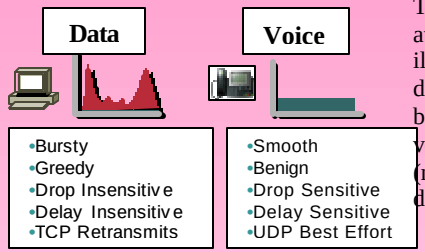
QoS: Quality of Service

Courtesy of Cisco Enterprise Marketing

Why should I care about quality of service?

Quality of service (QoS) refers to the perceived and measured performance of a network, typically thought of in terms of the sound quality of a voice call. Without implementing QoS strategy, applications such as IP telephony, videoconferencing, and mission-critical data are subject to "best effort (non-guaranteed) transmission. This can result in choppy voice or video during times of network congestion.



The diagram at left illustrates the differences between voice and (non-critical) data.

What are the Problems to be solved?

The Three Evils

The three parameters that define QoS are Loss, Delay, and Delay Variance (Jitter). Control of these three factors allows you to control the QoS:

LOSS refers to the percentage of packets dropped. In a highly available network this should be less than 1%. Voice network should approach 0% loss.

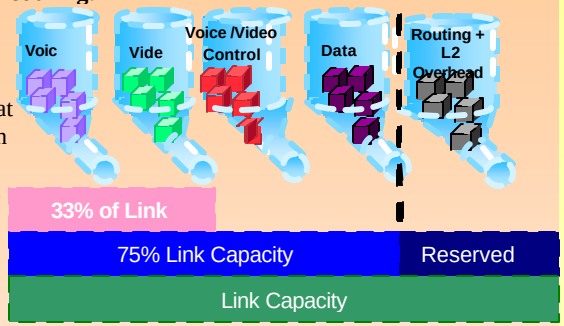
DELAY: Delay refers to the time it takes for a packet to reach the target destination. Delay is comprised of fixed delay (serialization, quantization, etc.) and variable delay (network congestion). The total time it takes for a voice packet to cross the network should be less than 150 ms.

DELAY VARIATION (or JITTER) is the difference in the delay times of consecutive packets. A jitter buffer is used to smooth out arrival times, but there are instantaneous and total limits on the buffers ability to smooth out arrival times. Voice networks cannot more than 30 ms of jitter.

Three Steps to Quality

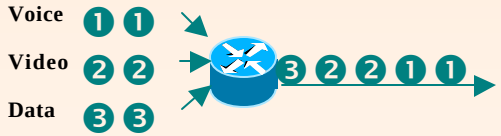
To mitigate the effects of the three evils, one must ensure that the network can properly handle time and drop sensitive packets. To achieve QoS, you must first leave room (bandwidth) for certain packets, you must identify which packets require special treatment, and you must maintain rules for how these packets should be treated. These three steps are also referred to as **provisioning**, **classification**, and **scheduling**.

Provisioning is the process of ensuring that the required bandwidth is available for all applications as well as for overhead traffic.



Classification refers to marking the packet with a specific priority denoting a requirement for special service from the network. This can be completed at Layer 2 or Layer 3. Typical classification schemes identify Critical (voice and mission-critical data), High (video), Normal (e-mail, Internet access), and Low (fax, ftp) priorities.

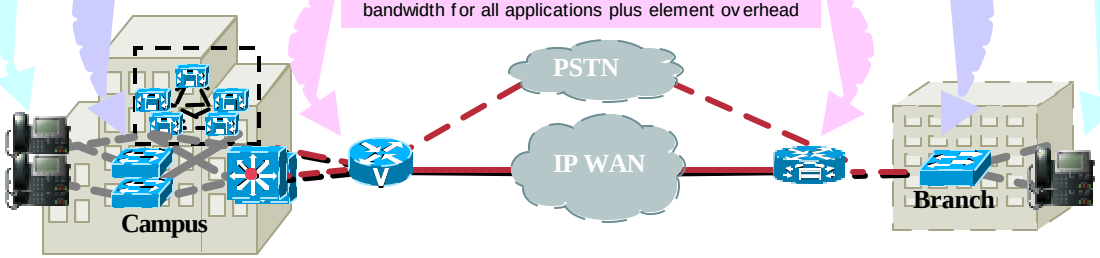
Scheduling is the process of assigning packets to one of multiple queues (based on classification) for priority treatment through the network. A good example of this is commercial airline boarding schemes. "Now boarding rows 40-50; first-class passengers and VIP members may board at any time."



Classification: Mark the packets with a specific priority denoting a requirement for class of service from the network
Trust Boundary: Define and enforce a trust boundary at the network edge

Scheduling: Assign packets to one of multiple queues (based on classification) for expedited treatment throughout the network; use congestion avoidance for data

Provisioning: Accurately calculate the required bandwidth for all applications plus element overhead

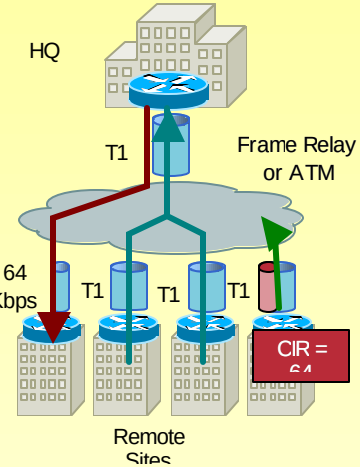


Low Bandwidth Tools

In addition to the three main steps to ensure QoS, some link-specific tools are also needed, such as Traffic Shaping and Link Fragmenting and Interleaving (LFI), especially when routing traffic over low-bandwidth (768kbs or slower) links.

Traffic Shaping is a method of throttling back packet transmission rates. If there are line speed mismatches between remote offices, the service provider connecting the offices may be forced to drop arbitrary packets traveling to the slower link. To avoid high-priority or drop-sensitive packets from being dropped, an enterprise can engineer its traffic to avoid over provisioning its traffic to the slower link. Traffic engineering also allows the enterprise to decide which packets can be should be dropped (low priority packets) when instantaneous congestion occurs.

The three most common cases for traffic engineering occur when there are:
Line speed mismatches
Remote to central site over-subscription
Traffic bursts above Committed Rate (CIR)



Link Fragmentation and Interleaving (LFI) . In addition to network congestion, one of the primary contributors to both delay and jitter is serialization delay. This is often caused by a time-sensitive packet getting "stuck in traffic" behind a large data packet (such as FTP). Link fragmentation is the process of breaking up large packets to allow smaller, more time sensitive, packets to proceed through the network in a timely manner. Interleaving is the processes of "weaving" in the time sensitive packets into the train of fragmented data packets.

