Your Short Cut to Knowledge

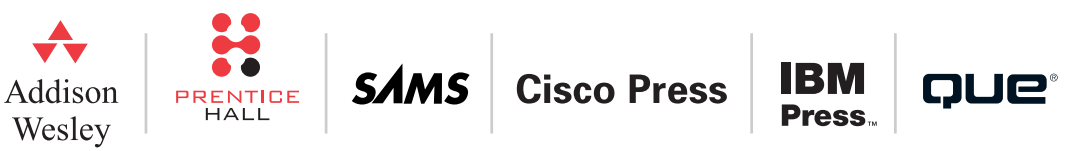
The following is an excerpt from a Short Cut published by one of the Pearson Education imprints.

Short Cuts are short, concise, PDF documents designed specifically for busy technical professionals like you.

We've provided this excerpt to help you review the product before you purchase. Please note, the hyperlinks contained within this excerpt have been deactivated.

Tap into learning—NOW!

Visit **www.informit.com/shortcuts** for a complete list of Short Cuts.



Quality of Service

Quality of service (QoS) is possibly the single most important feature to deploy to ensure a successful VoIP system. This section defines and describes why QoS is needed and explains how to configure and deploy a QoS solution using both the Modular QoS Command Line (MQC) and AutoQoS.

QoS Definition

QoS is defined as

> The ability of the network to provide better or "special" service to a set of users and applications at the expense of other users and applications.

Voice and video traffic is very sensitive to delayed packets, lost packets, and variable delay (jitter). The effects of these problems manifest as choppy audio, missing sounds, echo, or unacceptably long pauses in the conversation that cause overlap, or one talker interrupting the other. QoS configurations provide bandwidth guarantees while minimizing delay and jitter for priority traffic like VoIP. They do so not by creating additional bandwidth, but by controlling how the available bandwidth is used by the different applications and protocols on the network. In effect, this often means that data applications and protocols are restricted from accessing bandwidth when VoIP traffic needs it. This does not have much of an impact on the data traffic, however, because it is generally not as delay or drop-sensitive as VoIP traffic.

The areas that QoS can address to improve and guarantee voice quality are the following:

- Bandwidth
- Delay (including delay variation or jitter)
- Packet loss

 Please see page 147 for more details.

Bandwidth

A VoIP call follows a single path from end to end. That path may include a variety of LAN and WAN links. The slowest link represents the available bandwidth for the entire path—often referred to as a bottleneck because of the congestion the slow link can cause.

If congestion is occurring, there are several ways to fix the problem:

- **Increase the bandwidth:** If bandwidth is unlimited, congestion cannot occur. Realistically, however, increasing bandwidth is costly and is usually unnecessary if QoS is applied instead.
- **Queuing:** QoS employs advanced queuing strategies, which classify different traffic types and organize the classes into queues that are emptied in order of priority. The queuing strategies commonly used in Cisco Unified Communications include the following:
  - **Weighted fair queuing (WFQ):** WFQ dynamically assigns bandwidth to traffic flows as they arrive at the router interface. No configuration is necessary; the strategy is enabled by default on router links of T1 speed and below. This strategy is not appropriate for VoIP because it does not provide a bandwidth guarantee for the voice traffic, but instead allocates bandwidth fairly based on flow sizes (hence the name). VoIP needs a Priority queue (PQ) that guarantees it the bandwidth it needs, at the expense of all other traffic.
  - **Class-based weighted fair queuing (CBWFQ):** CBWFQ extends the WFQ algorithm to include user-defined classes for traffic. Instead of the router dynamically interpreting traffic flows and building queues for them, the admin classifies the traffic and assigns it to queues of configurable size and bandwidth allocation. There is still no priority queue, however, so CBWFQ is not appropriate for VoIP.
  - **Low-latency queuing (LLQ):** LLQ extends the CBWFQ system with the addition of a PQ. The PQ is typically reserved for voice traffic, and if any packets show up in the PQ, all packets in the queue are immediately sent while packets of other traffic types are held in their respective queues. This is the preferred queuing method for VoIP networks.

 Please see page 147 for more details.

- **Compression:** Several strategies are available to make the data that needs to be sent smaller so that it consumes less bandwidth:
  - **Payload compression:** By compacting the contents of a packet, the total size is somewhat reduced. This compression method does not affect the headers, which makes it appropriate for links that require the header to be readable to route the packet correctly (Frame Relay and ATM as examples).
  - **Link compression:** On point-to-point links where the header information is not needed to route the packet, link compression may be used.
  - **Header compression:** By specifying the use of compressed RTP (cRTP), the Layer 3 and 4 headers of a VoIP packet are dramatically reduced, from 40 to as little as 2 bytes. TCP header compression is also available for non-VoIP traffic using TCP transport.

Compression takes time and CPU resources, adding delay; this must be factored in to the decision of what strategies are appropriate for a given link.

Delay

Reducing end-to-end delay is a primary goal of QoS. Delay is calculated by adding the cumulative delay totals from source to destination and will be expressed as one-way or round-trip. Delay is classified in the following ways:

- **Fixed delay** is predictable and constant. Sources of fixed delay include the following:
  - **Propagation delay:** The amount of time it takes for the signal to transit the link. This is effectively the speed of light as it moves through copper or optical media. Light travels just less than a foot in one-billionth of a second, so long-distance links can impose significant delay that cannot be eliminated. L.A. to New York links routinely see 40 ms one-way propagation delay.

 Please see page 147 for more details.

  - **Serialization delay:** This is the time it takes to load bits onto the media; this relates directly to the speed of the link and cannot be altered unless that speed is changed.
- **Variable delay** includes processing and queuing delays; these will vary depending on the traffic load, the router performance, and many other factors that are not easily predictable or constant.

Minimizing delay employs the same strategies as improving bandwidth:

- Increase link speed.
- Use Priority queuing (such as LLQ) for delay-sensitive traffic.
- Employ appropriate compression techniques.

Packet Loss

Ideally, no packets of any type will be lost, but this is not realistic. We do need to minimize packet loss for VoIP traffic because it has no mechanism to retransmit lost packets (unlike TCP, for example). Packets are lost for a variety of reasons:

- **Tail drop:** When an output queue is full, no more arriving packets can be placed in the queue. Any packets that arrive when the queue is full are dropped from the last position (tail) of the queue and cannot be recovered. This is the most common source of packet loss.
- **Input drop:** If the input queue fills up, arriving packets are dropped and lost. This is rare, and it usually indicates an overloaded router CPU.
- **Overrun:** Also the result of CPU congestion, overruns happen when the router cannot assign the packet to a free buffer space.

 Please see page 147 for more details.

- **Ignore:** There is no buffer space available.
- **Frame errors:** Problems in transmission created CRC errors, giant or runt frames. This is usually related to EMI or failing interface hardware.

Minimizing loss can be achieved with QoS mechanisms like LLQ and compression or by increasing link speed. Some additional and complementary strategies known as Link Efficiency mechanisms will help to prevent congestion:

- **Traffic shaping:** Delays packets and sends them at a configured maximum rate. For example, if an FTP server is generating a 512 kbps stream, shaping could limit the output to 256 kbps, delaying the transmission of the excess traffic. This will add significant delay and might cause packets to be dropped, so it is not desirable to shape VoIP traffic, but shaping data traffic is an effective tool to complement voice QoS settings.
- **Traffic policing:** Drops packets in excess of a configured threshold. These packets may be retransmitted if the traffic is using TCP, but because VoIP does not, policing should not be applied to VoIP traffic. Again, policing is an effective complement to QoS configurations.

QoS Requirements for VoIP

There are some accepted targets for delay, loss, and jitter for VoIP traffic. These are the targets that QoS and Link Efficiency mechanisms help us reach:

- Delay should be less than 150 ms one way.
- Jitter (the variation in the delay between packets) should be less than 30 ms one way.
- Packet loss should be less than 1 percent.
- Each VoIP call requires between 17 kbps and 106 kbps of priority bandwidth, depending on the codec, compression, and Layer 2 in use; it also requires another 150 bps per call for signaling traffic.

 Please see page 147 for more details.