Supporting Real-time Traffic

Preparing Your IP Network for Video Conferencing

White Paper

Global Services

January 5, 2006

POLYCOM®

Table of Contents

1.0 Overview

Voice over IP (VoIP) and video conferencing over IP create real-time traffic streams with characteristics very different from run-of-the-mill data applications. If you are considering running voice or video conferencing over your IP network, you should prepare for this different and often challenging traffic type. The fact is that most enterprise networks are ill-equipped to carry video conferencing traffic, and will need to be evaluated, tested, and possibly reconfigured and upgraded to ensure acceptable quality.

Our goal is to identify the challenges posed by voice and video conferencing traffic, and provide a blueprint to prepare your enterprise network to support them. In this paper we recommend techniques for handling bandwidth, packet loss, jitter, latency and QoS implementation, as well as testing methodologies for network verification and ongoing network monitoring. Lastly, we describe an approach for specifying and managing the demands video conferencing places on IP network infrastructure.

An understanding of IP network design and deployment is helpful in understanding this guide, as is a general knowledge of IP network deployment (switching, routing, bandwidths, error mechanisms, etc.). Note that this paper does not address issues surrounding passing voice and video conferencing traffic through Network Address Translating Routers (NATs), or the issues associated with firewalls.

2.0 Real Time Traffic

Real-time traffic supports real-time interactive applications, the most prominent of which are voice and video conferencing. Both of these have users at each end of a connection who expect that what they say or do will be transmitted ‘instantly’ to the other end of the connection, and the conversation will proceed as if the two parties were in the same room. Some of the most difficult aspects of real-time traffic come from this need for speed. We will see later how aspects of a normal data network sometimes interfere with this requirement.

When we refer to real-time traffic in this document, it applies to voice over IP (VoIP) traffic, as well as to both the video and audio streams of a video conferencing application.

2.1 How is Real-Time Traffic Different?

The Internet Protocol (IP) is at the heart of all modern networks, and is responsible for connecting one endpoint to another. But by design, IP is an unreliable protocol. This means it is not designed to insure that all packets sent from one endpoint arrive successfully at the other endpoint.

Data applications require a reliable transport, one that will insure all the bits of the data being sent arrive successfully and correctly at the destination computer. To insure this result, our networks use Transmission Control Protocol (TCP) on top of IP (TCP/IP). TCP insures each packet that is sent arrives at the other end, and will send a packet again if one is lost. TCP then verifies all the data is correct before delivering it to the application.

Data applications tend to be very bursty in their utilization of the network bandwidth. When a file or block of data is ready to be transferred across the network, the sender wants to send the data as quickly as possible, and then move on to other tasks. The result is very short bursts of activity on the network, followed by periods of relatively low or no use of the network.

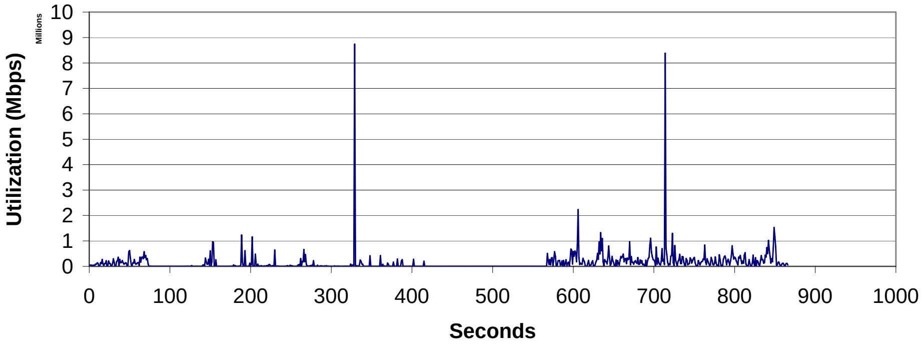

Figure 1 - Typical Data Bandwidth Utilization

Figure 1 is a graph of a typical data application, showing periods of low and high network utilization.

IP-based networks often experience packet loss, and this is a normal part of the network operations. In fact the TCP protocol uses packet loss as its flow control mechanism. TCP determines how quickly to send data by increasing its send rate until packet loss occurs, and then backing off. This occurs over and over for each TCP stream in the network.

Real-time traffic has very different characteristics. Real-time traffic results from a codec which is sampling a continuous real-world environment (speech or images), and transmitting constant updates of this information to reproduce a visual or auditory result. So the bandwidth utilization of voice and video is constant during the time the application is running. Figure 2 is a graph of a 384K video conference, showing both the audio and video streams, and their relatively constant use of bandwidth during operation.

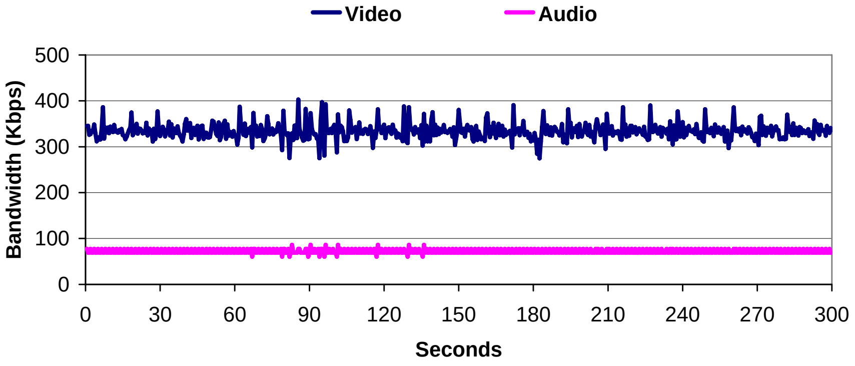

Figure 2 - Video Conferencing Bandwidth Utilization

A second characteristic of real-time streams is their sensitivity to delay. Because a real-time stream is sampling and reproducing a continuous event, such as speech, individual data samples must arrive at the destination end to be ‘played’ at the right time. If a packet is late, or is lost in transit, then there will be a gap in the information available to the player, and the quality of the audio or video reproduction will degrade. This degradation is significant, and occurs at relatively low levels of packet delay and loss.

Because real-time packets must arrive in a timely manner, it is not possible for the transport protocol to ask for a lost packet to be resent, and then wait for the source to try it again. The round trip delay to the source and back again is too long, and the packet will have missed its play window. Because TCP adds no value to these streams, they are carried instead with the User Datagram Protocol

(UDP), which has no recovery mechanism. Packets are sent by the sender into the network, and they either make it to the receiver on time, arrive late, or are lost in transit.

So somehow we must insure that the packets associated with voice or video conferencing make it though the network in a timely manner, without getting lost, and with no help from the transport protocol. This is the challenge of supporting real-time applications. Quality of Service (QoS) is the mechanism we deploy in our networks to give priority to the voice and video streams, to insure they will be delivered correctly. We will explore the different QoS approaches and how to deploy QoS for voice and video in this document.

3.0 What is Quality of Service (QoS)?

The term QoS has been used to describe many different ways of providing better service to some types of traffic in a network environment. These can include priority queuing, application specific routing, bandwidth management, traffic shaping and many others. We will be discussing priority queuing since it is the most widely implemented QoS approach. Using priority queuing is not the only approach that will work. Any approach that reliably delivers packets on time will support voice or video conferencing. But since priority queuing is the most available QoS, we will describe here how it works, and how to configure it to best support voice and video conferencing.

Enabling QoS in the network is only part of the problem. We will review here the four steps to insuring QoS works correctly in your network as follows:

- Network QoS Implementation
- Classification
- Bandwidth Demand and Bandwidth Availability
- Bandwidth Management

Queues are the primary contributors to packet loss and delay in a packet network. There are queues at each output port in each router and switch in the network, and packets have to enter and leave an output queue on each device through which it passes on its route. If queues are empty or nearly empty, the packet enters and is quickly forwarded onto the output link.

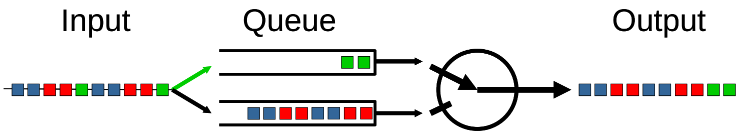

Figure 3 - Packet Queue Diagram

If momentary traffic is heavy, queues fill up and packets are delayed waiting for all earlier packets in the queue to be forwarded before it can be sent to the output link. If momentary traffic is too high, the queue fills, and then packets are discarded, or lost.

A priority queuing mechanism provides additional queues at each switch or router output port, dedicated to high priority traffic. In Figure 3 a simple 2-queue output port is diagrammed. Best effort traffic is all queued in the lower queue, while high priority traffic, here colored green, is queued in the upper queue. The queue policy determines how the queues will be emptied when they both have packets waiting.

A simple priority queue, sometimes called a low latency queue, is always emptied before any lower priority queue is serviced. So in Figure 3, if the queue policy is simple priority, both green packets are serviced first, and then the remaining packets in the best effort queue. If additional high-priority packets arrive while the lower queue is being emptied, service is immediately switched to the high priority queue.

A rate based queue behaves slightly differently. A rate based policy empties queues based on the bandwidth that has been allocated to each queue. If queue 1 is allocated 40% of the available bandwidth, and queue 2 is allocated 60% of the available bandwidth, then queue 2 is serviced 6/4 times as often as queue 1. Using a rate based policy, the queue is serviced often enough to keep the allocated flow moving rapidly through the queue, but if excess traffic arrives, it will back up in the queue while priority is given to the other queue. In our example in Figure 3 the green packets may or may not get forwarded first, depending on the queue bandwidth allocations, and depending on how much traffic has been previously emptied out of each queue.

3.1 Network QoS Implementation

The network must have a QoS mechanism that operates at each switch and router to prioritize real-time traffic. A number of different mechanisms exist in modern networks including IntServ (RSVP), DiffServ, IEEE 802.1p/Q, and IP Precedence. Additionally some enterprises rely on over provisioning, which really means not using a QoS mechanism, but instead insuring there is plenty of bandwidth. We'll discuss in a minute why this is a risky strategy.

To gain the maximum benefit, QoS must work from end-to-end. All routers and switches between the real-time sender and the real-time receiver must have a QoS mechanism available and enabled. Partial solutions will prevent packet loss or delay on the links they serve, but to insure quality delivery, all links should be served.

Let's review the QoS mechanisms available to determine which is best suited to supporting voice and video conferencing traffic.

Over Provisioning

In the discussion above about queues we said that if the queue is empty or nearly empty, that incoming packets will be forwarded without delay. Over provisioning takes advantage of this principle by providing more link bandwidth than necessary, so that queues will be empty most of the time. Enterprises who pilot video conferencing on high bandwidth campus links often are very pleased with the result without any further QoS implementation. Unfortunately this is a short sighted approach, and will inevitably lead to problems later on.

Implementing real-time traffic without a true QoS mechanism is a game of chance. We saw earlier the very bursty nature of data applications. An interesting characteristic of data applications is that as many applications are aggregated together, intuition would tell us that the peaks will be smoothed out. This turns out not to be true; data traffic keeps its bursty profile over many levels of aggregation and over many different time scales. So a large burst of data traffic is just waiting to happen, even on a 1 Gigabit or 10 Gigabit backbone link. When that burst occurs, it will cause delay or packet loss for the voice and/or video conferencing streams. Data traffic tends to be heaviest during the work day, which also coincides with when voice and video conferencing take place, again making the chances for packet loss more likely.

Layer 2 and Layer 3 Quality of Service

QoS is implemented both at layer 3 and layer 2 in the protocol stack. Layer 3 QoS will be recognized and handled by routers in the network. However, congestion also occurs in switch output queues, and so level 2 QoS is also required. This is less true in the WAN network, where the structure tends to be long links connecting one router to the next. However, in the Enterprise network it is often necessary to have both a layer 2 and layer 3 QoS deployment.

Layer 3 QoS
There are three methodologies available in most networks for implementing QoS at layer 3, **IntServ** (RSVP), **DiffServ**, and **IP Precedence**. Let's look at the difference between these approaches.

Integrated Services (**IntServ**) is a very comprehensive approach to QoS defined by the IETF to insure the proper treatment of high priority traffic in a converged IP network. IntServ, using the RSVP protocol, sends out a request to all routers on the path for a given real-time stream, requests the use of a priority queue and asks permission to use a certain amount of bandwidth. If each router responds indicating that those resources are available then the path is enabled, and the stream is given priority along that path. IntServ works well in environments where all routers support the protocol, and where the number of real-time streams is limited. Because RSVP requires each router to maintain information about each active real-time stream, network architects quickly realized that this solution would not scale to large sizes well. Too many resources are consumed in each router. Few enterprises today implement IntServ as their solution.

Differentiated Services (**DiffServ**) was developed as an alternative approach that would have better scaling properties. Rather than specifying resources for each real-time stream, DiffServ allocates resources for a class of traffic. All traffic allocated to that class is treated with the same policy, such as being queued in a high priority queue. The key difference is that whereas the network may have a bandwidth limit for this high priority class of traffic, the network is not managing the demand for this class, only serving the traffic in this class as best it can. So DiffServ makes the job of the network easier, but moves the problem of bandwidth management back to the application. We will discuss bandwidth management in detail later in this document.

IP Precedence is a methodology created in the original IP specification. It is a much simpler mechanism that gives precedence to IP packets marked as high priority. The bit positions in the IP header formerly used for IP Precedence and Type of Service (TOS) have been reassigned for use with DiffServ. There is no advantage to using IP Precedence over DiffServ unless routers are deployed that do not support the DiffServ standard.

Most Wide Area Network (WAN) vendors today are using DiffServ marking to manage QoS in their networks. WAN networks are inevitably a part of the enterprise network using voice or video conferencing, so there is value in choosing an approach that is consistent across both the campus and wide area networks. Careful selection of the DiffServ markings will insure that both the campus network and the WAN vendors treat voice and video conferencing streams with the correct policies.

Layer 2 QoS
The predominant technology in use today for providing QoS at layer 2 is IEEE 802.1p. Switches purchased in the last 7 years are very likely to have this capability built in. IEEE 802.1p functionality is often coupled with IEEE 802.1Q, which is the VLAN functions. Both technologies use the same bit field added to the Ethernet header, to specify a streams priority and VLAN association.

Switches implementing IEEE 802.1p use multiple output queues for each switch output port. Higher priority traffic is assigned to a higher priority queue, and those queues are serviced before lower priority queues, just as explained earlier for the routers.

Table 1 shows the mapping of Ethernet IEEE 802.1p priority mappings to the DiffServ codepoints. If a campus network is using 802.1p, this translation needs to take place wherever level 3 QoS meets the level 2 QoS such as at the core routers or the WAN router.

It is a common misconception that assigning traffic to VLANs completely separates it in the network, and that this is a viable solution for voice and video traffic. VLAN assignment insures that traffic will not be forwarded to portions of the network where those VLANs are not allowed, so traffic separation occurs from a permissions point of view. However, priority is assigned separately. Often VLAN assignment and priority assignment are coupled, e.g. any traffic assigned to the Red VLAN gets high priority. If this is true, then traffic in the high priority VLAN will get precedence at the switches, but

Table 1 - DiffServ to Ethernet IEEE 802.1p Priority Mapping

DiffServ Code Point (DSCP)	PPP Class Number
CS7, CS6	7
EF, CS5	6
AF4x, CS4	5
AF3x, CS3	4
AF2x, CS2	3
AF1x, CS1	2
DE, CS0	0

Ref: Nortel White Paper "Introduction to Quality of Service (QoS)"

the two assignments (VLAN and priority) are not necessarily coupled. If there are more than one 'high priority' VLANs passing through the same switch, they will contend for the same output queue resources.

VLANs are often used as a method for marking traffic. If the endpoint itself is unable to mark traffic, or is not trusted to mark traffic correctly, the VLAN assignment can be used to indicate this is high priority traffic. VLANs can be assigned to traffic arriving on a specified physical port, so any traffic arriving from a particular system, such as a room video conferencing system, can be assigned to a specific VLAN, and thus to a specific priority level in the network. This is discussed further in Section 3.2.

WAN QoS

A number of technologies exist for wide area networks, and each has its own QoS approach. We will discuss leased lines, Frame Relay, ATM, and MPLS.

Table 2 - DiffServ to PPP Priority Mapping

DiffServ Code Point (DSCP)	PPP Class Number
EF	7
CS7, CS6, CS5	6
AF4x, CS4	5
AF3x, CS3	4
AF2x, CS2	3
AF1x, CS1	2
DE, CS0	1

Ref: Nortel White Paper "Introduction to Quality of Service (QoS)"

Leased lines are the simplest technology to manage from a QoS perspective, but often are not the best choice for topology or cost. If an enterprise leases a T1 or T3 circuit between two facilities, than the enterprise router on each end is in full control of how traffic is scheduled onto that WAN connection, and the link behaves like any other link in the Enterprise network. In this situation, whatever QoS approach is chosen for the enterprise network can be used on these WAN links as well.

Table 2 shows the mapping from DiffServ Code Points to PPP class numbers. This mapping is done on the router that attaches to each end of a point-to-point link.

Frame Relay is a well established technology, and is used heavily by enterprises that require relatively low bandwidth connections. The user is cautioned against using a Frame Relay service to carry real-time traffic, as they often have difficulty maintaining sufficiently tight jitter specifications, and will usually not guarantee jitter within the specifications required for good real-time transport.

Frame Relay services provide classes of service that can be used to prioritize one traffic type over another. These classes of service work well to insure interactive applications like Telnet or Citrix get precedence over email transfers and file backup. But they are not designed to provide the kind of priority that real-time traffic requires. Furthermore, because frame relay services often have multiple PVCs using the same physical connection, it is difficult to get true priority on a priority queue. The router serving multiple PVCs creates a virtual port for each PVC, each having a high priority and best effort queue. There is no communications, however, between these two virtual ports, even though they are using the same physical port. Hence high priority traffic sitting in a queue on one virtual port cannot override traffic in the best effort queue on the other virtual port. This means true priority queuing is not happening, and leads to intermittent voice and video quality problems.

ATM (Asynchronous Transfer Mode) is a 1990s technology which was heavily deployed by wide area networking vendors. ATM has built-in sophisticated QoS mechanisms that do a good job of separating traffic.

Much of the technology for measuring and policing bandwidth flows was developed by the ATM Forum during the initial days of ATM deployment. These techniques are used today in DiffServ and MPLS implementations.

ATM service categories are not the same as DiffServ categories, but the DiffServ categories can be mapped into ATM categories at the network boundary. Table 3 shows the mapping between DiffServ code points and ATM service categories.

Table 3 - DiffServ to ATM QoS Mapping

DiffServ Code Point (DSCP)	ATM Service Category
CS7, CS6, CS5, EF	CBR or rt-VBR
AF4x, CS4, AF3x, CS3	rt-VBR
AF2x, CS2, AF1x, CS1	nrt-VBR
DE, CS0	UBR

Ref: Nortel White Paper “Introduction to Quality of Service (QoS)”

Enterprises using ATM within the corporation can use this QoS mechanism as well. Larger enterprises have ATM backbones between major sites. ATM QoS will give priority to classes higher on this list.

MPLS (Multi-Protocol Label Switching) is the latest WAN technology and appears to be the current or future plan of most WAN service providers. MPLS has many of the characteristics of ATM that allow services providers control over how traffic flows in their networks. This allows providers flexibility in how they offer services, and allows them to quickly adapt to new market needs for different services. Because MPLS allows providers this flexibility, many are offering services that include quality of service guarantees. This has led to the misconception that MPLS implies QoS, which it does not. Often a service provider will have built a good QoS offering on top of their MPLS implementation, but the details of this QoS implementation need to be understood to insure that real-time traffic gets the right treatment.

Service providers using MPLS create different classes of service by configuring their routers to either recognize bits in the MPLS tag, or by prioritizing specific routes based on their individual labels. For either method, there are a fixed number of classes, often eight or less. Thus there will again be some mapping that needs to be done between DiffServ markings and the specific traffic classes implemented by the MPLS core. Most MPLS-based WAN providers are using DiffServ to identify traffic classes. Work with your proposed MPLS-based vendor to understand how they map and support the different DiffServ markings into the classes of traffic they support, and understand the forwarding behavior they specify for each class in their network.

Recommendation – DiffServ Markings

A working group of the IETF has issued a draft document recommending DiffServ markings for specific types of traffic.

These recommendations are shown in Table 4. The intent of this document is to create consistency between WAN vendors to eventually allow quality of service to be carried across multiple WAN providers while maintaining similar forwarding behavior through each. Using these markings will make an enterprise compatible with WAN vendors following the IETF recommendations.

Table 4 shows Telephony (VoIP) using the EF marking, and video conferencing uses AF41, AF42 or AF43 markings. Telephony signaling is slotted in between these two, using the CS5 marking. Video conferencing signaling is not explicitly called out. It should go below the level of the video and audio streams, but not be considered ‘Standard’ (which is Best Effort), since people are waiting on the results of the signaling transaction. The High Throughput Data category (AF11, AF12 or AF13) works well for video conferencing signaling.

Table 4 - IETF Recommended DSCP Markings

Service Class	DSCP	PHB Used	Queuing	AQM
Network Control	CS6	RFC2474	Rate	Yes
Telephony	EF	RFC3246	Priority	No
Signaling	CS5	RFC2474	Rate	No
Multimedia Conferencing	AF41 AF42 AF43	RFC2597	Rate	Yes per DSCP
Multimedia Streaming	AF31 AF32 AF33	RFC2597	Rate	Yes Per DSCP
OAM	CS2	RFC2474	Rate	Yes
High Throughput Data	AF11 AF12 AF13	RFC2597	Rate	Yes Per DSCP
Low Priority Data	CS1	RFC3662	Rate	Yes
Standard	DF (CS0) + other	RFC2474	Rate	Yes

Reference Internet Draft draft-ietf-tsvwg-diffserv-service-classes-02.

The third and fourth column of Table 4 shows the forwarding behavior recommended, and refers to the appropriate IETF specification for details. Note that Expedited Forwarding (EF) is recommended for telephony, which is implemented with a priority queue. As discussed earlier, the priority queue is always emptied before other queues, giving the lowest possible latency to that traffic. It is suggested here that video conferencing be implemented using a high priority Assured Forwarding (AF) marking, which is implemented with a rate-based queue. This is appropriate for a converged network where both voice and video conferencing are being implemented, because it insures that the voice traffic has priority over the higher bandwidth and bigger packets of the video conferencing stream. A rate based queue will properly forward video conferencing traffic as long as the queue rates are set to exceed the worst case amount of traffic generated by the video conferencing streams. Management of this bandwidth is very important, and will be discussed in detail later.

Recommendations on separating the voice versus the video portion of a video conferencing stream vary. Some claim that there is no value in getting the voice component of a video conference delivered earlier than the video portion, since the receiving unit must delay the voice to synchronize it with the video signal. There is, however, value in giving audio better priority if it means it will get less interference with other streams, and thus be delivered more reliably. A video conference with poor visual quality can proceed as long as the audio remains clean, but without good audio, no conference exists.

It is possible to use Expedited Forwarding (EF) for video conferencing streams if video conferencing is the only real-time traffic in the network. Most enterprises are either making the transition to VoIP or are considering it in the near future, so using EF for video is not a good long term strategy.

Active Queue Management (AQM)

The last column of Table 4 indicates whether Active Queue Management (AQM) is recommended. Notice that for telephony, the answer is no. The predominant AQM mechanism is Random Early Discard (RED). This algorithm is designed to help manage multiple TCP flows when the link is near the congestion point. RED determines that the queue length is getting long, and then randomly selects packets within the queue for discard. When TCP streams lose packets, they back down their sending rate, thus reducing the congestion.

As we discussed earlier, real-time streams lose quality quickly when packets are lost, so inducing packet loss is not to their advantage. The IETF recommendation in Table 4 recommends that AQM be used on video conferencing streams, under the assumption that the video conferencing endpoints know how to reduce their sending rate when packet loss occurs. Assuming that functionality exists, video conferencing streams would change their video rates to a lower bandwidth when they detect packet loss, and this would relieve congestion in the network, making all streams again able to perform well.

A better approach is to insure that this condition does not happen in the first place. For a video conferencing environment where room-based video conferencing systems are the primary users of the video conferencing bandwidth, the bandwidth should be predicted, agreed upon and scheduled. Relying on the endpoints to negotiate reduced bandwidth means that there is a temporary interruption in the quality of the video conference, followed by a reduction in the video quality. This is not acceptable in a business level video conferencing environment.

If making ad hoc video calls using desktop video conferencing capabilities becomes popular in the future, it may be necessary to create two classes of video conferencing traffic, and to manage the bandwidth of each separately. Managing the bandwidth of ad hoc video conferencing users has to be done on an averaging basis, similar to the way it is done for voice calls. This group of users then may have to degrade their video quality when the network gets congested or receive a busy signal if insufficient bandwidth is available to place the video call. Business class video conferencing should be kept independent of this ad hoc class so that scheduled calls can be placed with the scheduled bandwidth, and obtain consistent, quality results.

3.2 Classification

Classification is the task of determining which streams deserve high priority treatment, and identifying them with a marking so that the network switches and routers will recognize them. All the previous discussion about QoS implementations assumes we know which streams ought to get preferential treatment, and which streams should just get best effort support. Classification is the job of making this decision and marking streams accordingly.

Classification by Endpoints

In many implementations it is possible for the endpoints themselves (phones, video conferencing endpoints, gateways or bridges) to identify the voice and video streams. The endpoint is the most knowledgeable component in the network, because the endpoint contains and understands the application. It is a simple matter for the endpoint to identify which streams contain voice or video content, and which streams are data transfers or control traffic. Most voice and video endpoints have a configuration option that allows the QoS marking to be specified, and that marking is then applied to the high-priority streams.

However, the network may not trust the endpoints to properly mark their streams. A video conferencing system may be properly identifying its traffic as voice, video, data or control, but a PC can also emulate these same markings for non-priority traffic, and take advantage of the better service. This is like driving in the commuter lane with only one person in the car.

Classification by the Network

Networks often avert this problem by classifying data streams themselves. This means the edge switch or router must look into the packets of the data stream to determine which ones are high priority and which are not, and mark them accordingly. A network will do this classification at the edge or access router; distribution, core and WAN routers will then trust the markings established at the edge. Classification can be done on the basis of:

- IP address (well known end-point)
- TCP or UDP port number
- Physical port

Often a combination of these parameters is used, perhaps in conjunction with the markings established by the endpoint itself. For instance, if a video and/or audio bridge is identified by its IP address, which is statically assigned, all UDP-based traffic can be marked as high priority traffic. Since the endpoint is a well-known IP address which is difficult to spoof, and the endpoint sends primarily real-time traffic streams, this classification rule will work well. Conversely, if desktop IP-phones or video endpoints are being user installed throughout the enterprise, it is difficult to have an up-to-date list of 'approved' high priority devices to use in this classification approach. Different portions of the enterprise may operate with different levels of trust with respect to endpoints making their own traffic classifications.

Enterprise Classification Policy

Each enterprise should develop a policy on how classification will be accomplished. The policy should scale well as the deployment of phones or video endpoints increases. The policy needs to reliably identify real-time streams without compromising the integrity of the network. The need to manage who is using the high priority service will be discussed in more detail in the bandwidth sections below.

Note that the default behavior of routers, if not configured to accept an endpoint classification, is to remark the incoming packets to the best effort category. Enabling QoS in the system and marking at the endpoints is often not sufficient. Insure that edge routers are properly configured to either trust the endpoints or to classify and mark packets themselves. It is useful to capture data on the receiving end and determine if packets maintained their QoS markings throughout their journey across the network.

4.0 Bandwidth Demand

Bandwidth use is an integral part of QoS. Sufficient bandwidth must be in place on each link to carry the expected real-time traffic. So the first question is what is the expected traffic? It is important to analyze expected demand so that proper bandwidth planning can be done to support video conferencing on the network links.

Bandwidth demand analysis can be done in a number of different ways. If a video conferencing service already exists in the organization, then there is some history available of how many calls are placed during the day, the locations called, the duration of those calls and the call bandwidth. This information can be compiled into a demand graph per link, and then the maximum values can be obtained. Figure 4 is an example of such an analysis. Video conferencing calls were mapped onto a network diagram to determine which WAN links were used for each call. The time of day, duration and bandwidth of the calls were noted. For each 30 minute period of the day, the bandwidth consumed by each call active on each link was summed up to show the total demand on that link for that period. Figure 4 shows those results for one of the WAN links. This link had a maximum usage of 6.5 Mbps for video conferencing.

A similar approach can be used for an enterprise that is currently transporting video conferencing traffic over ISDN lines. The call history can be used to determine what the demand on the IP network would have been if those calls were instead carried on the IP network.

If no existing call information is available, then call density and patterns must be estimated based on expected usage. If video conferencing will primarily be taking place from video conferencing rooms, then demand can be estimated by making some assumptions about room utilization. Call destinations will have to be estimated by someone with knowledge about the business and likely call patterns for the users.

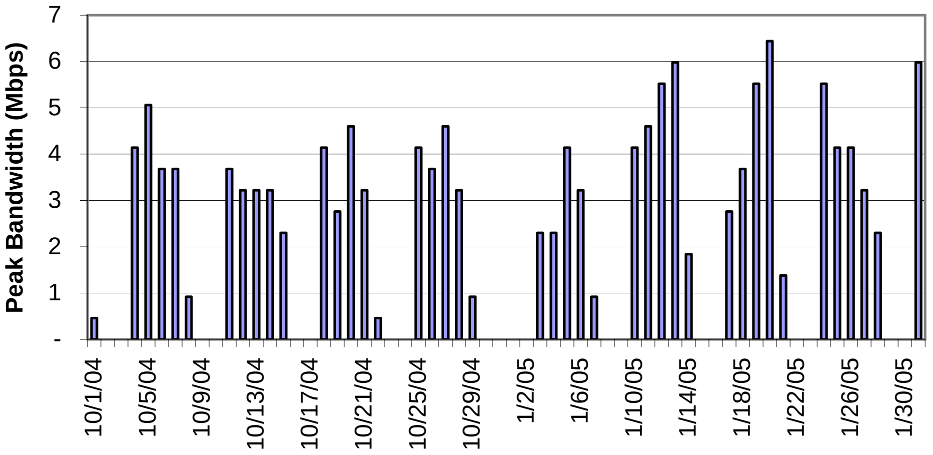

Figure 4 - Video Conferencing Bandwidth Demand

Video conferencing is usually used for a scheduled meeting, with duration of at least 30 minutes, and often more like an hour or an hour and a quarter. A heavy deployment of desktop video conferencing systems may change this dynamic, as users begin to use their video systems more like a telephone, for ad hoc meetings which start at random times, and have shorter durations.
Modeling telephone usage is done on a statistical basis using Erlang tables. Given the frequency and duration of calls, the Erlang calculation determines the amount of available bandwidth required to obtain a given level of bandwidth availability, i.e. in order to not have calls be blocked. More information on Erlang tables and a calculator can be found at www.erlang.com.

Bridge (MCU) Bandwidth Demand

Key infrastructure components of the video conferencing system need special consideration. First consider the video conferencing bridge (or MCU). If 20 video conferencing endpoints are engaged in a conference call, all 20 endpoints have established a full duplex connection to the bridge. The bridge network connection must be able to sustain the maximum number of endpoints that will be in all simultaneous conference calls. Thus the bridge should be placed near the core of the network where bandwidth is more plentiful. Furthermore, the bridge should be placed in the facility where the highest percentage of conference call users reside to minimize the WAN traffic required to support these conference calls.

Each client that connects to the bridge will have a traffic stream flowing from the client to the bridge at the bandwidth negotiated for that video conference. If each client has negotiated a 384K bandwidth call, and there are 20 clients, the bridge will be supporting 384K x 20 or 7.7Mbps of traffic. When we add the 20% additional bandwidth required for IP packet overhead, this now comes to 9.2Mbps.

Some video conferencing endpoints also support a built-in conferencing mode. If a video conferencing endpoint is acting as a bridge for a small conference, there will be a proportionate increase in the bandwidth to that client. A 4-person conference using one of the 4 clients as a bridge will generate three full duplex streams to the client acting as a bridge. The other three clients will see a single full-duplex stream.

Gateway Bandwidth Demand

A gateway is often used to connect an IP-based voice or video system to the PSTN. Calls being placed from within the IP infrastructure pass through the gateway to an ISDN or POTS connection. Gateways can be stand alone units, or are sometimes incorporated into a Bridge.

Bandwidth demand for the gateway should be calculated based on the number of simultaneous calls being placed through the gateway to the PSTN. This calculation is often straight forward, because there is a limited amount of PSTN bandwidth. If the gateway is incorporated into the bridge, this bandwidth should be added to the bandwidth due to conferencing.

IP and ATM Overhead

Bandwidth calculations need to take into account the overhead of an IP-based transport. In an ISDN environment, a 384Kbps video conferencing call will consume 384Kbps of the available transport bandwidth. In an IP environment the same 384Kbps bit stream is broken up into packets, and those packets carry the additional overhead of an RTP header, a UDP header, an IP header and a layer 2 transport header. This overhead must be added to each call bandwidth to determine the real impact on the IP network. Bandwidth overhead can be approximated using the values shown in Table 5. Voice overhead for lower bandwidth calls like G.729 is higher than the values shown for voice in this table. Header compression should be considered when many low bandwidth voice calls are carried across WAN trunks, to help minimize overhead.

Table 5 - IP Bandwidth Overhead

Type	Ethernet/ MPLS	PPP	ATM
Voice G.711	36%	25%	40%
Video	20%	18%	25%

The video in Table 5 is the overhead for the combined video and audio streams of a video conference. For example, a 384K video conference consumes 384Kbps x 1.2 = 460Kbps of network bandwidth.

5.0 Available Bandwidth

Once the bandwidth demand has been calculated, an evaluation of existing network bandwidth and utilization is required to determine if there are sufficient resources to support the new real-time load. Each link of the network needs to have sufficient bandwidth to support the voice and video traffic expected, plus the existing data applications that use those same connections.

Although this sounds like a daunting task, in practice it usually means evaluating the wide area network links, the backbone connections of the bridge, and client connections where there may be 10Mbps Ethernet or shared Ethernet connections. Often much of the infrastructure of an enterprise does not need detailed bandwidth analysis, just these key elements.

Client connections should all be upgraded to 100 Mbps full duplex if possible. If the video conferencing endpoint does not support full-duplex operation, it is preferable to run at 100Mbps half-duplex. If the endpoint supports full duplex, but does not support 100 Mbps, it is preferable to run at 10 Mbps full duplex. When video conferencing runs on a half-duplex link, such as older Ethernet links using a hub, the video conferencing application consumes a larger portion of the available bandwidth. A 10 Mbps full duplex Ethernet connection supports 10 Mbps between the client and the network switch, and another 10 Mbps between the network switch and the client. If the client is running at 384K, it consumes 460Kbps in each direction, or 4.6% of the available bandwidth. If the same client is running on a half-duplex Ethernet, it consumes 9.2% of the available bandwidth.

There are two parameters to consider when evaluating the WAN links. First, the expected voice and video (real-time) load should never exceed 35% of the link capacity. Priority-based QoS mechanisms begin to lose their effectiveness at this level. Running with more than 35% high-priority traffic means that the traffic starts to compete with itself, and reliable delivery is compromised.

The second parameter is the total bandwidth utilization of the link, including the real-time components and the data components. It is straight forward to determine the bandwidth demand of the real-time applications, but determining the needs of data applications is much more difficult. Data applications, as we saw in Section 0, are very bursty, and when many of those applications are aggregated on a link their profile is still very bursty. Data applications depend on bandwidth overhead to get good performance. If the bandwidth of a link is limited to the average consumption of the data applications, the applications themselves slow down, creating user frustration and reduced productivity.

One method of determining the utilization required by existing data applications is to measure current utilization during the busy hour of the day. Look at the utilization of each important link with as fine a resolution as possible. Typical bandwidth monitoring tools average utilization over some period of time (15 minutes, or an hour). This averaging smoothes out the utilization peaks, and gives a false impression of how much bandwidth is needed for proper data application performance. If monitoring can be done at a finer granularity (5 minutes or even one minute) more accurate results are obtained.

Another useful metric is to talk to application users, and determine if they notice a reduction in application performance during the busy hour of the day. This would indicate that bandwidth is already a scarce commodity during those times. It may be useful to test those network links using a synthetic traffic generator to simulate the expected video conferencing load, and determine the impact on existing application performance. The testing should be done during the busy hour when those applications are running at their peak load. If there is concern about impacting the business, implement the tests slowly, adding additional synthetic bandwidth and monitoring application performance.

If bandwidth is scarce, it is often valuable to evaluate the traffic flowing across critical links during the busy hour to determine if some of that traffic does not support legitimate business purposes. Finding and eliminating these unwanted traffic streams can often free up bandwidth.

Giving specific utilization values that are acceptable is difficult because each enterprise has a different traffic mix and different needs. On the low end, 35% total utilization means good performance for all applications. Seventy percent (70%) utilization is a high end limit in almost all cases. But this leaves a wide range of choices.

Background tasks that are not time sensitive, such as email transfer, backups, downloads or database synchronization, will work at higher load percentages. Applications where users are waiting for an immediate response, such as HTTP-based applications, client-server applications or keystroke applications like Citrix and Telnet, will be less tolerant. There is often value in extending the QoS strategy to give these interactive applications a priority higher than the background tasks, but lower than voice and video streams (see Table 4.)

6.0 Demand Management

If the network testing determines that there is insufficient bandwidth on critical links, the enterprise has a few options to resolve the conflict:

- Bandwidth upgrade
- Reduce voice or video conferencing demand
- Compression / Application Acceleration Appliances
-

Bandwidth Upgrade - A bandwidth upgrade is always possible and may be the only solution if insufficient bandwidth is available to carry the required voice or video conferencing load.

Limit Conferencing Demand - The second option is to limit the video conferencing demand. This can be done in a number of ways. First, the bandwidth used by video conferencing calls can be limited. Better video quality can be obtained at 1 Mbps or 512 Mbps, but quite good quality can be obtained at 384Kbps, and even at 256Kbps. Use of the new H.264 video compression algorithm enhances these lower bandwidth calls and provides better quality. So if the expectation of a remote office was that they would be able to make calls at 512K, perhaps introduction of H.264 and a reduction in call bandwidth to 384K or 256K will reduce demand sufficiently.

A second way to reduce demand is to manage call volume so that a limited number of calls can occur simultaneously across each link. If a remote office has three video conferencing units, but the bandwidth of the link can only support two simultaneous calls, a scheduling policy can be put in place to insure that only two systems are being used concurrently. The simplest case of this policy is to insure that the remote office only has the number of video conferencing endpoints that the link can support.

Voice demand can be reduced by using a lower bandwidth codec (e.g. G.729) and by implementing header compression on low bandwidth links.

The voice or video conferencing gatekeeper can also be used to help manage bandwidth utilization. The gatekeeper can be assigned a maximum bandwidth available between pools of endpoints, which relate to the topology of the network. The gatekeeper will then only allow calls across that link up to the available real-time bandwidth allocated to that link. The bandwidth value given to the gatekeeper is the maximum amount of real-time traffic allowed on that link, not the link capacity. Once the link utilization reaches this maximum amount, the gatekeeper will refuse additional call requests.

Overflow strategies can also be considered for those environments where a connection refusal is not appropriate. If a data network has redundant paths to a remote office, application specific routing can be employed to take advantage of the additional paths, to support a higher call volume. Another approach is to have ISDN connections available, and to route overflow traffic through the ISDN connections when necessary. Monitoring ISDN usage during busy hours will provide a simple business case indicating when it is cost effective to add more IP bandwidth.

Compression – One more option is to compress the existing data traffic. A new class of data appliances is appearing on the market that use various tricks to both reduce data traffic and increase application performance simultaneously. These appliances use compression, caching, TCP termination, transparent turns reduction and other techniques to accomplish their goals. There is a bit of work to determine which approach best suits the data streams employed for each situation, but these appliances can often make room on the link so that video conferencing or voice traffic can be introduced without requiring a bandwidth upgrade.

Scheduling – The epitome of bandwidth management is to be able to schedule bandwidth at the same time that a conference room and conference bridge ports are scheduled. Scheduling bandwidth insures that the executive conference to be held next week will have the appropriate bandwidth waiting for it when the call is set up, and that ad hoc voice or video conferencing users have not pushed link bandwidth to the maximum just before the meeting is to start. In a centrally managed video conferencing environment, this kind of bandwidth management is possible through a manual process. Conferencing schedulers can insure that no more than the maximum number of conferences are scheduled to use a particular network link during each half-hour period of the conference, for example. Making this work requires conferencing schedulers to understand the network topology as it applies to video conferencing connections.

In a more dynamic environment, with users making ad hoc calls without scheduling, a more sophisticated approach is required. Polycom offers tools like Polycom Conference Suite (PCS) and ReadiManager, which include endpoint, MCU port and bandwidth scheduling.

7.0 WAN Vendors and Technologies

Connecting enterprise locations is often done with the help of a Wide Area Network (WAN) service provider. The service provider links become an integral part of the enterprise network, and of the real-time traffic support. Thus it is important that we take a close look at how service providers implement QoS on their access links and within their own core networks, and understand the impact on the transport of the enterprise real-time streams.

7.1 QoS in the WAN Core

WAN service providers have a high speed core network, often using links at OC48 (2.4Gpbs) or OC192 (9.7 Gbps) data rates. Jitter is minimal at these data rates because queues are emptied so quickly. Some vendors claim that they can provide high quality transport for real-time traffic without implementing QoS in their core networks. The reader is cautioned to think this through carefully before deploying with one of these vendors.

Remember our discussion in section 0 about over provisioning? This is the strategy that these vendors are implementing. By providing a very high bandwidth core, they hope to have minimal packet loss and jitter, and thus be able to provide good quality without the overhead of implementing a QoS algorithm. Often these vendors will provide a Service Level Agreement (SLA) that indicates they meet specifications sufficient to support real-time traffic. We will discuss below how to read and evaluate these SLAs.

There are WAN vendors today who are providing classes of service for different types of streams. The reader is encouraged to seek them out and implement a WAN strategy using these service providers, because they can give better guarantees for quality transport of real-time traffic.

7.2 QoS in the WAN Access Links

Of the many network links between two voice or video conferencing endpoints, the WAN access link is likely to be the connection with the least available bandwidth. This means it is the link with the highest probability of causing packet loss or jitter problems. Queuing problems occur where high speed links are being switched through to a lower speed link, which is the case at each end of the access link. Implementing QoS on this link is critical.

Prioritizing traffic entering the WAN is straightforward, because this traffic comes from the enterprise LAN and thus is under control of the enterprise routers. If proper classification is done within the LAN, the real-time traffic can be prioritized as it leaves the LAN and enters the WAN access link.

Prioritizing traffic as it leaves the WAN core and enters the access link has to be the responsibility of the WAN vendor. If the WAN vendor is implementing QoS, this feature will be supported. If the WAN vendor is not implementing QoS in the core, insure that at the very least they are implementing QoS on traffic as it leaves the core and enters the access link. Also insure that the vendor is carrying the QoS markings across the core so they are still available to traffic at this egress router. A good test is to capture traffic as it enters the enterprise from an access link, and verify that QoS markings are intact.

7.3 SLA Interpretation

The Service Level Agreement (SLA) provided by a WAN vendor will have specifications for availability and performance as well as specifics of responsiveness when there are problems on the link and how the enterprise will be compensated for those problems. We will concentrate here on the performance specifications.

Performance specifications should include a value for latency, packet loss and jitter. Latency is affected by the geographic distances involved with the WAN connection due to the limitations of the speed of light. An SLA that includes connections from continent to continent will have longer latency specifications than for an SLA that only includes domestic connections.

Check carefully to see if the SLA covers traffic flowing through the access links between the enterprise and the WAN vendor. Some SLA specifications only guarantee traffic specifications between the edge nodes of the WAN vendor's network.

Packet loss and jitter specifications are often given as an average over some period of time. Take careful note of this averaging period, it is often quite long, as much as a month. Like bandwidth averaging, this averaging means the details of how the link is performing during important busy-hour periods can be buried in the longer time-period specification. As an example, a packet loss specification of 0.1% over a month period could mean:

Nights and Weekends loss = 0.02%
Work day loss (except busy hours) = 0.05%
Busy hour loss (2 hours/day) = 1.3%

So instead of getting the expected 0.1% packet loss during the important busy hour period, the network is failing at 1.3% loss, causing significant video and voice impairment, and all within the SLA specification. A more appropriate SLA specification would have performance guarantees over much smaller periods of time, like an hour.

7.4 Enterprise to WAN QoS Handoff

Most WAN implementations today recognize the IETF DiffServ codepoint markings, and use those to map traffic into specific traffic classes. It is important to align the enterprise QoS classes with the WAN vendor classes, and to have consistency between different WAN vendors where more than one vendor is employed within an enterprise. It is possible to remap QoS markings at the edge of the WAN cloud so that enterprise markings can be properly aligned with WAN classes, but this adds unnecessary complexity if it is not needed. If a QoS strategy is being developed, poll the current and likely future WAN vendors to learn their QoS marking strategy, and align the enterprise strategy accordingly.

7.5 Real-Time over VPNs or the Internet

Many small to medium sized enterprises today are taking advantage of Virtual Private Networks (VPNs) to connect their geographically distributed offices. VPNs create an encrypted tunnel through the public Internet. The advantage of a VPN is that the cost is often much less than a dedicated connection. Enterprise VPNs come in two flavors, those that connect two offices through a single WAN provider, and those that use the open Internet, so they may use more than one service provider and their associated peering points.

Carrying real-time traffic through these VPNs is as risky as using over-provisioning for QoS. There is usually no QoS capability offered in the VPN connection. Quality can often be reasonable in the case where a single service provider is providing connectivity at both ends, but again no guarantees about bandwidth, loss or jitter are available. Some enterprises use this approach anyway, because the value of a voice call or video conference to an Asian manufacturing plant or a European development center justifies the risk, and because the users can be tolerant of failures. If the quality expectation is high, supporting management staff meetings, sales updates, presenting to clients or other high visibility uses, than the risk of quality degradation and call failure may be too high to allow this use.

Using the Internet for real-time traffic carries the same risks as a VPN, with less control. When connecting to another party via the Internet, multiple carriers may be involved, and the user has no control over how the call is routed. Hot-potato routing algorithms often insure that traffic flowing one direction will take a different route than traffic flowing in the reverse direction. Educational and research institutions have had some luck using the Internet where they have very high bandwidth connections, but the risk of having a poor quality connection is high.

8.0 Network Verification

Testing the IP network for real-time support is the only way to get a real understanding of the state of an IP network. A number of vendors offer testing tools that can be used to test the network, or a consultant can be contracted to do a network audit. This step provides direct evidence of the ability of the network to handle the bandwidth and timing requirements that real-time traffic imposes.
Synthetic test tools use a hardware or software agent, installed at locations around the network to represent voice or video clients. These agents are then coordinated by a central console to conduct real-time tests. Tests can be constructed to mimic the real-time traffic expected in the target network, so voice calls and video conferencing calls of various quantities and bandwidths can be mimicked.

Traffic flowing between these two agents stresses the network with the bandwidth consumption it creates. If the consumption of this additional bandwidth impacts the performance of other business applications, this can be noted and further bandwidth investigation started to solve those issues.

The receiving agent for each test stream also checks the packet stream for latency, packet loss and jitter. These three key characteristics determine the network's ability to support the timely delivery of real-time data. Where test streams indicate poor real-time performance, diagnosis work can be done to determine where packets are lost, or where jitter is introduced.

Network verification often finds both high level problems and lower level problems. High level problems are ones that may have been missed during the design stage, such as enabling the appropriate QoS, losing the classification markings at an edge router, or insufficient bandwidth on a link. Lower level problems may include routers with an out-of-date software revision, an incorrect access list definition, or insufficient memory or CPU resources. Poor LAN links, such as shared 10Mbps Ethernet connections or CAT 3 cables can also cause poor real-time performance. A common issue is a poorly functioning Ethernet negotiation, where one end of the link resolves to a half-duplex configuration and the other end resolves to a full-duplex configuration. Data traffic will often work well across this mismatch, but real-time traffic will fail.

Table 6 - Real Time Test Parameters

Parameter	Value
Packet Loss	< 0.1%
Packet Latency	<= 100 ms
Packet Jitter	< 40 ms

Network verification should be done well ahead of real-time traffic deployment, so that any issues uncovered during the verification process can be resolved before real-time traffic begins to flow.

Table 6 shows good target values for an enterprise network supporting voice or video conferencing. Meeting these goals will insure quality voice and video conferencing transport.

Packet Loss is often the most difficult parameter to meet. Many vendors' products will indicate that they can deliver good quality voice or video with much higher packet loss values than shown here. Each of those approaches uses some kind of algorithm to mask the effect of lost information. Whenever information is lost, the quality degrades, so the products have to make guesses as to what the missing audio or video information was doing during the time represented by the missing packet. Although these approaches can yield good results, the best approach is to not lose the information in the first place. Because of the dynamic nature of the IP network, there will be plenty of opportunities to test packet loss concealment algorithms when something goes wrong. Design to make the network clean, and use packet loss concealment as a backup plan.

Latency affects the quality of an interactive voice or video conferencing call by slowing down the response from the other party. After latency reaches 200 ms the delay effect is noticeable, and parties have trouble interrupting each other, and maintaining the flow of a normal conversation. The 200 ms value represents the delay from the speaker to the listener, including all the delays of the encoding, transmission and decoding of the signal. The network is only involved with the transmission portion of this delay. Network latency should be kept below 100ms to insure that speaker-to-listener latency stays below 200ms.

Latency is affected by congestion and by geographic distance. Congestion can be managed through bandwidth management and QoS, but geographic distance and the speed of light are harder to manage. Using a satellite incurs a long delay because of the distance signals have to travel up and back to a geostationary satellite. In some global routing cases it is possible to get better routes. The path traffic takes from Asia to Europe, for instance, often flows through the United States. Finding carriers that will route this traffic by a more direct geographic route can lower the latency impact. Enterprises not using satellites and working within a single continent will not experience latency problems due to distance.

Jitter is the variation in packet delay as packets cross the network. Jitter is usually measured as packet inter-arrival time, which isn't quite the same thing, but works well for a packet stream that is sent at a periodic interval like a voice stream. Jitter is primarily caused by queue delays. If a packet passes through a nearly empty queue, its delay is short. If the next packet behind it finds the same queue nearly full, it waits a long time before being forwarded. Jitter management is done by managing queue depths. A shorter queue means packets won't have to wait long, but it also means bursts of traffic will cause packet loss sooner. So by managing jitter properly, it pushes the problem back onto packet loss.

Jitter is managed by a receiving endpoint with a jitter buffer. This buffer holds an on-time packet for some period of time before playing it. The depth of the jitter buffer determines the amount of time an on-time packet is held. A late packet will be moved through the jitter buffer quickly, to bring it up to the right 'play' time. Jitter is specified in Table 6 as 40 ms because some Polycom conferencing systems have a 40 ms jitter buffer. This means packets arriving as much as 40 ms late can still be played on-time, but packets arriving later than this will be discarded. The network needs to keep jitter within this bound to prevent more packets being dropped, and the subsequent degradation of voice or video quality.

9.0 Network Monitoring

Network monitoring is the ongoing version of network verification. Networks are very dynamic; changes and additions are made to the network every day. The management tools used for data networking are often insufficient to track and manage issues with real-time support, because they are not measuring the right parameters, and are not measuring with sufficient granularity. To properly maintain a network supporting real-time traffic, a real-time measurement tool needs to be deployed.

Network verification tools need the following characteristics to properly monitor the health of an IP network carrying real-time traffic:

- Testing end-to-end across the network
- Testing packet loss, jitter and latency
- Storing historical data into a database for forensic analysis
- Thresholds available to cause alarms when quality degrades

Network monitoring tools built for supporting VoIP will also calculate and report Mean Opinion Score (MOS), a measure of the quality of the voice call. A similar measure for the perceived quality of a video conferencing image is being standardized by the ITU, and vendors will be including this calculation in tools in the near future. This allows the tool to present a test result that directly represents the quality of the call, giving IT managers a way to judge the effectiveness of their IT infrastructure.

Network monitoring can be done with purpose-built hardware and software tools. A number of vendors supply test tools to support the testing of networks carrying real-time traffic (See Appendix 1). These tools place hardware or software probes around the network which generate small amounts of synthetic traffic, and measure the ability of the network to properly deliver that traffic. Some of these tools will also passively monitor real-time traffic streams to determine their quality.

An alternate approach is to take advantage of the voice or video conferencing endpoints involved in real-time connections. Often these devices monitor the quality of the traffic delivered to them from the far end, and record this information in a Call Detail Record (CDR.) Quality information may also be available dynamically by reading MIB data from the endpoint, or opening a management connection to the endpoint during a voice or video call. Long term collection of the CDR information is useful for tracking the overall quality of the service being offered, and helps find correlations between poor quality and the endpoints or geographies involved in those calls.

Often a combination of these two monitoring approaches provides the best and most timely information to the network team about the ability of the network to carry real-time traffic. Determining and implementing a monitoring strategy is important to managing real-time call quality.

10.0 Enterprise Service Level Agreement (SLA)

Introducing real-time traffic into the enterprise changes the way both the voice/video team and the network team operates. A new set of expectations is suddenly applied to the network, and traditional approaches no longer work. This change can cause difficulties between the teams, and slow down the deployment of voice or video. It also interferes with the prompt resolution of problems, as each team blames the other for the current issues.

Developing an Enterprise SLA is a way of bridging the gap of misunderstanding between the teams, during the transition period. It helps both teams be successful while the organization learns how to adjust to the new reality of real-time traffic on the IP network.

An Enterprise SLA is a written agreement between the voice/video team and the network team, which defines the characteristics required of the IP-network to support real-time traffic. It is very similar to the agreement written with an external WAN provider. This document specifies the following parameters required of the network:

- Bandwidth – Expected real-time traffic load by link
- Latency – Maximum latency between any two video or voice endpoints
- Packet Loss – Maximum allowed packet loss during any portion of the business day
- Jitter – Maximum allowed packet jitter during any portion of the business working day

Within the specification of these items will be included the timeframes over which these measurements are taken, and with what granularity the information is captured (hourly, quarter-hourly, etc.) Specifications should be written so that quality can be maintained during the busiest hours of the work day.

These four items are the most critical. Additional items in the SLA can include process definitions for how users will report problems, and how the two organizations will communicate when the problems are real-time related. Additionally the SLA needs to have a process for renegotiating these parameters, especially bandwidth, as the real-time deployment grows.

The SLA document gives the voice/video team a network specification against which they can test. If testing tools are in place to test this SLA, the voice/video team can quickly determine if a reported problem is due to a network failure or an equipment failure. This eliminates the finger pointing, and gets the problems resolved in a timely manner.

For the network team, the SLA document defines the resources required to support voice or video conferencing. Transport over the IP network is often considered ‘free’, but the IT team knows that real money is required to support new applications. The specific requirements defined in the SLA document allow the network team to define the infrastructure support required to meet the specification, and to request the appropriate resources from management to support it.

11.0 Checklist

The following checklist is a summary of this document, and can be used to determine how well the network is prepared for a voice/video implementation.

- □ Identify real-time bandwidth demand by link
- □ Identify data bandwidth demand by link (utilization)
- □ Assess available bandwidth by link, and upgrade as necessary
- □ Determine enterprise QoS classification approach
- □ Determine enterprise QoS technology for the Campus networks (LAN)
- □ Determine enterprise QoS technology for the WAN, and select WAN vendors
- □ Verify WAN support for enterprise QoS approach through testing
- □ Deploy QoS in all portions of the network supporting real-time traffic
- □ Verify network support of real-time traffic with synthetic testing
- □ Implement a bandwidth management methodology
- □ Implement real-time network monitoring capability
- □ Create, negotiate and sign an enterprise SLA

12.0 Conclusions

Deploying applications using real-time traffic, like voice over IP or video conferencing, creates a new and different challenge for the IP-network team. A successful deployment requires careful attention to the requirements of real-time traffic. If each of the steps outlined in this document are tackled and then incorporated into the daily operations of the network, the enterprise can not only have a successful deployment, but maintain a high quality service over the IP-network through the inevitable changes in applications, locations, and the network itself.

Appendix 1

Vendors of network tools:

Network Qualification Tools

- Apparent Networks (www.apparentnetworks.com)
- Clarus Systems (www.clarussystems.com)
- Ixia Chariot (www.ixiacom.com)
- NetIQ Vivinet Assessor (www.netiq.com)
- Viola NetAssessor (www.violanetworks.com)

Network Monitoring Tools

- Brix (www.brixnetworks.com)
- Clarus Systems (www.clarussystems.com)
- Computer Associates (www.ca.com)
- Corvil Networks (www.corvil.com)
- NetIQ Vivinet Manager (www.netiq.com)
- Opticom (www.opticom.de)
- Prominence (www.prominencenet.com)
- Qovia (www.qovia.com)
- RADcom Performer (www.radcom.com)
- Telchemy Vqmon (www.telchemy.com)
- Visual Networks (www.visualnetworks.com)
- Volia NetAssessor (www.violanetworks.com)

Network Diagnostic Tools

- Computer Associates (www.ca.com)
- Ixia Chariot (www.ixiacom.com)
- NetIQ Vivinet Diagnostics (www.netiq.com)
- Touchstone Technologies (www.touchstone-inc.com)
- Visual Networks (www.visualnetworks.com)
- WildPackets (www.wildpackets.com)

References

"DiffServ, The Scalable End-to-End QoS Model", Cisco Systems, 2001, updated Aug. 2005, http://www.cisco.com/application/pdf/en/us/guest/tech/tk766/c1550/ccmigration_09186a00800a3e2f.pdf
"Implementing QoS Solutions for H.323 Video Conferencing over IP, Cisco Systems, Document ID 21662, http://www.cisco.com/warp/public/105/video-qos.pdf
"VoIP over PPP Links with Quality of Service (LLQ /IP RTP Priority, LFI, cRTP)", Cisco Systems, http://www.cisco.com/warp/public/788/voice-qos/voip-mlppp.pdf
"Low Latency Queuing", Cisco Systems, IOS Release 12.0(7)T
http://www.cisco.com/univercd/cc/td/doc/product/software/ios120/120newft/120t/120t7/pqcbwfq.pdf
"Voice over IP, Per Call Bandwidth Consumption", Cisco Systems, May 2005, http://www.cisco.com/warp/public/788/pkt-voice-general/bwidth_consume.pdf
"Cisco IOS Quality of Service Solutions Configuration Guide", Cisco Systems, http://www.cisco.com/univercd/cc/td/doc/product/software/ios122/122cgcr/fqos_c/qcfbook.pdf
"Blueprint for Convergence", Nortel Networks brochure, http://www.nortelnetworks.com/solutions/conv/collateral/nn108081-052004.pdf
"Introduction to Quality of Service (QoS)", Nortel Networks White Paper, http://www.nortelnetworks.com/products/02/bstk/switches/bps/collateral/56058.25_022403.pdf
"QoS Recommendations for VoIP", Nortel Networks White Paper, Ralph Santitoro
"Nortel Guide for Planning and Deploying Converged VoIP Networks to Enterprises", Nortel Networks, November 2005, http://www142.nortelnetworks.com/bvdoc/bestpractice/Nortel_Guide_for_Planning_and_Deploying_Converged_VoIP_Networks_to_Enterprises_1.2.pdf

Proprietary and Confidential

The information contained herein is the sole intellectual property of Polycom, Inc. No distribution, reproduction or unauthorized use of these materials is permitted without the express written consent of Polycom, Inc. Information contained herein is subject to change without notice and does not represent commitment of any type on the part of Polycom, Inc. Polycom and are registered trademarks of Polycom, Inc.

Notice

While reasonable effort was made to ensure that the information in this document was complete and accurate at the time of printing, Polycom, Inc., cannot assume responsibility for any errors. Changes and/or corrections to the information contained in this document may be incorporated into future issues.