

***Best Practices Handbook  
for Ensuring Network  
Readiness for Voice and  
Video Over IP***



# Best Practices Handbook for Ensuring Network Readiness for Voice and Video Over IP

---



**PACKETEER®**

## Table of Contents

Executive Summary.....	3
Introduction.....	3
<i>Important QoS Parameters</i> .....	4
Assessing the Current Network.....	4
<i>VoIP versus VoIP</i> .....	5
<i>Advanced Network Monitoring</i> .....	6
Bandwidth Analysis and Planning.....	7
<i>Bandwidth Utilization Analysis</i> .....	7
<i>Application Performance Analysis</i> .....	8
<i>Bandwidth Planning for Video and Voice Applications</i> .....	9
Implementing Control Procedures.....	10
<i>Packet Marking and Queuing</i> .....	10
<i>Traffic Shaping</i> .....	11
<i>Partitioning</i> .....	12
<i>Facilitating an MPLS WAN</i> .....	13
<i>Adaptive Response Control</i> .....	14
Monitoring and Reporting.....	15
<i>Avoiding the SLA Trap</i> .....	15
Real-World ROI.....	17
Conclusion.....	18
<i>About Packeteer</i> .....	18
<i>About Wainhouse Research</i> .....	19
<i>About the Author</i> .....	19

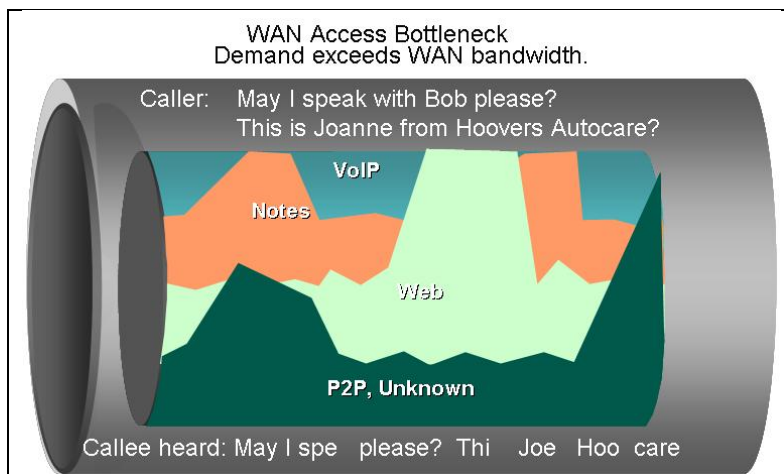
## Executive Summary

Preparing for voice and video over IP requires understanding current network traffic, determining network policies for traffic, and applying quality of service (QoS) controls over the wide-area network (WAN). Many organizations simply overprovision the WAN bandwidth, hoping to obtain adequate QoS. Some enterprises upgrade to an MPLS VPN believing that classes of service in MPLS will solve QoS problems. A holistic approach for enabling voice and video over the IP network includes network assessment, bandwidth analysis and planning, implementing control procedures, and monitoring and reporting; properly employing these four practices may allow organizations to postpone costly bandwidth upgrades and ensure that the network will provide the QoS required for voice and video over IP.

## Introduction

The biggest technical challenge in transitioning from traditional circuit-switched voice and video systems to the new, more economical voice and video over IP packet-switched technologies is obtaining adequate quality of service (QoS) over the wide area network. Quality of service is the capability built into the network to guarantee that information traverses the network in a timely manner. Most existing data networks were designed for bursty applications that are delay-insensitive, meaning that if a data packet arrives within a reasonable amount of time, both the application and the user are satisfied.

Voice and video data, on the other hand, are very sensitive to delay; if a packet arrives more than approximately 200 milliseconds (ms) after it is transmitted, the packet is worthless as a carrier of real-time communication because it will arrive too late to be used in the conversation or video image. Consequently, networks carrying IP voice and video must be designed and configured properly to ensure that real-time packets traverse the network efficiently.



*Figure 1: When numerous uncontrolled data-intensive applications compete for scarce WAN bandwidth, mission-critical and/or real-time applications like voice and video may perform poorly.*

The challenge of obtaining adequate quality of service is exacerbated when a data packet must traverse the WAN. Typical local-area networks (LANs) run at 10 Mbps, 100 Mbps, and some even have bandwidth of a gigabit per second (1000 Mbps) and higher. However, because bandwidth over the WAN is significantly more expensive than over the LAN, many wide-area networks operate at T1 speeds (1.45 Mbps) and slower, creating a huge bottleneck at the LAN/WAN interface. For normal data packets like email, Web browsing, client-server programs, and a host of other applications, this LAN/WAN bottleneck is a nuisance, but not an application killer because these applications can withstand delay and still function satisfactorily. However, when voice and video packets must compete with regular data packets for transmission over a bandwidth-constrained WAN, the voice and video applications may be rendered useless unless steps are taken to insure voice and video QoS.

## Important QoS Parameters

For IP networks supporting voice, video, and data applications, the quality of service objective is to preserve both the mission-critical data in the presence of voice and video and to preserve the voice and video quality in the presence of bursty data traffic. Network quality of service is evaluated by measuring four key parameters: bandwidth, end-to-end delay, jitter, and packet loss.

- **Bandwidth:** The average number of bits per second that can travel successfully through the network.
- **End-to-end delay:** The average time it takes for a packet to traverse the network from a sending device to a receiving device.
- **Jitter:** The variation in end-to-end delay of sequentially transmitted packets.
- **Packet loss:** The percent of transmitted packets that never reach the intended destination.

For IP voice and video communications systems to work properly, the bandwidth should be as large as economically possible while the end-to-end delay, jitter, and packet loss should be minimized. Lower end-to-end delay leads to a more satisfactory, natural communications experience, while large delay values lead to unnatural conversations with long pauses between phrases or sentences. Target values for delay, jitter, and packet loss are < 200 ms, < 50 ms, and < 1% respectively.

Organizations wishing to maintain management control of their networks when adding new applications, including voice over IP (VoIP) and video over IP, usually follow some variation of a four-step process: 1) perform an assessment of the current network, 2) determine the bandwidth and performance characteristics required for the new applications, 3) implement control procedures, and 4) monitor and report network behavior.

## Assessing the Current Network

Conceptually, performing a network assessment is straightforward. A network administrator needs to develop a network baseline by identifying which applications are currently

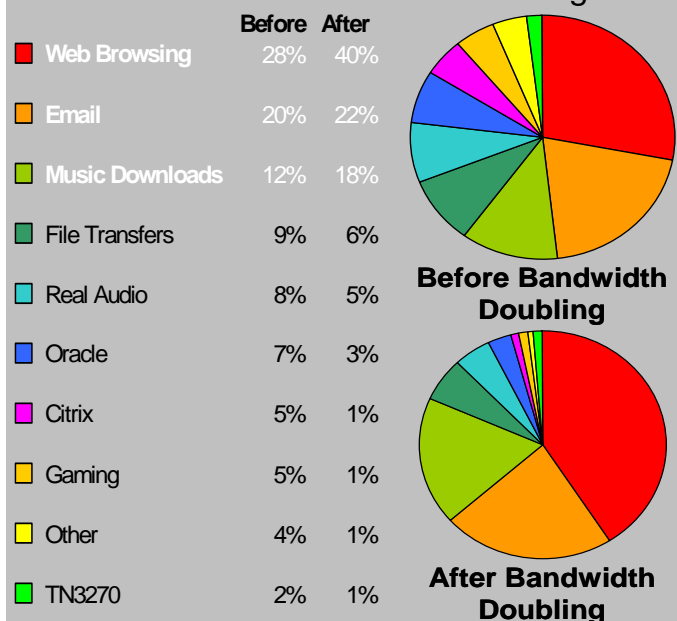
## You Can't Control What You Can't See

One of the most common mechanisms network administrators use to improve wide area network performance is to increase the WAN bandwidth. Although this can help, it is a short-term solution to the network performance issue. Most popular applications including web browsing, email, and music downloads rely on TCP's slow start algorithm to steadily increase their own bandwidth utilization until there is a problem. Consequently, these types of applications, unrestricted, quickly consume all available network bandwidth, leaving VoIP, video, and other time sensitive applications suffering.

In a network utilization study done by IDC, they found that most organizations unknowingly allow the least important applications to control the network. As shown in the figure below, doubling the WAN bandwidth haphazardly allocates increased amounts to the most demanding, bandwidth-hungry applications which are probably not the most urgent and critical applications the organization wants running over their network.

Armed with a knowledge of each application's resource consumption and coupled with an understanding of the organization's priorities for bandwidth utilization, network administrators become empowered to control network performance.

### The Effect of Bandwidth Doubling



Source: IDC.

running on the network, when they run, how much bandwidth they use, and if they are performing satisfactorily. A variety of tools can be used for beginning the assessment including packet sniffers and network probes. Sniffers and probes come with a wide diversity of capabilities, and they can be either hardware- or software-based. Most of these devices use a standard monitoring specification known as Remote Monitoring, or RMON, that enables various network monitors and console systems to exchange network-monitoring data. Using RMON, network administrators are able to monitor numerous network parameters including the number of packets dropped, packets sent and received by a particular device, packet collisions, which devices send the most data, which devices receive the most data, counters for packet size, and so forth. In addition to sniffers and probes, many routers allow network administrators to measure delay and jitter on network segments between the routers.

However, relying only on low-level protocol analyzers, such as those that look at IP packet headers to classify traffic gives no indication of which applications are actually putting data on the network and may preclude the discovery of some significant traffic trends. Moving up to the application level and differentiating one application's traffic from another's permits individual application behavior to be distinguished. For example, Skype<sup>1</sup>, Ares<sup>2</sup>, BitTorrent<sup>3</sup>, viruses, worms, and a number of other non-HTTP-based applications that run over the network consume significant bandwidth while hiding in an HTTP tunnel using port 80 to traverse the firewall. IT administrators facing network capacity problems would have difficulty distinguishing legitimate Web browsing traffic from illegitimate peer-to-peer (P2P) traffic without more sophisticated monitoring tools that examine network behavior at the application level. This is particularly true for videoconferencing and VoIP, which begin using well-defined ports to establish call setup, but branch out to use several dynamically defined ports (up to eight or more depending upon the video endpoint used) to exchange media. Knowing which applications are running, what low-level network resources they use, and how they behave is critical to establishing a baseline and maintaining network control.

## VoIP versus VoIP

As enterprises embrace IP telephony by deploying IP PBXs, they are beginning to see the need to protect sanctioned, enterprise critical, voice applications from not only the data traffic, but from other unsanctioned VoIP applications arising from instant messaging programs such as AOL IM, Yahoo!, and MSN Messenger. Furthermore, there is a possibility that future P2P applications may masquerade as VoIP traffic in order to traverse NATs and firewalls and to receive priority treatment when traversing the WAN.

Instant messaging VoIP applications are of two main types:

1. Those that use standard VoIP protocols like SIP and RTP. The issue here is that the edge router may be programmed to prioritize WAN traversal for any voice packet using these protocols; consequently, the enterprise voice system may have intermittent quality problems because it must compete for bandwidth with individuals using rogue VoIP applications, like AOL or MSN Messenger, who knowingly or unknowingly may be using high priority enterprise bandwidth for communicating with family and friends.
2. Those that use proprietary VoIP protocols. The challenge with these applications is that they are typically more difficult to distinguish from regular data applications; therefore, a packet prioritization scheme based on marking a certain type of traffic, i.e. RTP, would fail. This may be fine except in the case where the proprietary VoIP application is sanctioned and is considered a business critical application.

---

<sup>1</sup> Skype is a free P2P Internet telephony technology.

<sup>2</sup> Ares is a free peer to peer file sharing program that enables users to share any digital file including images, audio, video, software, documents, etc.

<sup>3</sup> BitTorrent is a free P2P file sharing program capable of "swarming" downloads across unreliable networks (like the Internet). It trades pieces of a file you have with pieces your peers have.

Besides putting unsanctioned voice traffic on the network, IM clients often support a basic set of features that includes text chat, file transfer, and videoconferencing. Text chat is not normally a big network burden; however, large file transfers and videoconferencing can put significant amounts traffic on the network. File transfers can clobber a network link if not controlled, and a videoconference stream may inadvertently be put in the priority VoIP queue, thrashing corporate voice conversations because it is identified as an application using SIP or the RTP protocol.

### Advanced Network Monitoring

Advanced network monitoring tools are available from several vendors which will allow enterprises to distinguish and profile the applications running on the network. Of these Packeteer, Inc., offers one of the most advanced systems. These tools are particularly useful to network administrators because they have the capability to *autodiscover* many applications running on the network, and they can do so to a remarkable degree of detail. For example, they are able to distinguish between various kinds of Oracle databases being used by an application, which protocols (SIP or H.323) and which compression algorithms are used in voice and video over IP applications, which Citrix applications are being used, along with the high- and low-level network resource usage of each. By having this kind of detail, network administrators get a clear picture of how much bandwidth each application actually uses and whether it is contributing to network performance problems.

Once the different applications running on the network have been identified, network administrators and application users need to distinguish between traffic types, classifying them to identify which applications are to be prioritized according to an organization's business needs. For example, voice and video traffic need a high priority in order to maintain the real-time nature of an interactive conversation.

Most network administrators classify traffic according to one of two classification schemas: application-based or location-based. Application-based schemes are often used when the network administrator is considering a single location and wants to identify and classify traffic over the network in that location by application. Location-based classification schemes are usually used at central data centers where connections from a number of branch locations meet; location-based schemes allow administrators to

## The World of Voice Protocols

Rogue voice applications may be a problem where you work today. Could your network tools distinguish between the various types of voice protocols and control them? Below is a list of the more common VoIP protocols and sub-protocols.

CiscoCTI	Cisco Computer Telephony Interface
Dialpad	Dialpad Internet Telephone service group
Dialpad-Ctrl	Dialpad Internet Telephone — control traffic
Dialpad-Stream	Dialpad Internet Telephone — RTP stream
H.323	Internet telephony standard service group
H.323-GKD	H.323 Gatekeeper Discovery
H.323-H.245	H.323 call control
H.323-Q.931	H.323 call setup
H.323-RAS	H.323 Gatekeeper Control
I-Phone	Vocaltec Internet telephone service
Megaco	Media Gateway Control (H.248)
Megaco-Text	Media Gateway Control (H.248) Text
Megaco-Bin	Media Gateway Control (H.248) Binary
MGCP	Media Gateway Control Protocol
MGCP-Gateway	Media Gateway Control Protocol Gateway
MGCP-CallAgent	Media Gateway Control Protocol CallAgent
MGCP-KpAlive	Media Gateway Control Protocol KeepAlive Connection
Net2Phone	Net2Phone CommCenter
Net2Phone-TCP	Net2Phone Call Setup and Control
Net2Phone-UDP	Net2Phone Internet Phone Calls
RTCP-B	Real-Time Control Protocol (Broadcast)
RTCP-I	Real-Time Control Protocol (Interactive)
SIP	Session Initiation Protocol
SIP-UDP	Session Initiation Protocol - UDP
SIP-TCP	Session Initiation Protocol – TCP
Skinny	Cisco's Skinny Client Control Protocol (SCCP)
Skype	Skype P2P Telephony Application
SkypeCommand	Skype Command
SkypeData	Skype Data
VDOPhone	Service group for Internet telephone service group
VDOPhone-a	Internet telephone application — TCP port 1
VDOPhone-b	Internet telephone application — TCP port 2
VDOPhone-UDP	VDOPhone real-time media

classify traffic by user group, subnet, IP address, or host lists, and then subdivide these by application types.

## Bandwidth Analysis and Planning

Network traffic discovery and classification reveal what is on the network and how important each application actually running on the network is to the organization; bandwidth analysis provides network administrators with the critical data necessary for creating strategies to manage bandwidth, improve network topologies, and plan for future capacity. Two primary types of analysis are utilization analysis and performance analysis.

### Bandwidth Utilization Analysis

Utilization analysis identifies which applications and individuals or host machines are using the bandwidth. Utilization analysis will reveal

- If the right applications are getting sufficient access to bandwidth and if the wrong ones are getting too much access,
- If the organization is receiving all the bandwidth they have contracted for from their service provider, and
- If the organization has enough bandwidth for the applications that need to run over the network and how much bandwidth is really needed for critical applications if the network were managed more effectively.

One key measurement is the top bandwidth users by classification type. This data is used to determine if the right applications are getting an appropriate share of the bandwidth and if the wrong ones consume too much. For example, a network administrator may determine that the web traffic class consumes 75% of the network bandwidth and may want to divide the web traffic into critical and casual classes and prioritize each accordingly. Conversely, administrators will not want voice and video traffic to bloat, consuming all the network bandwidth and causing other important applications to perform poorly. In this instance, the administrator would consider putting limits on the amount of bandwidth consumed by VoIP and video applications so that all enterprise critical applications are able to perform satisfactorily.

Another useful measurement is determining which applications transmit the most data (talkers) and which receive the most data (listeners). If one of the top talkers is a Web page on the organization Web site, appropriate steps can be taken; if a top listener is pointed to music downloads or streaming video, steps can be taken to curb bandwidth for these applications.

Using Bandwidth Utilization Data	
If You See...	Then Consider...
Very high peak rates with low average rates	Control features to smooth bursts and increase throughput.
Low usage rates rarely or never approach capacity and applications perform well	Whether you are paying for more bandwidth than you need. You may be able to reduce bandwidth expenses.
Frequent peaks to a level that is consistent, but lower than your supposed capacity	Verifying whether you are getting all that you are paying for. Consider using control features to make the most efficient use of the bandwidth you have.
Consistently high usage rates that remain near capacity	First, adding devices with control and/or compression features to make the most efficient use of your bandwidth (such as Packeteer's PacketShaper Xpress model). Afterward, if the problem still exists, consider buying more bandwidth.

Table 1: Guidelines for using bandwidth utilization data to make network control and capacity decisions.

Monitoring peak and average bandwidth utilization rates provides a very useful glimpse into the behavior of all traffic classes. One key is to ensure that the measurement time is short enough to really capture application behavior and that peak utilization is detected. Routers and other network devices can give an artificial sense of security when they offer metrics for average network usage over lengthy periods of time. Devices tracking peak usage and those which use more frequent measurement intervals can highlight a hidden capacity problem. For example, an average rate can mislead an administrator into thinking that usage never approaches total network capacity while a peak-rate line can reveal frequent spikes over the same time period. Alternatively, a peak-rate line can reveal the opposite – perhaps the organization uses less bandwidth than it is paying for.

## Application Performance Analysis

Application performance analysis quantifies what has often been subjective, anecdotal information about measuring a user's perception of performance – response time – once the Enter key has been pressed or a mouse button clicked. This type of analysis provides network administrators a mechanism to compare both actual and anticipated application performance and a way to verify service-level compliance. It also gives administrators the ability to quantify and validate performance claims that can help justify new equipment and assure that voice and video applications function within specified QoS target values.

<b>Application Performance Analysis Parameters</b>		
<b>Metric</b>	<b>Description</b>	<b>Uses</b>
<b>Total Delay</b>	<p>The number of milliseconds a transaction requires, beginning with a client's request and ending upon receipt of the response.</p> <p>This is what most people mean when they say "response time," corresponding to the end user's view of the time it takes for a transaction to complete.</p>	<p>Verify users' perceptions of slow performance.</p> <p>Spot good or bad trends by tracking historical averages over time.</p>
<b>Network Delay</b>	<p>The number of milliseconds spent in transit when two devices exchange data.</p> <p>If a transaction requires a large quantity of data to be transferred, it is divided and sent in multiple packets. Network delay includes the transit time for all packets involved in a request-response transaction. The amount of time the server uses for processing a request is not included.</p>	<p>Determine if a slowdown is due to a problem in transit (as opposed to an overloaded server, for example).</p> <p>This parameter is a critical measurement for voice and video over IP.</p> <p>Determine if packet-shaping control features can resolve a particular performance slowdown or prevent future occurrences.</p>
<b>Server Delay</b>	<p>The number of milliseconds the server uses to process a client's request after it receives all required data.</p> <p>The server delay is the time after the server receives the last request packet and before it sends the first packet of response (not receipt acknowledgment, but actual response content). This is the time the server takes to process the client's request.</p>	<p>Determine if servers are to blame for a slowdown.</p> <p>For voice and video, this would include multipoint control units and gatekeepers.</p>
<b>Round Trip Time (RTT)</b>	<p>The number of milliseconds spent in transit when a client and server exchange one small packet. Even if a transaction's data is split into multiple packets, RTT includes only one round trip of a single packet between client and server.</p>	<p>Determine if a major network delay is a result of large transactions or a slow network. For example, if users suddenly begin using the file-sharing features with extremely large files, a slowdown might be caused by the quantity of information being transmitted rather than a change in network status. In this case, the network delay would grow, but the RTT would not.</p>

*Table 2: Application performance analysis parameters and how to use them to detect network performance problems.*

Application performance analysis consists of measuring the delays that occur in both the network and any servers, hosts, endpoints, or other devices used by an application. The best performance analyzers can track and associate delay with a particular application or a sub-process within an application (i.e. Citrix, VoIP, or video applications). Typical parameters to review are shown in the following table:

## Bandwidth Planning for Video and Voice Applications

When planning bandwidth for VoIP or video applications, it is important to understand the true amount of bandwidth required. These applications typically consist of several components: audio, video, control, and protocol overhead. The bandwidth specified when configuring the application may or may not account for all of the data used by the application. The actual bandwidth consumed is often greater when all other components are considered. The only reliable method for determining accurate application capacity requirements is by measuring them with network test equipment.

Bandwidth for VoIP is dependent upon which compression algorithm is used. Bandwidth consumption for several common VoIP codecs is shown in the table below. Note that these values include packet header protocol overhead, but an additional 10% may be required for traversing the WAN using ATM or frame relay.

Bandwidth Requirements for Several Common VoIP Compression Algorithms <sup>4</sup>		
Codec	Bit Rate (Kbps)	Nominal Ethernet Bandwidth (Kbps)
G.711	64	87.2
G.729	8	31.2
G.723.1	6.3	21.9
G.723.1	5.3	20.8
G.726	32	55.2
G.726	24	47.2
G.728	16	31.5

*Table 3: Different VoIP compression algorithms consume different amounts of bandwidth.*

For video over IP, a conservative rule of thumb is to assume the nominal value plus 10% for LAN overhead and an additional 10% for WAN overhead (i.e. the ATM "cell tax"). For example, to support a 384 Kbps video call over the WAN, provide at least 460 Kbps of bandwidth; to support a 512 Kbps video call, provide at least 615Kbps of bandwidth.

Network managers must also be sensitive the fact that different types of VoIP calls may use different bandwidth. For example, in an IP PBX environment, the network manager may set the system up for a low bandwidth protocol, like G.723, when dialing between IP phones on the network. However, a PSTN call entering the network through a gateway may use the 64 Kbps G.711 protocol. Network managers must account for the difference in the possible protocol bandwidth allowed on the network when planning a VoIP implementation.

In an 'unloaded' environment (i.e. the voice or video application is not competing with other data applications), network sizing is simple -- account for the 10% protocol overhead plus another 10% for WAN inefficiencies. However, if the WAN is shared with other applications, then allowances are necessary to account for the overhead associated with operating in a shared environment. It is recommended that voice and video traffic consume no more than 50% to 70% of the network capacity in an environment that supports multiple application types. This means that at least 30% of network capacity is reserved for data applications. Even if the expected bandwidth of the competing data

<sup>4</sup> Source: Cisco Systems. See

[http://www.cisco.com/en/US/tech/tk652/tk698/technologies\\_tech\\_note09186a0080094ae2.shtml#topic1](http://www.cisco.com/en/US/tech/tk652/tk698/technologies_tech_note09186a0080094ae2.shtml#topic1).

applications is less than 30%, this allocation level is recommended to minimize potential queuing delays and any resulting audio or video degradation.

IT administrators can take some comfort when adding bandwidth for video applications because new capacity actually creates more bandwidth for all applications running across the network. In a video application, the video compression engine sends three types of frames: 1) key frames that contain all the information in a particular image, p frames that contain only what changed in the video image from frame to frame, and b frames that are like p frames, but which are built from both key frames or p frames. Because the video bandwidth required is sized based on when the large amount of data in a key frame is sent, and because most of the time, smaller p and b frames are used, more bandwidth is actually available to all other applications except when a key frame is sent.

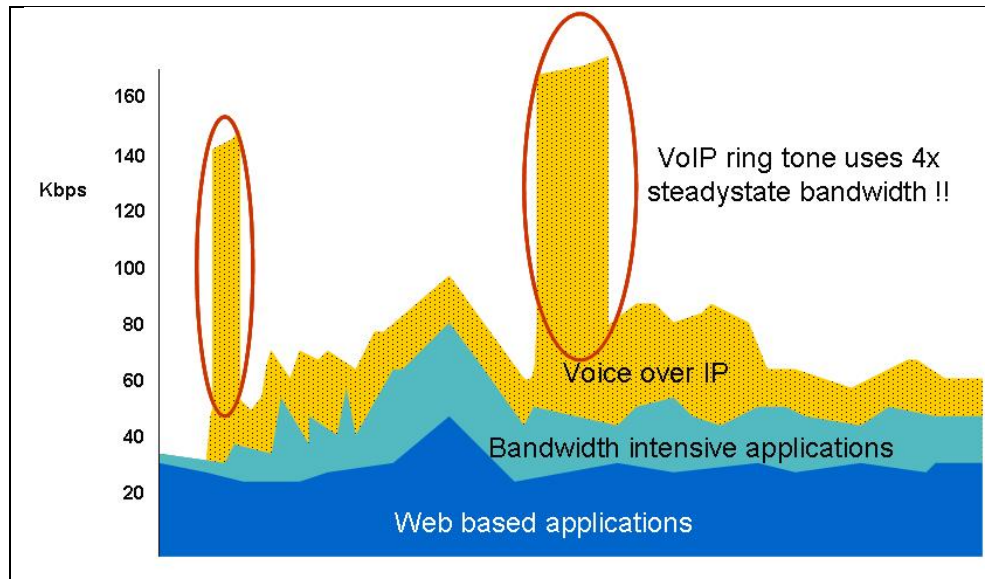


Figure 2: Voice and video over IP call set up rates may be several times the steady state value.

## Implementing Control Procedures

There are a number of control-based measures to provide adequate quality of service. One key element in any control strategy is defining an overall concept of network operations to provide a business foundation upon which administrators can build precise definitions of the features and performance desired in the network. Failing to develop an operational concept for network management can lead to network instability or network management by exception due to shifting end user demands. Administrators will often need to unify inconsistent performance expectations from the various application users.

## Packet Marking and Queuing

On networks that carry data as well as voice and video, some network owners put QoS mechanisms in place at the IP level to ensure proper treatment of real-time voice and video packets. Mechanisms at this level allow individual packets from high-priority applications to be identified and treated preferentially in the network transport mechanisms using packet-marking and queuing.

### IP Precedence and DiffServ

IP precedence and DiffServ rely on similar mechanisms for marking packets to provide quality of service. Both of these traffic-marking schemes modify certain bits in the data packet header. Upon arrival at an IP precedence or DiffServ-enabled router or switch, packets with the header bits set appropriately are given priority queuing and transmission.

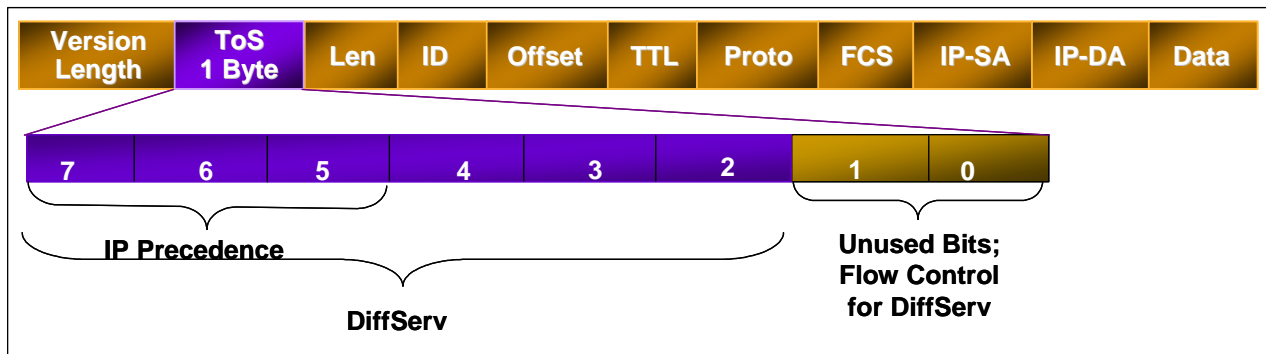


Figure 3: IP Precedence and DiffServ packet classification schemes.

In the IP packet header, bits 9 -11 are reserved as IP precedence bits; these three bits support eight different classifications ranging from seven at the highest priority to zero at the lowest priority<sup>5</sup>. IP precedence is not consistently implemented from vendor to vendor; consequently, care must be taken to assure that networks with mixed vendor equipment function properly.

DiffServ uses IP packet header bits 9 -16 to help routers prioritize which packets to send more rapidly and which to drop in the event of congestion. DiffServ is designed to have broader classification flexibility than IP precedence with 64<sup>6</sup> different classifications available.

With either IP precedence or DiffServ, the network must be designed so that the scheme is consistently implemented within the entire network. Some service providers are beginning to provide classes of service with differing levels of service quality based on the DiffServ classification.

In a network controlled by packet marking, voice packets would be given the highest priority since they are very sensitive to delay and jitter, even though voice is not particularly bandwidth-intensive. Video packets are given a slightly lower priority while email and Web surfing packets, for example, are given the lowest priority.

### Queuing

Queuing occurs in routers and switches. Different queues or buffers are established for the different packet-marking schemes. One of the queues, for example, might be established for delay- and drop-sensitive information like voice and video data. Voice and video packets marked with certain IP precedence or DiffServ values will be placed in these high-priority queues. Packets in the high priority queue are always transmitted first, followed by any packets in lower-priority queues. Network administrators establish the queues and priorities across their networks, and networking infrastructure companies like Cisco have recommendations for how to do this.

### Traffic Shaping

Queuing-based solutions have a number of drawbacks. Of these, one of the most significant is the lack of any feedback mechanism for determining how applications are competing for bandwidth. Consequently, data traffic for applications on networks with queuing mechanisms in place cyclically ramp up and back off transmission rates based upon packets being discarded. This causes chunks of data that accumulate at the LAN/WAN interface where speed conversion occurs.

<sup>5</sup> Organizations running Cisco equipment exclusively should set their IP precedence bits as follows: for voice over IP, set the header bit to 5; for IP videoconferencing, set the header bit to 4; for all other data applications, set the bits between 3 and 0 as needed. Priorities 7 and 6 should be reserved for routing protocol packets.

<sup>6</sup> Although DiffServ uses the "ToS" octet in the IP packet header consisting of bits 9 – 16, the last two bits (15 and 16) are currently unused; hence, there are really only 6 bits used which allows 64 different classifications. For more information, please refer to <http://www.qosforum.com/docs/faq/>.

One way to eliminate these chunks of data is by using a special technology called TCP Rate Control. Developed by Packeteer, TCP Rate Control paces or smoothes network data flows by detecting a remote user's access speed, factoring in network latency, and correlating this data with other rate and priority policies applied to various applications. Rather than queuing data in a switch or router and metering it out at the appropriate rate, TCP Rate Control induces the sending applications to slow down or speed up, thus sending data just-in-time. By shaping application traffic into optimally sized and timed packets, TCP Rate Control can improve network efficiency, increase throughput, and deliver more consistent, predictable, and prompt response times.

TCP Rate Control uses three methods to manage the rate of application transmission:

- Real-time flow speed detection,
- Metering of TCP acknowledgments going back to the sender
- Modification of the TCP-advertised window sizes sent to the sender

It works by detecting and continuously monitoring the connection speed of servers and clients involved in a network transaction, adjusting bandwidth management strategies in real-time while conditions are changing in the network. A sliding window flow-control mechanism controls TCP packet throughput over the network. As the receiving device acknowledges initial receipt of data, it advertises how much data it can receive, called its window size. The sending device can transmit multiple packets, up to the recipient's window size, before it stops and waits for an acknowledgment. The sender fills the pipe, waits for an acknowledgment, and fills the pipe again. TCP Rate Control monitors the acknowledgments from the receiver as well as the data window sizes, modifying them in real-time to smooth out data bursts and controlling when the sending application transmits information and how much data is sent.

Most voice and video applications use UDP rather than TCP for transmitting real-time communications data. Unlike TCP, UDP sends data to a recipient without establishing a connection, and UDP does not attempt to verify that the data arrived intact. Therefore, UDP is referred to as an unreliable, connectionless protocol. The services that UDP provides are minimal — port number multiplexing and an optional checksum error-checking process — so UDP requires less processing time, and lower bandwidth overhead than TCP. This allows UDP packets to traverse the network more rapidly, which is a desirable characteristic for voice and video applications.

However, because UDP doesn't manage the end-to-end connection, it does not get feedback regarding transmission conditions; consequently, applications transmitting UDP packets cannot prevent or adapt to congestion. Therefore, UDP can end up contributing significantly to an overabundance of traffic, impacting all traffic on the network. This may cause latency-sensitive flows, such as voice and video over IP, to be so delayed as to be useless. In these instances the voice or video application may still continue to transmit data, oblivious to the fact it is contributing to the delay problem.

## **Partitioning**

The rate of UDP transmissions over the WAN can be controlled by a process called partitioning. Partitioning is a special case of rate control in which specific amounts of bandwidth are set aside for the most important classes of traffic. Partitioning can also be overlaid on top of TCP rate control for TCP based applications.

Partitioning is administered by a packet-shaping device that examines all packets traversing the network. By identifying a particular application as a member of a particular partition class, the "packet shaper" is able to control how much bandwidth each application or class of applications uses, and it can ensure that a particular partition always gets sufficient bandwidth. When the bandwidth within a particular partition is not fully utilized, the excess bandwidth can be reallocated to partitions serving other important applications. Administrators can also specify UDP flow maximums so that one large flow -- videoconferencing or streaming video for example -- does not consume all bandwidth on the network. In environments where multiple video devices are deployed, organizations will want to couple partitioning with a SIP proxy's or H.323 gatekeeper's call bandwidth controls to manage how many simultaneous

video calls can be placed. Using a gatekeeper will avoid overloading a particular video partition, which would cause all of the video applications running in that partition to fail.

## Facilitating an MPLS WAN

One of the fastest growing markets for network service providers is transitioning customers from Frame Relay and ATM to fully-meshed MPLS virtual private networks (VPNs). While MPLS promises additional control, particularly for delay- and drop-sensitive traffic like voice and video over IP, the enterprise must still do significant planning and control on the LAN side of the network in order to gain the benefits an MPLS network can provide.

A typical MPLS network will have four or five different classes of service, with priority being given to the packets in the highest service class. The MPLS service class to place a packet into is determined using IP precedence or DiffServ setting in the IP packet header (the TOS byte). Because carrier networks are designed to transmit packets very rapidly, the carrier typically works with the enterprise to assure that packet priorities are already marked by the enterprise CPE router before traversing the last mile between the enterprise edge router and the carrier edge router. If packets are not marked with the right priority before being sent to the carrier MPLS network, they may be placed into the wrong class of service when traversing the WAN. Consequently, the enterprise must enable some sort of packet marking rules prior to sending the packets to the MPLS network.

A variety of means are available for the enterprise to mark packets; four of the most common are

1. The network "trusts" an end device with a particular IP address to set its own priority. In this scenario, the end device sets its own priority and the CPE router does not change it before transmitting the packet to the MPLS VPN. Often this is controlled by an access control list in the router so that only certain IP addresses are trusted to set their own packet priorities. An example would be a videoconferencing unit with a fixed IP address. The network administrator would know that traffic coming from the IP address of that video endpoint would be high priority video.
2. Looking at the physical port number a given packet came from and marking the packet accordingly. A VoIP phone may have a specific physical port number, and any packets from this physical port will be marked with the specified high priority.
3. Checking the protocol. Some, but not all, routers can look deep into a packet and determine if it is carrying data with a known protocol such as SIP or H.323. If the router recognizes that a particular packet is carrying data containing one of these protocols, it will mark the packet with a high priority.
4. Checking the logical port. Several legitimate voice and video applications use specified TCP and/or UDP ports. If the router detects traffic coming from one of these specified logical ports, it can prioritize the packet accordingly.

Most enterprises run hundreds and sometimes thousands of applications over the WAN, and some of these applications will be required to share MPLS service levels. Unless applications are prioritized properly and controlled, there may be contention between applications within a given class of service.

The enterprise edge router can do packet marking; however, looking deep into a packet payload to determine what type of protocol the packet is carrying can put a big burden on the router's processing capability. Furthermore, routers are not usually designed to recognize packets at the application level; consequently, packets from rogue voice or video applications using a protocol known by the router, or packets from applications that hop to a high priority logical port, will not be detected. These rogue packets may then be given a high priority, allowing them to complete with legitimate applications for MPLS priority transmission.

In considering how to choose the right MPLS class of service for a given application, it is necessary to uniquely identify the application and measure its performance. Monitoring the application packet flow allows an analysis of the application over a period of days to ensure that all characteristics of the

application are accounted for. A second, equally important measurement is the application response time that leads to satisfactory performance. Measuring the Round Trip Time, which is the average number of milliseconds spent in transit when a client and server exchange SYN (synchronize sequence numbers flag) and its corresponding ACK (acknowledge flag), allows a network manager to know an application is performing satisfactorily. Combining these measurements provides insight into which MPLS class of service a given application should be placed.

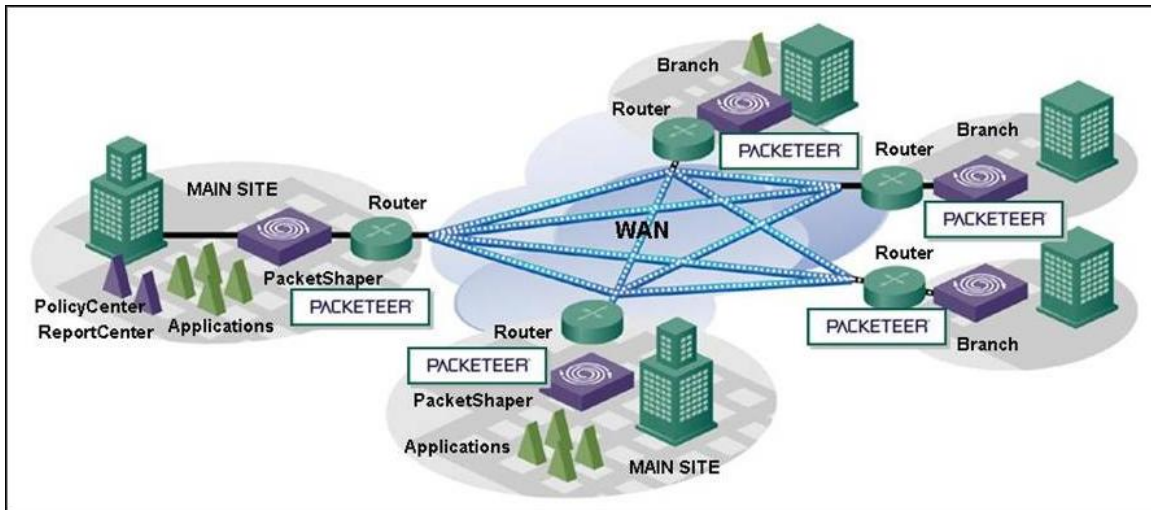


Figure 4: PacketShapers are situated on the LAN side of the enterprise edge router.

One way to measure application performance, take the burden off the edge router, and provide additional control is to use traffic shaping technology, like that offered by Packeteer. These devices, situated before the CPE router, identify packets at the application layer, control TCP traffic flows and UDP application partitions, and align all packet priorities to MPLS service classes based on application type and intended priority to provide a consistent and predictable user experience in a manageable, accountable, and efficient manner.

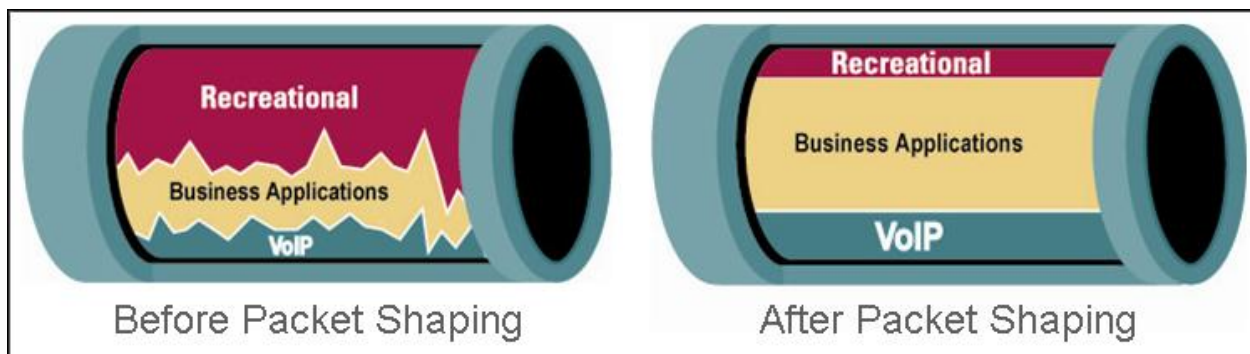


Figure 5: Application flow over the MPLS WAN links before and after packet shaping.

### Adaptive Response Control

Automatic problem detection, notification, and flow control are extremely useful to network managers; however, problem detection and notification can occur during breaks and at other times when network administrators are away. To help combat this problem, a mechanism called adaptive response control can be used. Adaptive response automatically monitors for network anomalies; once an anomaly is found, the control device performs any corrective actions specified in advance by the network administrator to resolve the particular problem.

As an example of Packeteer's adaptive response control, consider the following. Suppose SAP is a mission critical network application running over an enterprise MPLS WAN supporting four classes of service. Under normal conditions the network has been designed to support a service level such that 92% of SAP transactions complete within 1.5 seconds. To enable this service goal, SAP packets are normally marked to run in the third MPLS service class and a SAP partition has been defined to guarantee that SAP packets have at least 15% capacity of the WAN link.

Now, suppose that FTP bursts occur at a time when the network manager is away (i.e. 4:00 a.m.), and that these bursts begin to impact the SAP application's response time and service-level compliance. Adaptive response control is able to automatically adjust the partitions and policies' definitions to solve the problem in the absence of the network manager. In this example, the manager may have predefined a solution so that the PacketShaper adjusts the SAP partition to 18% and marks the SAP packets so that they traverse the MPLS network in the next higher service class. Adaptive response control allows network managers to mitigate problems, keeping users happy and application performance within specification, until the network administrator can analyze and correct an unforeseen network traffic problem.

## Monitoring and Reporting

Network environments change constantly, and the capacity requirements of the business applications running through them are in flux. As a result, administrators must monitor and report on network performance to ensure that bandwidth is utilized appropriately. A variety of tools are available to provide insight into real-time and historical performance, load, efficiency, and the effectiveness of the organization's network management strategy.

The key parameters to measure and report for VoIP and video are those previously mentioned: bandwidth capacity and utilization, delay, jitter, and packet loss. These must be measured on a per-endpoint level to assure proper performance system-wide.

## Avoiding the SLA Trap

Network service providers offer some very compelling service level agreements. We are aware of providers that offer 100% packet delivery, 100% availability, and very low latency, jitter, and packet loss values. Unfortunately, as one drills down into these service agreements, we find that they usually only extend from the carrier edge router ingress point to carrier edge router egress point: they do not cover the enterprise CPE or the local area network where most of the delay and drop problems occur. Furthermore, latency, jitter, and packet loss are values averaged over a time period, and may not reflect instantaneous maximum values that can occur in the core. Consequently, caution must be exercised when considering what the true service level will actually be for any given application because most of the delay, jitter, and packet loss will occur over the last mile, from the CPE router to the carrier edge.

One way to evaluate if your carrier's SLA is sufficient is as follows<sup>7</sup>:

1. Determine the required packet loss rate for high and low priority traffic, based on the applications you must support and the performance requirements of those applications.
2. Diagram your network and determine the number of router hops in the access network(s) as packets move from one enterprise location to another.
3. Next, take the carrier's packet loss rate SLA for the core network, and degrade it by an order of magnitude to account for the extensive averaging used. This number is the best possible SLA value for the busy hour periods of your business, assuming your busy hour periods coincide with the majority of users of the WAN core.
4. If this number is not sufficient for your SLA, search for another carrier.

---

<sup>7</sup> Method source: *Business Communications Review*, October 2004, P. 17.

To more accurately determine if an SLA is adequate, it is possible to measure the performance of application traffic from within the LAN, over the WAN, back on the destination LAN, and back to the originating application. Three parameter measurements are important:

- Response Time Management (leading directly to Service Level Compliance)
- Per application performance, peak and average throughput.
- Network Efficiency

Response-time management (RTM) helps network managers understand the compliance of an application within a class of service. RTM can be used to substantiate user complaints and evaluate performance problems before they disrupt the business. With a mechanism to compare actual and anticipated performance, service-level agreements become measurable and enforceable, and data is available to help quantify and validate performance claims for justifying new equipment and to assist with future IT planning.

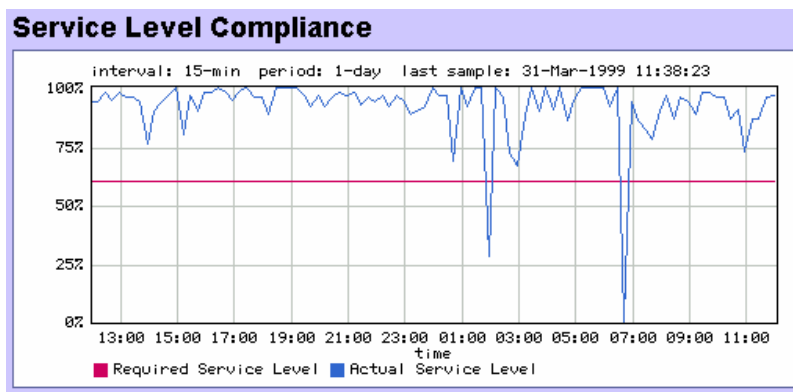


Figure 6: RTM can be used to validate service level compliance.

Each response-time measurement can be broken into network delay (time spent in transit) and server delay (time the server used to process the request). By understanding which element of the network (CPE routers, MPLS WAN, etc.) contributes to delay it is possible to set acceptability standards and track whether performance adheres to them.

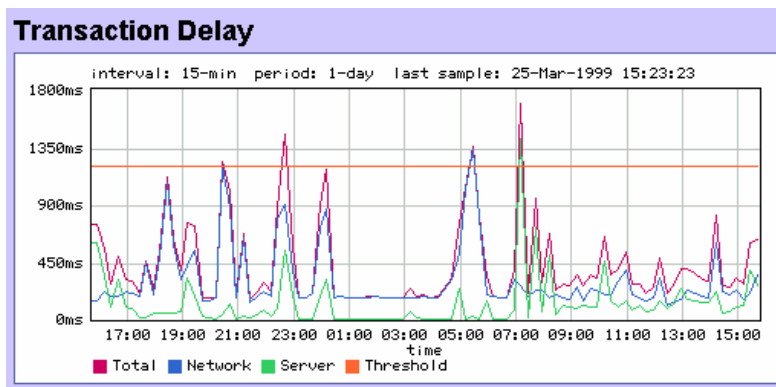


Figure 7: Transaction delay shows an application's (traffic class') average response times along with detailed measurements of overall, network, and server delay.

SLA reporting should display usage over time for different applications, branch offices, the entire link, or other criteria. Viewing average bandwidth rates over long measurement periods can give an artificial sense of security leading network managers to think that usage never approaches capacity. It is essential to track *peak* usage using more frequent time intervals than a service provider will normally report. Peak-

rate line measurements may reveal spikes that use the entire link, resulting in traffic not meeting the desired performance. It can also reveal the opposite — perhaps an organization using less bandwidth than it's paying for, allowing CoS sizes to be cost optimized.

Retransmissions, traffic that must traverse the network multiple times for successful arrival, should optimally be as close to zero as possible. Retransmissions spike when router queues deepen as a result of congestion that causes packet loss; it also spikes as latency increases the frequency of time-outs, and in some cases, retransmission spikes when a busy IP network behaves precisely as designed under heavy loads. Increasing bandwidth to support a high rate of retransmissions may be a costly, wasteful situation.

Tracking the rates of throughput and retransmissions for any particular application, allows a calculation of the percentage of wasted bandwidth and helps managers determine if additional bandwidth is really required or if the applications need additional control. If information is being collected on a stateful basis (on a per application, per user basis) it is possible to gain insight into the efficiency not only of the link as a whole, but by application, protocol, subnet, user, server, or web destination.

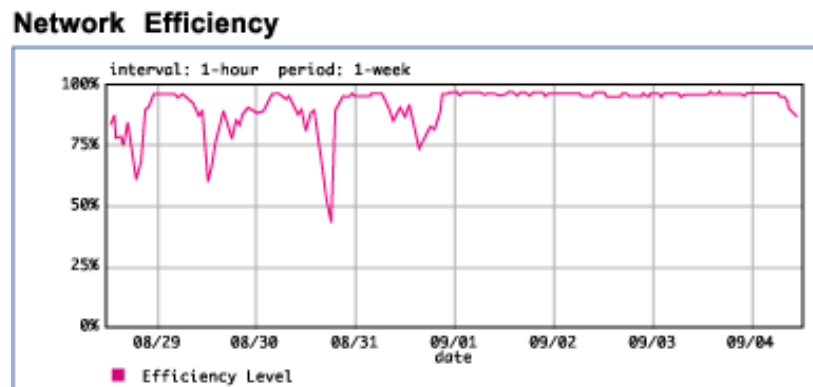


Figure 8: Network efficiency shows the percentage of wasted bandwidth.

## Real-World ROI

Packet-shaping technology can be a huge boon to companies that want the reliability and cost savings associated with transitioning voice and video from circuit-switched to IP packet-switched data networks. Of course, this transition can only proceed if the quality of service on the IP network is sufficient to enable high-quality IP videoconferencing experiences. Packet-shaping products assist in this transition.

One key to a successful transition is transmitting voice and video traffic over the existing corporate data network without performing a network upgrade. For example, suppose a company with an existing T1 WAN wanted to implement IP voice between six locations: Atlanta, Chicago, New York, Los Angeles, Denver, and Washington D.C. In many instances, the company would probably upgrade the network by adding a second T1 in each location to provide sufficient "uncontrolled" bandwidth for QoS. However, by using packet-shaping technology to provide application visibility and bandwidth control, the second T1 may be avoided. If the organization was able to avoid the network upgrade by investing in packet-shaping technology from a provider like Packeteer, the ROI would be calculated as follows:

<b>Cost Assumptions</b>	
Recurring Upgrade WAN T1 Costs (6 New T1s)	\$6,600 <sup>8</sup>
Cost for 6 PacketShaper 1200 Units (2 Mbps each)	\$17,700
ROI = \$17,700/\$6,600	
<b>ROI = 2.7 Month</b>	

*Table 4: ROI calculation for using Packeteer's PacketShaper technology to avoid upgrading six T1s when implementing IP voice and video.*

## Conclusion

Transitioning to voice and video over IP requires rock-solid quality of service over the wide-area network link. One of the major IP packet congestion points is at the LAN/WAN interface, where bandwidth availability can often decrease by two or more orders of magnitude. Provisioning additional WAN bandwidth may provide temporary relief, but it may be an expensive short-term solution because additional bandwidth is often consumed by more aggressive non-business applications like Web surfing, peer-to-peer file sharing, streaming of rich media files, and more.

Because of the unpredictability associated with today's multi-service business networks, it is extremely important to conduct an IP readiness assessment to ensure that investments in voice and video over IP pay off. This method consists of the following steps:

- Perform an assessment of the current network's traffic mix
- Determine bandwidth and performance characteristics of current and new applications
- Implement controls based on the organization's overall concept of network operations
- Measure and report

Visibility and control are essential in enabling network quality of service. Traffic classification, performance analysis, and a clear delineation of organizational data priorities allow network administrators to develop appropriate management policies which can then be implemented to gain control over network utilization and business application performance – specifically voice and video traffic. Devices like Packeteer's PacketShaper<sup>®</sup>, which automate most of the classification and analysis, and which offer significant control functionality, provide a clear path toward provisioning and managing IP voice and video networks with high-quality service-level compliance.

Exploiting today's packet-shaping technology may significantly decrease the cost of transitioning to voice and video over IP by eliminating the need to invest in costly bandwidth upgrades and extending existing resources in an intelligent, business-focused manner.

## About Packeteer

Packeteer, Inc., (NASDAQ: PKTR) is the global market leader in Application Traffic Management for wide area networks. Deployed at more than 7,000 companies in 50 countries, Packeteer solutions empower IT organizations with patented network visibility, control, and acceleration capabilities delivered through a family of intelligent, scalable appliances. For more information, contact Packeteer at +1 (408) 873-4400 or visit the company's Web site at [www.packeteer.com](http://www.packeteer.com).

<sup>8</sup> T1 cost is typical U.S. retail WAN pricing obtained from numerous industry sources. Actual pricing varies between carriers and is affected by carrier discounts, length of contract, and current market conditions.

## **About Wainhouse Research**

Wainhouse Research (<http://www.wainhouse.com>) is an independent market research firm that focuses on critical issues in rich media communications, videoconferencing, teleconferencing, and streaming media. The company conducts multi-client and custom research studies, consults with end users on key implementation issues, publishes white papers and market statistics, and delivers public and private seminars as well as speaker presentations at industry group meetings. Wainhouse Research publishes a number of reports detailing the current market trends and major vendor strategies in the multimedia networking infrastructure, endpoints, and services markets, as well as the segment reports *Comparing IP Video Network Service Providers Versus the Naked Internet* and *Surviving in the Conferencing Reseller Channel* and the free newsletter, *The Wainhouse Research Bulletin*.

## **About the Author**

**E. Brent Kelly** is a Senior Analyst and Partner at Wainhouse Research. Brent was formerly VP of marketing at Sorenson where he launched the company's live streaming and IP videoconferencing products. He has authored articles on IP conferencing and has developed seminars on implementing IP-based Rich Media Conferencing. As an executive in a manufacturing firm, he developed and implemented a marketing and channel strategy that helped land national accounts at major retailers. Brent has significant high tech product management and development experience, working on the team that built the devices that Intel uses to test their microprocessors. He has also led teams developing real-time data acquisition and control systems and adaptive intelligent design systems for Schlumberger. He has worked for several other multinational companies including Conoco, and Monsanto. Mr. Kelly has a Ph.D. in engineering from Texas A&M and a B.S. in engineering from Brigham Young University. He can be reached at [bkelly@wainhouse.com](mailto:bkelly@wainhouse.com).