Quality of Service In Internet Protocol (IP) Networks
Prepared for the International Communications Industries Association
To support Infocomm 2002

By E. Brent Kelly
Senior Analyst and Consultant, Wainhouse Research, Brookline, MA, USA

E. Brent Kelly has more than 15 years experience developing and marketing high technology products. Prior to consulting with Wainhouse Research, Brent held senior marketing, managerial and engineering positions with several large and small high technology companies. He has published many articles and research papers on IP-based multimedia communications, videoconferencing, and artificial intelligence. Brent has also created and teaches seminars on multimedia communications and videoconferencing. Mr. Kelly holds B.S. degree in engineering from Brigham Young University and a Ph.D. in engineering from Texas A&M University.

A revolution is occurring as organizations of all sizes begin to implement IP-based voice and video communications systems. Termed converged networking, these new IP technologies allow enterprises to convert from separate circuit-switched telephone and ISDN video networks to all IP networks where data, voice, and video all traverse the same network infrastructure.

Within a converged network, Quality of Service (QoS) is by far the most important implementation consideration. QoS is a networking term that specifies a guaranteed network data performance level. In practical terms, QoS is a mechanism to assure that audio and video data traverse the network with minimum delay. If network QoS is not in place, IP voice or videoconferencing calls will be unreliable, inconsistent, and often unsatisfactory.

QoS Elements: Bandwidth, End-to-End Delay, Jitter, and Packet Loss

Network quality of service is evaluated by measuring four key parameters: bandwidth, end-to-end delay, jitter, and packet loss. Bandwidth, typically specified in kilo or mega bits per second (kbps or Mbps), is measured as the average number of bits per second that can travel successfully through the network. End-to-end delay is the average time it takes for a network packet to traverse the network from one endpoint to the other. Jitter is the variation in the end-to-end delay of sequential packets. Packet loss is measured as the percent of transmitted packets that never reach the intended destination1.

For IP voice and video communications systems to work properly, the bandwidth should be as large as possible while the end-to-end delay, jitter, and packet loss are minimized. Lower end-to-end delay leads to a more satisfactory, natural-feeling conferencing experience, while large delay values lead to unnatural conversations with long pauses between phrases or sentences. Large jitter values may cause network data packets to arrive in the wrong sequence causing jerky video or stuttering, popping audio as will packet loss greater than 1%.

1 Transmitted packets may be lost for several reasons, but the primary cause is due to congestion in the network routers. When too many packets are simultaneously sent to a router, it will simply discard some packets, assuming that the application that sent the packet will retransmit it.
The ITU standard G.114 states that end-to-end delay should be no more than 150 milliseconds (ms). However, experience has shown that an end-to-end delay of 200 ms is still usually satisfactory for most users. Jitter should not be more than 20 to 50 ms.

Delay from the endpoints should not exceed one hundred milliseconds. Total latency, which includes end-to-end network delay and endpoint processing time, therefore should not exceed approximately three hundred milliseconds. Otherwise, users of the system will be able to detect the latency, particularly in the audio, and they will have an unpleasant experience.

**Three Fundamental QoS Enablers**

Most existing networks have been designed for data applications that do not require real-time performance. Because audio and video data must be received in real-time, QoS must be designed into the network. Three fundamental concepts affecting real-time data transmission must be considered while designing the IP network for audio and video data. These are network provisioning, queuing, and classifying.

**Network Provisioning**

The most common approach to quality of service today is to over-provision the network bandwidth. Over provisioning the network simply means installing more network bandwidth or capacity than is actually needed for all of the audio, video, and regular data applications that will run over the network.

Bandwidth refers to the “speed” or throughput of the network, typically specified in kbps or Mbps. Two common forms of Ethernet deployed inside the enterprise are 10 and 100 Mbps, while a T-1 connection capable of carrying 1.5 Mbps of data traffic is often used for enterprise wide area networks (WAN) or corporate connections to the Internet.

Rich media communications can consume significant bandwidth on the enterprise network and network provisioning is an important part of any rich media implementation plan. Usually the least cost solution, and the first step to be taken in any IP rich media communications environment, is to boost the network bandwidth by migrating to a 100 Mbps switched Ethernet architecture and by segmenting the LAN into multiple sub-nets so that the available bandwidth is shared by a smaller number of endpoints.

![Network Provisioning](image)

*Figure 1.* Many organizations try to achieve QoS by over provisioning the network bandwidth. Some very high quality IP audio and video calls may be configured to use 768 kbps or more network bandwidth. This number of kbps refers to the actual amount of data that will be transmitted by each endpoint. When designing the network for QoS, consideration must also be made for network overhead. A typical video call will use approximately 20% network overhead. Thus, a call made at 768 kbps actually may consume as much as 920 kbps on the network. At this bandwidth, only a single call would satisfactorily traverse a T-1 connection over the WAN.

As a general rule of thumb, the maximum bandwidth required for all applications added together, including voice and video, should not exceed 75% of the available network bandwidth. Consequently, over provisioning the network to some extent is necessary; however, by itself, over

---

2 768 kbps is used only for illustration purposes. Many video calls are made at 384 kbps and some as low as 128 kbps with excellent results.
provisioning is not sufficient to guarantee adequate quality of service.

Queuing

Network designers have come to understand that buffering, not bandwidth, is the key QoS issue. Transmission buffers in network switches and routers tend to fill rapidly in high-speed networks. This causes packet drops, which in turn causes voice or video clipping and skipping. As shown in the figure below, every point in the network where there is a router or a switch may be a source of transmission or buffering challenges, each potentially giving rise to poor quality of service.

Figure 2. Network designers have come to understand that buffering, not bandwidth, is the QoS issue. Transmit buffers tend to fill rapidly in high-speed networks causing audio and video packet loss or excessive delay.

Buffering issues may be overcome by enabling separate voice and video data queues in the network switches and routers. Separate queues allow time critical data such as audio and video to be transmitted in a priority fashion.

To enable queuing, application data or time sensitive voice and video data must be classified in some manner prior to entering the switch. Based upon each data packet’s classification, the packet is placed into an appropriate transmission queue; time critical voice and video data are classified such that they are placed into a delay and drop sensitive queue.

This may mean that any data arriving simultaneously with the audio and video may be lost. However, since application data typically is not real time, lost data will simply be retransmitted by the application initiating the data packet, and there will be no noticeable interruptions in the normal data flow on the network. Queuing gives the delay sensitive voice and video data a higher priority in the network switch or router insuring that the voice or video packet will be transmitted in a timely manner.

Figure 3. Establishing priority queues in network switches allows time sensitive voice and video data to be placed in transmission queues in a preferential fashion with respect to non-real-time data.

Network hubs do not support data queuing and may lead to increased packet collisions, thereby causing packet loss or delay; consequently, switches are preferred over hubs in a network designed to support QoS. Also, not all network switches support separate queues or classifying schemes. Those that do not will need to be upgraded when implementing QoS.

Classifying

Queuing is enabled by some type of packet classification or prioritization scheme. Several different schemes currently exist for providing priority to network packets. These include Resource Reservation Protocol (RSVP), IP precedence, differentiated services (DiffServ), and Multi-Protocol Label Switching (MPLS).

---

3 Application data is transmitted using TCP, which is a non-real time protocol that prescribes lost packet retransmission. Voice and video data use UDP transmission protocol, a real-time protocol that does not allow retransmission of lost packets.
RSVP

RSVP is a flow-based protocol that provides a guaranteed quality of service for each data flow. Each unidirectional data stream between two applications is considered a separate flow. In a typical videoconference, there would be four flows: audio transmit and receive and video transmit and receive.

In practice, RSVP is somewhat cumbersome to implement. The reason for this is that every device along the data flow path, which include servers, PC’s, routers switches, etc. must be able to signal the RSVP specified QoS requirements. Hence, RSVP is difficult to scale to very large implementations.

IP Precedence and DiffServ

IP precedence and DiffServ rely upon similar mechanisms for providing quality of service wherein certain bits in the data packet header are modified. Upon arrival at an IP precedence or DiffServ enabled router or switch, packets with the header bits set appropriately are given priority queuing and transmitting.

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. 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 $64^4$ 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 quality of service dependent upon the DiffServ classification.

Multi-Protocol Layer Switching (MPLS)

Conventional routers make packet-forwarding decisions by performing the complex task of looking up the routing information based upon the destination IP address in each packet. Each router along the routed path makes an independent forwarding decision by analyzing the packet header and forwarding the packet from one router to the next until the packet reaches its final destination. The choice of the next hop for a packet is based on the header analysis and the result of running a routing algorithm. This approach is sometimes insufficient to support today’s networking demands, because routers can become QoS bottlenecks, even when IP Precedence and DiffServ schemes are employed.

MPLS\(^5\) defines a different approach to improving and simplifying the packet forwarding function and to providing sufficient network guarantees to support quality of service. MPLS is designed to bring the speed of OSI layer 2, the link/switching layer, up to layer 3, the network protocol layer. Each packet is assigned a routing label based upon several factors including the packet priority and the ultimate packet destination. Label-based switching is faster because it allows routers to make forwarding decisions based upon

---

\(^4\) 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/.

\(^5\) For more information on MPLS see www.mplsrc.com and www.qosforum.com.
the contents of a simple label, rather than performing the complex task of routing lookup based upon the destination IP address.

MPLS brings a number of other benefits to IP-based networks including RSVP-like guaranteed QoS; nevertheless, few MPLS networks are actually functioning today because the specifications are still slightly in flux. MPLS-enabled networking equipment is now available, however, and some network service providers are implementing MPLS.

**Packet Shaping**

IP precedence and DiffServ require the packet header bits to be modified. Some, but not all, videoconferencing endpoints allow network administrators to set the IP precedence or DiffServ header bits. Should an organization have endpoints that do not allow the administrators to set these bits, a technology called packet shaping can be employed to set the bits as needed. Packet shaping devices are placed on the network prior to the router or switch and simply check the contents of each packet traversing the network. Packet shapers can be configured such that they recognize audio and video packets and set the IP precedence or DiffServ header bits as necessary to provide each packet the priority required to traverse the network routers and switches in a timely manner.

**Conclusion**

Network quality of service is a critical element of a successful converged networking design. Although over-provisioning the network bandwidth may provide adequate QoS temporarily, additional mechanisms including queuing and classification should be designed into the network infrastructure.