run over a T1 link because now the deepest jitter buffer available is 245ms. The MFSR for a T1 line based on a 40 KB maximum frame size is 208ms, which can be accommodated using the deepest jitter buffer.

Jitter Target

In order to maintain acceptable experience-level latency, Cisco recommends that customers design their networks with a target of no more than 50ms of video frame jitter, in each direction, between any two TelePresence systems. Given the desire to deploy TelePresence over the smallest, and hence least expensive, amount of bandwidth possible, and the need in some circumstances to implement shaping within the routers along the path to conform to a service provider's contractual rates and policing enforcements, 50ms of jitter of video-frame-level jitter is not always possible. Cisco TelePresence will alert the network administrator when video-frame-level jitter exceeds acceptable levels.

Packet Loss

Loss is defined as packets that did not arrive (i.e., were dropped somewhere along the network path) and is measured by each TelePresence system by comparing the sequence numbers of the RTP packets it receives with the sequence numbers it expected to receive. Packet loss can occur anywhere along the path for a variety of reasons. The three most common reasons are:

- Layer 1 errors on the physical interfaces and cables along the path, such as a malfunctioning optical interface.
- Misconfigured network interfaces along the path, such as Ethernet speed/duplex mismatches between two devices.
- Bursts of packets exceeding the buffer (queue) limit or policer configurations on network interfaces along the path, such as Ethernet switches with insufficient queue depth or oversubscribed backplane architectures, or WAN router interfaces that police traffic to conform to a service provider's contractual rates.

A closely related metric is late packets, which are packets that arrived but exceeded the jitter buffer (i.e., arrived too late to be decoded) and hence were discarded (dropped) by the receiving TelePresence system. Lost packets and late packets are tracked independently by Cisco TelePresence Systems, but they both result in the same outcome: noticeable pixelization of the video.

Loss is by far the most stringent of the three metrics discussed. Latency can be annoying to the users but their meeting can still proceed, and jitter is invisible to the user, but loss (including packets that arrived but exceeded the 165ms jitter buffer and manifest into late packets) is immediately apparent. Consider the following calculation:

1080p resolution uncompressed (per screen)

2,073,600 pixels per frame

× 3 colors per pixel

× 1 byte (8 bits) per color

× 30 frames per second

= 1.5 Gbps uncompressed

The Cisco TelePresence System will use the H.264 codec to compress this down to 4 Mbps (per screen).
This represents a compression ratio of > 99%. Therefore, each packet is representative of a large amount of video data, and a very small amount of packet loss can be extremely damaging to the video quality. Cisco TelePresence implements a new technique known as Long-Term Reference Frames. This allows the system to recover from packet loss significantly faster by retransmitting the differences in the current frame relative to the original reference frame, instead of transmitting an entirely new reference frame. Long-Term Reference Frames are only implemented for point-to-point meetings between two TelePresence systems. They are not implemented in the Cisco TelePresence Multipoint Switch in this release, but are planned for a future release.

Loss Target
To maintain acceptable experience-level video quality, Cisco recommends that customers design their networks with a target of no more than 0.05% packet loss, in each direction, between any two TelePresence systems. This is an incredibly small amount, and given the complexity of today's global networks, 0.05% loss is not always possible to accomplish. Cisco TelePresence alerts the network administrator when packet loss (or late packets) exceeds acceptable levels.

In most deployments, packet loss is attributed to the following:

- **Queue drops**—Usually a result of a shaper or interface buffer, as a result of the available bandwidth on the teleworker router, but also can be due to congestion in the ISP.
- **Outages**—Packet loss is common when there is a failed link and the routing protocol must detect and install an alternate route in the routing table.
- Received but Discarded (RxDisc)RxDisc can come from an out-of-order packet or a packet that arrived too late. Most load-sharing techniques do not load-share on a per-packet basis, minimizing the occurrence of out-of-order packets. Late arrival is attributed to high jitter.
- **Faulty hardware or Interface misconfiguration**—In Ethernet topologies, a duplex mismatch (one side of the link is half duplex, while the other is full duplex) can exhibit high percentages of packet loss. Faulty or failing hardware can also contribute to packet loss.

Bursts

Network engineers are used to expressing bandwidth in “per-second” averages (e.g., 15 Mbps). Queuing, shaping, and policing algorithms in routers and switches, however, measure bandwidth on much smaller intervals of time (milliseconds). Cisco TelePresence never consumes more than its theoretical maximum over 1 second, but you have to analyze the traffic per millisecond to understand how policer and shaper algorithms treat it.

Routers and switches measure traffic in bytes over some time interval. The mean rate (R) is the amount of bytes a router expects to receive over some number of milliseconds (Tc) for a given bit rate.

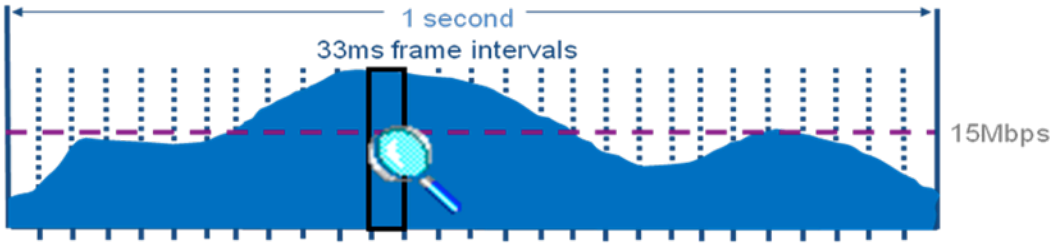
To obtain the mean rate in bytes per millisecond, use the following formula:

- Mbps x 1000 (converts Mbps to kbps) x 1000 (converts kbps to bps) ÷ 8 (converts bps to bytes/sec) ÷ 1000 (converts bytes/sec to bytes/ms)

Using a Cisco TelePresence System 3000 Release 1.2 @ 1080-Best (15.3 Mbps) as an example, the mean rate the router expects to receive is 1.913 KB per millisecond, or 63,129 bytes over 33ms.

Figure 2 illustrates bytes a single codec (per screen) can transmit on a per-millisecond basis.

Figure 2.



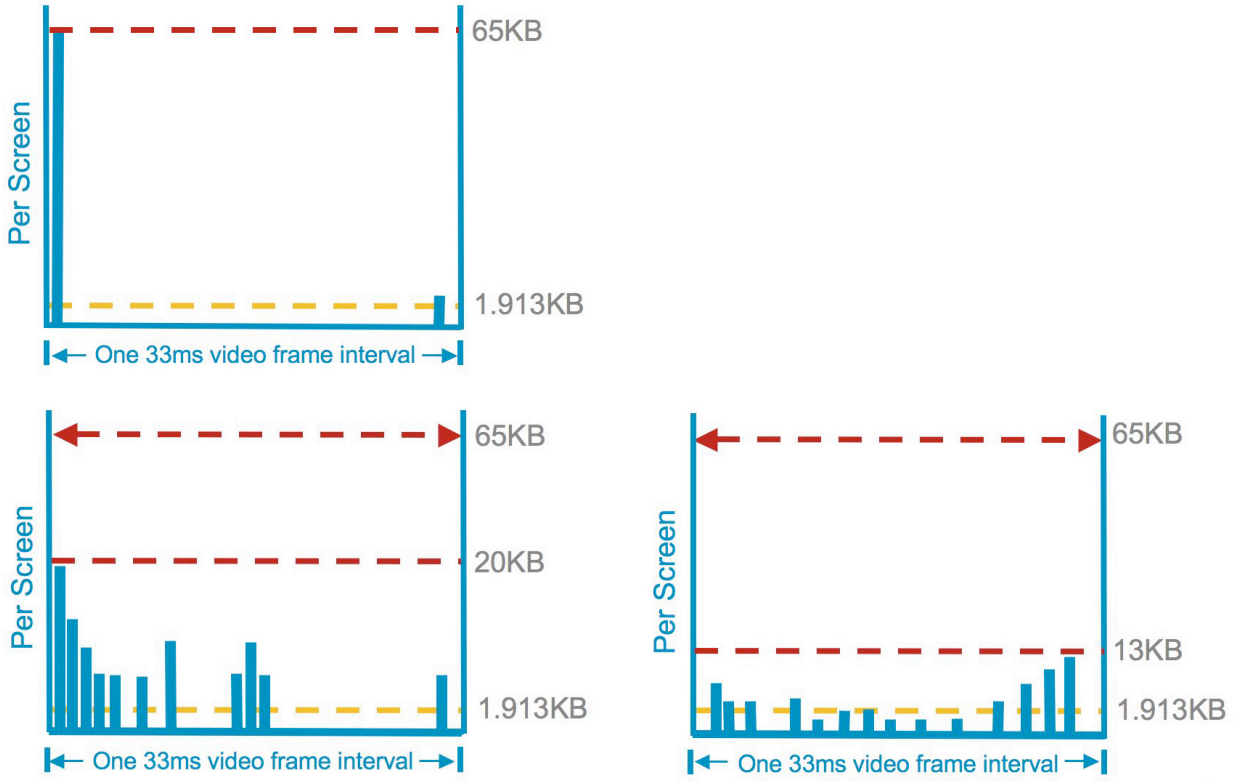
Resolution	1080p	1080p	1080p	720p	720p	720p	720p
Motion Handling	Best	Better	Good	Best	Better	Good	Lite
CTS-1000 *max* bandwidth over one second (Mbps)	5,553 TX 5,707 RX	4,953 TX 5,107 RX	4,353 TX 4,507 RX	4,353 TX 4,507 RX	3,153 TX 3,307 RX	1,953 TX 2,107 RX	1,397 TX 1,550 RX
CTS-3000 *max* bandwidth over one second (Mbps)	15,307	13,507	11,707	11,707	8,107	4,507	
CTS-1000 *mean* rate per millisecond the router expects (Bytes)	688 TX 713 RX	613 TX 638 RX	538 TX 563 RX	538 TX 563 RX	388 TX 413 RX	250 TX 263 RX	250 TX 263 RX
CTS-3000 *mean* rate per millisecond the router expects (Bytes)	1,913	1,688	1,463	1,463	1,013	563	

* Audio Traffic Not Included for Simplicity

As you will see in the next section, a Cisco TelePresence unit can generate much more than 1.913 KB per millisecond, even within the 33ms mean. In the graph in Figure 3, everything above the pink line(1.913KB/sec) is considered burst by the router.

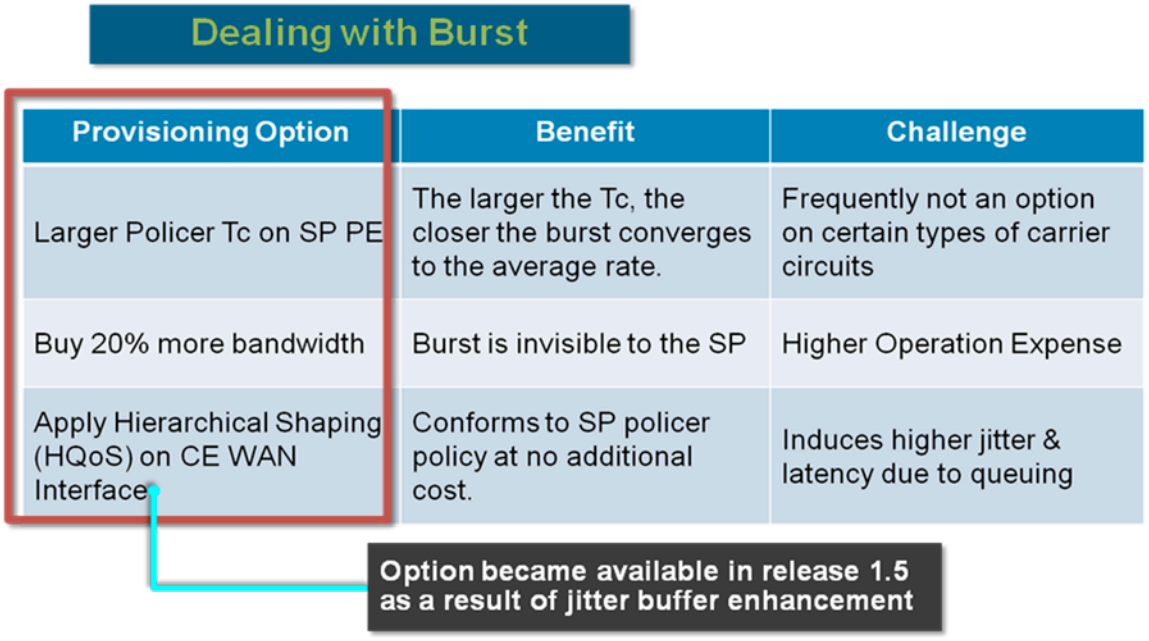
In all three of the graphs shown in Figure 3 below, a maximum of 65 KB is sent over each 33ms interval (average 16 KB over each 33ms interval). The difference is simply that the 65 KB is spread over the entire interval instead of being sent all at once, leaving lots of unused milliseconds of space in between. With Cisco Virtual Office and Release 1.6, we have been successful in achieving the effect in the third graph, wherein the traffic is spread over a period of 33ms, rather than sending it all at once, as shown in the first graph.

Figure 3. Graphs indicating traffic sent over 33ms frame interval



When choosing a teleworker solution, it is important to consider how that solution addresses bursts. Due to stringent traffic requirements, it is essential to choose a solution that uses hierarchical shaping to deal with bursts. Figure 4 shows Various options for dealing with burst.

Figure 4. Options for Dealing with Burst



Summary

Figure 5 provides a quick summary of Cisco TelePresence traffic characteristics. Figure 6 provides summary of latency, delay and jitter thresholds.

Figure 5. Summary- CTS Traffic Characteristics

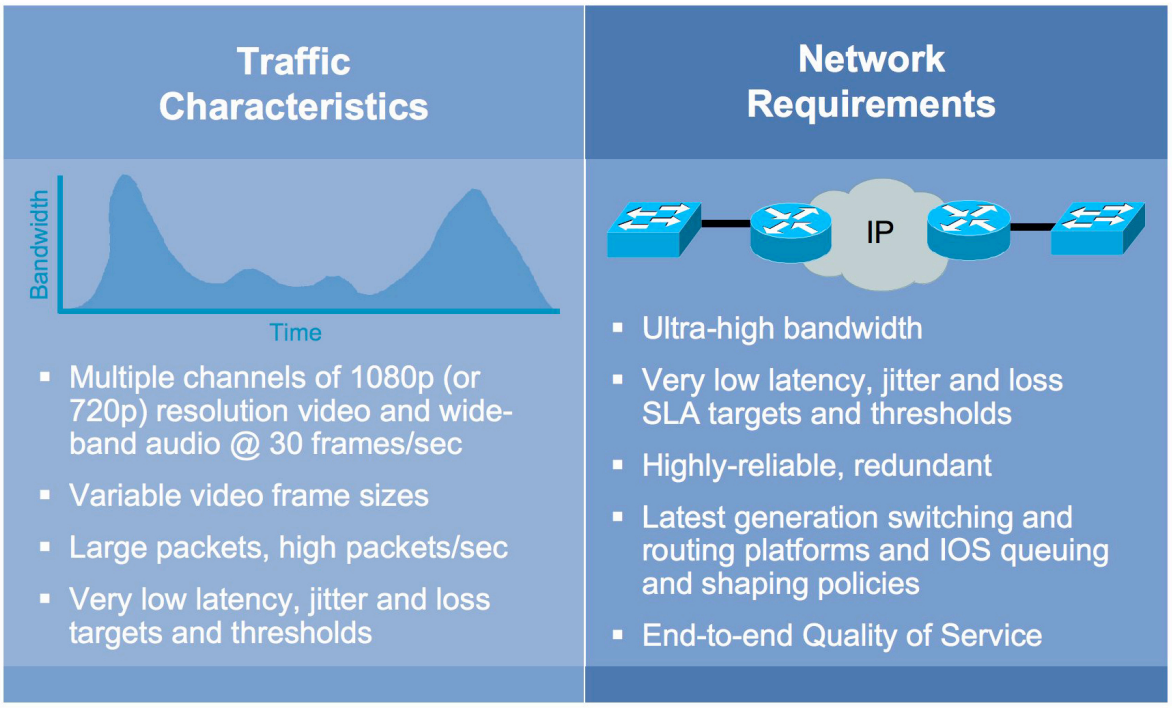
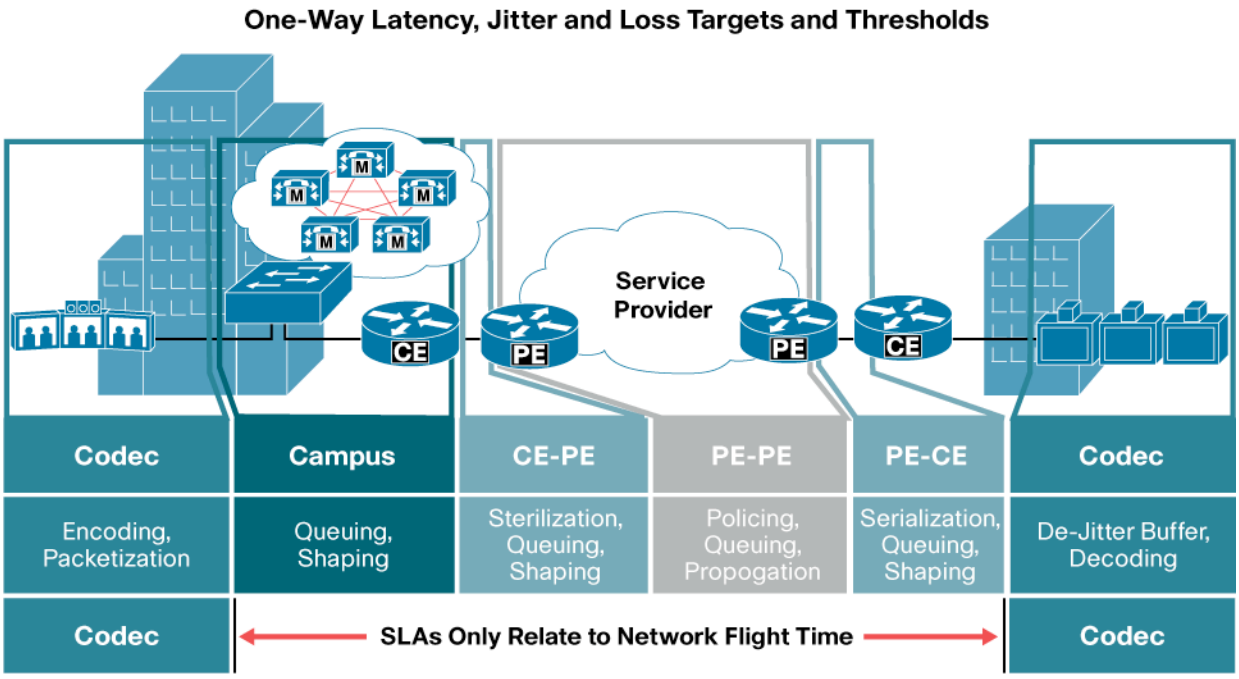


Figure 6. Latency, Jitter and Loss Thresholds



Metric	Target	Thresholds				Triggered Action on Threshold	
		1st	2nd	3rd	4th	1st	2nd
Latency	150ms	250ms	400ms			None	None
Jitter	50ms	85ms	125ms	165ms	245ms	None	None
Loss	0.05%	1%	10%			Network Bar Change	1. Reduce Quality 2. Drop Call

Spoke Router Recommendations

This section shows and discusses the configuration excerpts and output of the QoS and interface statistics of the remote teleworker router. The remote teleworker router and headend routers are based on a Cisco Virtual Office deployment. For more information, refer to http://www.cisco.com/web/go/cvo.

The following QoS-related show commands include overhead associated with encrypted tunnels. The DMVPN tunnels are protected with a transform set using 3DES encryption and SHA hash.

```
#show cry ipsec sa | inc interface|settings|transform
interface: Tunnel100
transform:  esp-3des  esp-sha-hmac  ,
in use settings ={Transport UDP-Encaps, }
interface: Tunnel200
transform:  esp-3des  esp-sha-hmac  ,
in  use  settings  ={Transport  UDP-Encaps,  }
```

If the teleworker router is deployed behind a NAT device, NAT Traversal (NAT-T) is negotiated and is evident from the specification of UDP-Encaps in the above display. Because the QoS policy references Differentiated Services Code Point (DSCP) values to classify packets, and the encryption encapsulation process copies the ToS byte of the plain text packet to the encrypted packet header, QoS pre-classify is not required on the tunnel interfaces. Based on your own corporate design, however, QoS pre-classify may be required.

Encryption Overhead

The encryption and encapsulation of the TelePresence calls in DMVPN tunnels adds overhead to the plain text packets. In deployment testing, Triple DES (3DES) and SHA are configured. Many deployments are interested in using Advanced Encryption Standard (AES) or Rijndael instead of 3DES because AES supports 128-, 192-, and 256-bit keys while 3DES supports a 168-bit key with an effective length of 112 bits. AES is an encryption standard adopted by the U.S. government.

This illustration assumes DMVPN (GRE Transport mode) ESP AES-128/SHA or 3DES/SHA with NAT-T. AES adds slightly more overhead than 3DES. AES uses an Initialization Vector (IV) that is 16 bytes while 3DES uses an V of 8 bytes. This is a factor of the block size, AES uses a 16-byte block where 3DES uses an 8-byte block. Either encryption algorithm encrypts plain text on a block-by-block basis. If the plain text is not an even multiple of the block size, padding must be added to the plain text to fill the last block. Assuming a plain text packet of 1024 bytes, (and therefore no padding of the last block), the resulting encrypted packet would be 1090 bytes or an approximately 6 to 7 percent increase in packet size. The average packet size of a TelePresence call is over 1024 bytes per packet.

QoS Configuration

In this section, the QoS configuration of the remote teleworker router is shown. The DSCP values from plain text packets are copied to the IP header of the encrypted packet as part of the encapsulation process.

- Queuing policies, by default, do not engage at sub-line rates.
- To ensure that transmission rates do not exceed the contracted rate, a shaper must be used.
- Cisco IOS allows for Hierarchical QoS (HQoS) policies. One QoS policy may be “nested” within another; thus, a queuing policy may be nested within a shaping policy, in which case packets are prioritized within a sub-line (shaped) rate.
- As with policers, Cisco IOS shapers operate on a token-bucket principle using the formula:
  - Burst (Bc) = Shaped Rate * Shaping Time Interval (Tc)
- Testing has shown the optimal shaping interval for TelePresence to be 20ms.
- The shaping interval (Tc) cannot be explicitly set within Cisco IOS. It is instead calculated from the value for Burst (Bc), as follows:
  - Burst (Bc) = Shaped Rate * Shaping Time Interval (Tc)

Class Map

The Cisco Medianet Application Classes DiffServ QoS Recommendations (RFC 4594-based) is configured on the remote router. To accommodate Cisco Unified Video Advantage (formerly known as Cisco VT Advantage), AF41 has been included in the voice class as well as CS4 for TelePresence.

```
!
class-map  match-any   TELEPRESENCE
  match  ip  dscp  cs4
class-map  match-any   LOW-LATENCY-DATA
  match  ip  dscp  af21 af22  af23
class-map  match-any   HIGH-THROUGHPUT-DATA
  match  ip  dscp  af11 af12  af13
class-map  match-any   BROADCAST-VIDEO
  match  ip  dscp  cs5
class-map  match-any   NETWORK-CONTROL
  match  ip  dscp  cs6
```

```
class-map  match-any   MULTIMEDIA-CONFERENCING
  match  ip  dscp  af41 af42   af43
class-map  match-any   OAM
  match  ip  dscp  cs2
class-map  match-any   VOICE
  match  ip  dscp  ef
  match  ip  dscp  af41
  match  ip  dscp  cs4
class-map match
  any SCAVENGER
  match ip dscp
  cs1
class-map match-any
  CALL-SIGNALING match
  ip dscp cs3
class-map match-any
  MULTIMEDIA-STREAMING match
  ip dscp af31 af32 af33
```

Note: Class-default is the 12th class and does not require an explicitly defined class-map statement. This QoS policy assumes the end user will not be on a VoIP call and a Cisco TelePresence System 500 call simultaneously.

Policy Map

In testing the TelePresence network, traffic is shown configured in either a bandwidth queue, a priority queue, or a low-latency queue. Any configuration is acceptable and will produce a quality video experience.

The Committed Information Rate (CIR) value used in testing is a fraction of the provisioned and measured bandwidth. The service provider advertises a 2-Mbps uplink. The measured actual throughput of the link is approximately 1.9 Mbps. The CIR value is configured at approximately 95 percent (1.8 Mbps) of the measured uplink throughput. Configuring a shaper CIR value at 95 percent of the measured actual throughput is consistent with the Enterprise Class Teleworker[6] design guide. The goal is to shape at a rate that will minimize indiscriminate drops by input policers of the service provider's broadband aggregation switches/routers.

The tested and configured shaper measurement intervals (Bc) ranged from 4ms to 25ms, all producing acceptable results.

```
shape  average  cir  [bc]  [be]
shape  average  1800000  45000       45000  25ms
shape     average      1800000  18000      18000  10ms   <----  Value  illustrated
in  config  samples
shape     average      1800000      7200   7200   4ms    <---  IOS  default  value
```

If the configuration command does not specify Bc and Be, the default value is a 4ms measurement interval; therefore, the Bc will be CIR * (4/1000).

Bandwidth Queue

The policy-map configuration using the bandwidth queue is shown below:

```
!
policy-map CVO-teleworker
class VOICE

bandwidth percent 85
class CALL-SIGNALING

bandwidth percent 2
class NETWORK-CONTROL

bandwidth percent 5
class class-default

fair-queue

random-detect dscp-based
policy-map Shaper

class  class-default

shape average 1800000 18000 18000 queue-limit 1024 packets

service-policy  CVO-teleworker
```

Priority (LLQ) Queue

The policy-map configuration using the priority queue is shown below:

```
!
policy-map CVO-teleworker
description CVO-teleworker

class VOICE

priority percent 85

class CALL-SIGNALING

bandwidth percent 2

class NETWORK-CONTROL

bandwidth percent 5

class class-default
```

```
fair-queue

random-detect dscp-based

policy-map Shaper

class  class-default

shape average 1800000 18000 18000 queue-limit 1024 packets

service-policy  CVO-teleworker

!
```

WAN Interface

The service policy is applied to the WAN (outside) interface of the teleworker router. The max-reserved-bandwidth default value of 75 percent must be increased to 100 percent (as shown below) because of the amount of bandwidth required by the TelePresence call.

```
!
Interface FastEthernet4 description Outside
ip  address  dhcp
ip access-group INPUT_ACL in
ip nat outside
ip inspect CBAC out
ip virtual-reassembly

max-reserved-bandwidth 100
service-policy output Shaper
end
```

Deployment Results

In this section, the output from the show policy-map interface command is included to demonstrate the effectiveness of the QoS service policy. As a baseline, a initial test is run with only the TelePresence call active from the teleworker router. A capture of this traffic from the LAN interface perspective is shown in Figure 7.

Figure 7. Baseline Test

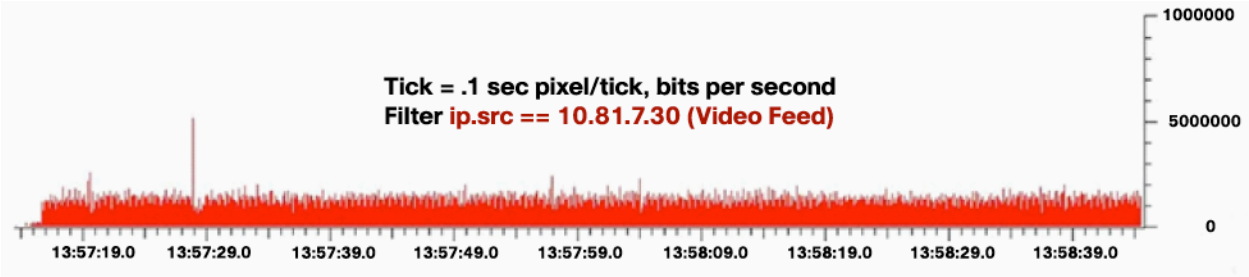


Figure 2 illustrates that the voice, video, and AUX port (slide sharing) traffic from the Cisco TelePresence System 500 is approximately 1.5 Mbps of sustained traffic. There is an occasional burst in traffic. Please refer to the appendix to see a scenario where the CISCO TELEPRESENCE feed is stopped during an HTTP transaction, as well as its behavior when the traffic is reinitiated.

Bandwidth Queue Configuration

In this example, the voice and video traffic is configured in a bandwidth queue of 85 percent of the 1.8 Mbps shaper with a measurement interval of 10ms. The interface statistics show that the combined throughput is approaching the 1.8 Mbps value. Recall that the output interface statistics as well as the policy-map values include the encryption overhead.

```
#show interfaces fastEthernet 4 | include rate
  30 second input rate 1323000 bits/sec, 225 packets/sec
  30 second output rate 1799000 bits/sec, 303 packets/sec
```

The parent shaper policy-map shows the shaper active by the presence of packets currently queued. Note that the TelePresence network traffic is matching DSCP AF41 rather than CS4. The marking of CS4 is recommended by the Cisco Medianet Application Classes DiffServ QoS Recommendation. However, in this implementation, the network manager has decided to mark both TelePresence and Cisco Unified Video Advantage as AF41.

```
#show policy-map interface fastEthernet 4

FastEthernet4

Service-policy output: Shaper

Class-map: class-default (match-any) 96168 packets, 95948483 bytes

30 second offered rate 1799000 bps, drop rate 3000 bps Match: any

Queuing

queue limit 1024 packets

(queue depth/total drops/no-buffer drops) 26/322/0 (pkts output/bytes output)
95846/95554955

shape (average) cir 1800000, bc 18000, be 18000 target shape rate 1800000

Service-policy : CVO-teleworker

Class-map: VOICE (match-any) 47749 packets, 32302766 bytes

30 second offered rate 1437000 bps, drop rate 0 bps Match: ip dscp ef (46)

0 packets, 0 bytes
```

```
30 second rate 0 bps Match: ip dscp af41 (34)

47749 packets, 32302766 bytes

30 second rate 1437000 bps Match: ip dscp cs4 (32)

0 packets, 0 bytes

30 second rate 0 bps Queuing

queue limit 64 packets

(queue depth/total drops/no-buffer drops) 1/20/0 (pkts output/bytes output)
47730/32286500 bandwidth 85% (1530 kbps)

Class-map: CALL-SIGNALING (match-any))

[omitted]

Class-map: NETWORK-CONTROL (match-any)

[omitted]

Class-map: class-default (match-any)
47401 packets, 63490483 bytes
30 second offered rate 360000 bps, drop rate 3000 bps
Match: any
Queuing
queue limit 64 packets
(queue depth/total drops/no-buffer drops/flowdrops) 23/302/0/0
(pkts output/bytes output) 47101/63103185
Fair-queue: per-flow queue limit 16
Exp-weight-constant: 9 (1/512)
Mean queue depth: 19 packets
dscp      Transmitted  Random drop Tail/Flow drop  Minimum  Maximum
Mark
   pkts/bytes   pkts/bytes   pkts/bytes   thresh thresh prob
default   54020/68936025      329/428374  0/0  20 40     1/10
cs5 561/79398     0/0  0/0  30 40     1/10
```

In class-default, Weighted RED (WRED) is enabled and the best-effort class drops are random rather than tail drops.

Priority Queue Configuration

The priority queue configuration demonstrates results that are similar to the previous example.

```
#show policy-map interface fastEthernet 4

FastEthernet4

Service-policy output: Shaper

Class-map: class-default (match-any) 55226 packets, 48774960 bytes

30 second offered rate 1788000 bps, drop rate 3000 bps Match: any

Queuing

queue limit 1024 packets

(queue depth/total drops/no-buffer drops) 28/107/0 (pkts output/bytes output)
55068/48592668

shape (average) cir 1800000, bc 18000, be 18000 target shape rate 1800000

Service-policy : CVO-teleworker

queue stats for all priority classes: Queuing


queue limit 64 packets

(queue depth/total drops/no-buffer drops) 0/51/0 (pkts output/bytes output)
36690/24813756

Class-map: VOICE (match-any) 36741 packets, 24868798 bytes

30 second offered rate 1330000 bps, drop rate 0 bps Match: ip dscp ef (46)

0 packets, 0 bytes

30 second rate 0 bps Match: ip dscp af41 (34)
36741 packets, 24868798 bytes

30 second rate 1330000 bps Match: ip dscp cs4 (32)

0 packets, 0 bytes

30 second rate 0 bps
```

```
Priority: 85% (1530 kbps), burst bytes 38250, b/w exceed drops: 51
```

The remaining output display is eliminated for brevity.

TelePresence and Auth-Proxy

A remote-office network may not be physically as secure as the corporate environment, meaning that guests and family members may also have access to the devices connected to the spoke router. Authentication Proxy provides a way to identify legitimate users and limit access to the corporate network to only those users. Auth-proxy can be used to provide role-based access permissions to the users. All access to the corporate network is denied by an inbound access control list (ACL) applied on the inside interface of the router. To initiate the authentication process, the user will have to first access a corporate website using a web browser. This access will be intercepted by the router and will be replaced with a web-based user authentication prompt. The user will be allowed to have access to the corporate site only if correct credentials are provided. The credentials are verified by an AAA server. Upon verification of the credentials, appropriate permit access control entries (ACEs) are downloaded and applied to the auth-proxy inbound ACL on the spoke router. It is possible to download a "permit ip any any" for all users or to download specific ACEs based on the group to which the user belongs. This way the network administrator can implement role-based access control.

Cisco TelePresence considerations:

- Cisco TelePresence Systems cannot display the Authentication Proxy prompt, so they cannot be authenticated using auth-proxy. One solution to this problem is to use Context-Based Access Control (CBAC). Configuration download protocols include those needed by the Cisco TelePresence System codecs as well as their attached IP phones to upgrade system load images and download device configuration files. IP phones usually download their initial configuration using Trivial File Transfer Protocol (TFTP). Unlike IP phones, TelePresence codecs utilize HTTP over TCP port 6970 to download system images and configuration files. In that case, TFTP and TCP port 6970 need to be opened in the auth-proxy inbound ACL.
- **SIP Registration Protocol**: Once the Cisco TelePresence System codec(s) and associated IP phone have completed downloading their configuration files and possibly upgrading their system images, both perform a SIP registration with the Cisco Unified Communications Manager cluster. SIP signaling uses either TCP or UDP port 5060. However, the connection-oriented nature of TCP makes it preferred for TelePresence deployments. These ports need to be opened on the auth-proxy inbound ACL. IP inspection dynamically opens holes for RTP streams when a phone call is made. By opening only UDP 5060 and 5061, the IP phone and codec work without doing any authentication.

  (Also note that for management of the end Cisco TelePresence System units, HTTP, HTTPS, and SSH could be used and correspondingly configured.)
- For details on configuration of auth-proxy, visit http://www.cisco.com/en/US/prod/collateral/iosswrel/ps6537/ps6586/ps6660/prod_white_paper0900aecd8046cbc4.html

TelePresence and dot1x

When using port-based IEEE 802.1x, all IP devices connecting to the router's switchports must have dot1x supplicants and must present valid credentials before getting an IP address, and thus access to the network. Once authenticated, the device gets network access. If the validation fails, the port is shut down.

Cisco IP phones can request a voice VLAN. If a voice VLAN is enabled on the router, the Cisco IP phone is automatically placed in that VLAN, and bypasses 802.1 x authentications. However, the current Cisco TelePresence System codec cannot be placed automatically in a voice VLAN, and will not be bypassed with dot1x enabled natively. The workaround is to either disable dot1x on those spokes that are planning to have a Cisco TelePresence System codec, or to enable dot1x using MAC authentication bypass just like the way third-party phones are bypassed.

For more information on configuration of dot1x, please visit http://www.cisco.com/go/cvo.

Spoke to Spoke

DMVPN traffic from one spoke to another can be sent via the hub or directly between the spokes using spoke-to-spoke configuration. Spoke-to-spoke with extended reach is recommended if the Cisco TelePresence System calls need to be placed in a zone serviced by a single service provider. However, when you need to go from one site to another served by a different service provider, it is recommended to send Cisco TelePresence System traffic from one spoke to another spoke via the hub. Moreover, spoke-to-spoke tunnels can be used only for point-to-point calls. In case of multipoint calls, all Cisco TelePresence System traffic should be sent to the Cisco TelePresence Multipoint Switch located at the central site, and traffic should go over the spoke1 → hub → spoke2 route. For more information on configuration of DMVPN, please refer to http://www.cisco.com/go/cvo.

HUB Recommendations: Per-Tunnel QoS Deployment

This section provides detailed design and implementation information for deployment of per-tunnel QoS features with Cisco Virtual Office. This section assumes basic knowledge about the Cisco Virtual Office deployment solution and basic QoS features and concepts. Please refer to the Cisco Virtual Office overview (http://www.cisco.com/go/cvo) for more information about the solution, its architecture, and all of its components.

Per-Tunnel QoS

Per-tunnel QoS allows traffic from a hub to spokes to be regulated on a per-spoke basis. It allows QoS to be configured on tunnel interfaces used for DMVPN and IPsec, while previously QoS was restricted to physical interface support.

Per-tunnel QoS solves two problems commonly found in hub-and-spoke topologies:

1. Prevents lower-end spoke routers from being overrun by higher-end hubs
2. Prevents some spokes from hogging hub resources and starving other spokes.

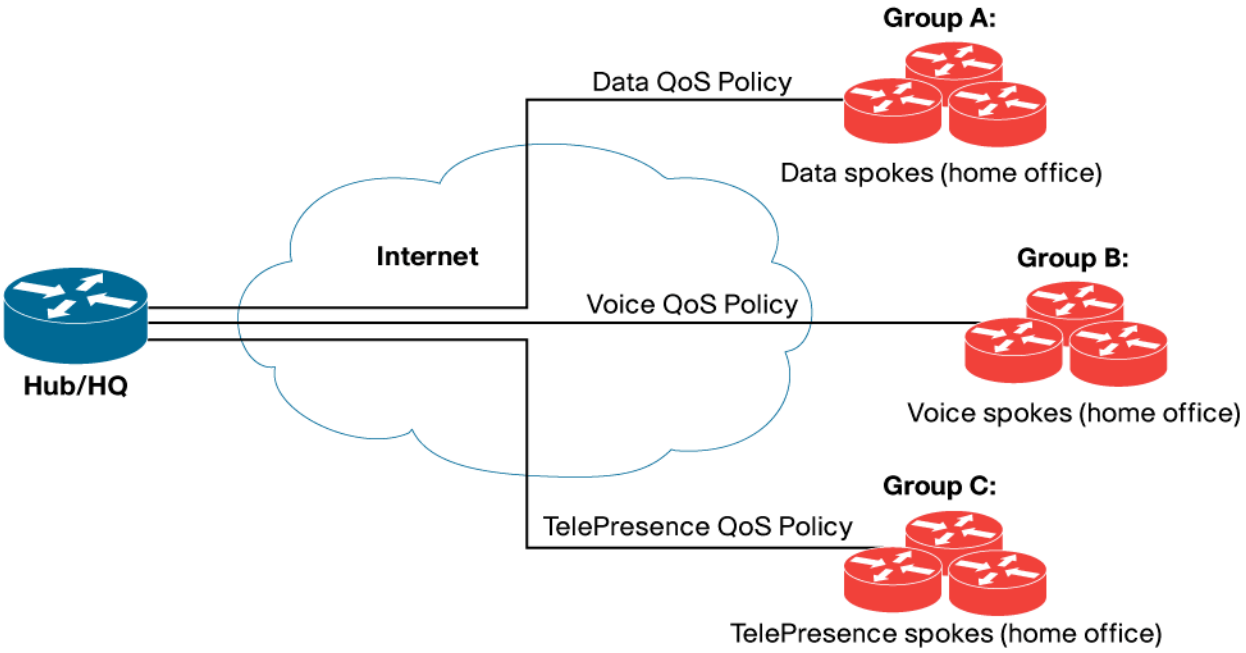
Case 1 occurs in many DMVPN networks, as the hub is usually a high-end router, while spokes are often lower-end routers. In this case, if the link between the hub and the spoke is unregulated, the hub may send traffic at a higher rate than the spoke can handle, causing congestion and packet loss. In Case 2, one spoke may be performing traffic-intensive operations, such as VoIP calls. Since that spoke is requesting a lot of traffic, it may overrun the link between the hub and itself, not allowing traffic to flow from the hub to other spokes.

Previously, QoS could be configured on spokes, but in these instances, the hub isn't aware of the policies and cannot prevent the problem indicated in Case 2 above. Alternatively, past configurations of QoS on the hub cannot adjust for multiple classes of service for the spokes.

Per-tunnel QoS is most advantageous when different levels of service are desired for separate subsets of spokes, or when the spoke routers themselves vary in capability. One of the first design considerations is to divide the spokes into groups, which will have different QoS policies.

In the Cisco Virtual Office configuration, there are three different class levels to which spokes belong: data, voice, and TelePresence. Figure 8 shows the Cisco Virtual Office hub-and-spoke topology and the spoke groups.

Figure 8. Cisco Virtual Office DMVPN Hub-and-Spoke Topology and Per-Tunnel QoS Setup



In Figure 8, Group A comprises spokes with data traffic being the primary traffic flowing from hub to spoke. Users in Group B make frequent calls and so require a QoS policy that can guarantee good voice and video quality for the calls, while users in Group C are executives who use Cisco TelePresence and need guaranteed bandwidth to ensure adequate TelePresence quality. With the deployment of per-tunnel QoS, different QoS policies can be applied to each tunnel depending on which group the spoke belongs to, ensuring that users in each group receive adequate bandwidth for their applications without overstressing the hub or their spokes.

Per-Tunnel QoS Configuration

The bulk of per-tunnel QoS configuration occurs on the hub, which uses Cisco Policy Language to set up a hierarchy of class-maps and policy-maps. On the DMVPN hub interface, these policy mappings are tied together through the use of a nhrp-group command. Similarly, on the spoke side, the nhrp-group command specifying the policy name is configured on the spoke router DMVPN interface. The hub receives the nhrp-group string from the spoke in the periodic NHRP registration requests. The nhrp-group string is then mapped to the QoS policy defined on the hub, and the policy is applied to the tunnel from the hub to the spoke.

The following is a sample Cisco Virtual Office configuration for per-tunnel QoS. Based on your own corporate policies, some parts of the configuration may need to be redefined.

```
!!!   Hub configuration !!!
! The QoS configuration on the hub side follows a parent-child hierarchy in which a
! child policy is defined and then can be called by the parent policy.
! This configuration creates 3 classes to support data; voice and video;
! and Telepresence QoS
```

```
! Define parent class 'data,' which provides default QoS policy for spokes
requiring
! just data traffic
policy-map data
 class class-default
    shape average 1000000

! Configure the default behavior.  Parent classes of 'data,' 'voice,' and
! 'Telepresence' will all call the default class as it is assumed data traffic will
be
! common to all the groups.
class class-default
    fair-queue
     random-detect

! Define parent class 'voice,' which provides voice and video QoS policy
policy-map voice
 class class-default
    shape average 2000000
  service-policy voice_and_video

! Child policy 'voice_and_video' is called by parent policy 'voice' above
policy-map voice_and_video
 class voice
    priority 384
 class class-default
    fair-queue
     random-detect

! Define parent class 'tp,' which provides QoS policy for spokes needing
Telepresence
! support
policy-map tp
 class TelePresence
    priority 10000
 class class-default
    fair-queue
     random-detect

! Child class 'TelePresence' is called by parent policy 'tp' above
class TelePresence
    priority 500000
 class class-default
    fair-queue
     random-detect

! Policies are attached to the DMVPN tunnel interface using the 'nhrp map group'
! command
```

```
interface Tunnel300
 ip nhrp map group persa_data service-policy output data
 ip nhrp map group persa_voice service-policy output voice
 ip nhrp map group persa_tp service-policy output tp


!!! Spoke configuration !!!

! Spokes requiring data-only QoS policy should configure the following on their
DMVPN
! tunnel interface
interface Tunnel300
 ip nhrp group data

! Spokes requiring voice and video QoS policies should configure the following on
! their DMVPN tunnel interface
interface Tunnel300
 ip nhrp group voice

! Spokes requiring Telepresence QoS policy should configure the following on their
! DMVPN tunnel interface
interface Tunnel300
 ip nhrp group tp
```

Summary

In summary, a Cisco TelePresence System 500 using 720p Lite can effectively be implemented on a business-class broadband connection when provisioned with a recommended minimum of 2 Mbps uplink with Cisco Virtual Office. The 720p Lite configuration requires approximately 1.5 Mbps with the overhead of DMVPN/IPSec using 3DES/SHA in transport mode.

As a best practice, the broadband connection should be speed-tested and the shaper should be configured with an average CIR value of approximately 95 percent of the measured bandwidth. In testing, a shaper of 1.8 Mbps was used on a service advertised as 2.0 Mbps and measured at 1.9 Mbps.

It is important to use network management tools such as the CiscoWorks Internetwork Performance Monitor and IP SLA to measure the network performance on an ongoing basis to verify that the service provider continues to offer the subscribed bandwidth. As a guideline, for an effective user experience, latency (one-way) should be less than 50 ms, jitter less than 10 ms, and packet loss less than 1/2 of the 1 percent.

It must be emphasized that the service level of broadband providers and the Internet in general will vary by time of day. Usage usually increases mid-morning to early evening. Additionally, in periods of area-wide disasters, which increase telecommuting load, expect that the service levels may be adversely impacted.

References

- **Cisco TelePresence System 500 data sheet**: http://www.cisco.com/en/US/prod/collateral/ps7060/ps8329/ps8330/ps9599/data_sheet_c78-46851_7.html
- **Cisco Virtual Office**: http://www.cisco.com/go/cvo
- **Cisco Unified Communications—Poor Voice Quality**: http://docwiki.cisco.com/wiki/Cisco_Unified_Communications_--_Poor_Voice_Quality
- **Cisco IOS Quality of Service Solutions Command Reference, Release 12.2**: http://www.cisco.com/en/US/docs/ios/12_2/qos/command/reference/qrfcmd9.html#wp1077189
- **Cisco TelePresence**: http://www.cisco.com/go/telepresence

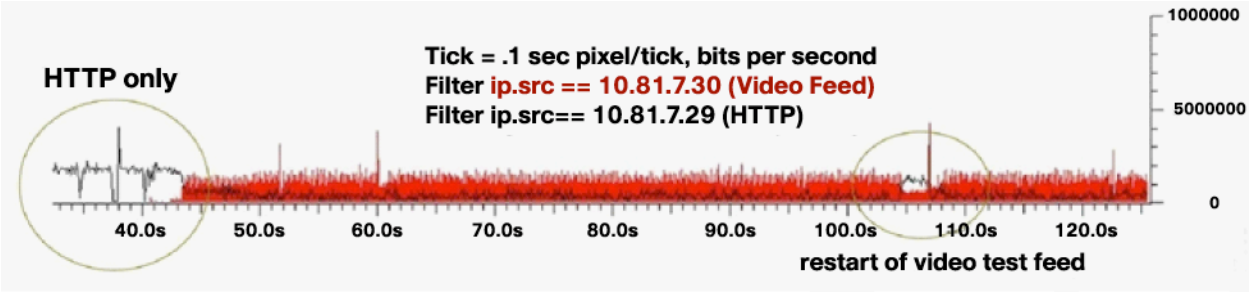
Appendix A

Pilot Testing Results

One observation from pilot testing and the review of output queue drops on the teleworker router and I/O graphs from packet captures is that most of the packet loss occurs during the first seconds of a call.

In testing, network traffic on the local LAN segment is captured. The TelePresence conference (720p Lite) is a pre-recorded video feed substituted as input to the encoder rather than usual camera input. Slide sharing by way of the AUX port is enabled with a slide change every 10 seconds. Also on the LAN segment is a PC using HTTP to upload a file.

Figure 9. I/O Graph of Voice, Video, and Data



QoS is configured on the uplink (CBWFQ) with the shaper on uplink at 1,800,000 bps with interval at 10ms. Figure 9 illustrates the how effective QoS is at managing the video feed over the data traffic. When the video feed is not present on the network, the bandwidth consumed by the PC approaches the capacity limit of the shaper. When the video feed is initiated, the HTTP application throughput decreases accordingly.

In Figure 9, at the point labeled “restart of video test feed,” the video input is solid black while the recorded video feed restarts its loop. During that period, the HTTP throughput increases because the video bandwidth consumption decreases. When the video loop begins again, a burst of traffic is encountered. Some packets may be lost during this state change. By modifying the target shaped rate, queue size, and Committed Burst size (Bc) / Excess Burst size (Be), this loss could be avoided. However, bandwidth for the teleworker is a finite resource and attempting to eliminate all video packet loss may not be economically feasible or practical. The goal is to enable the use of TelePresence where bandwidth is limited by distance and cost. The QoS configuration parameters shown in the above sections are a reasonable balance between quality and accessibility.

Americas Headquarters
Cisco Systems, Inc.
San Jose, CA

Asia Pacific Headquarters
Cisco Systems (USA) Pte. Ltd.
Singapore

Europe Headquarters
Cisco Systems International BV
Amsterdam, The Netherlands

Cisco has more than 200 offices worldwide. Addresses, phone numbers, and fax numbers are listed on the Cisco Website at **www.cisco.com/go/offices.**

Printed in USA C07-581216-00 01/10