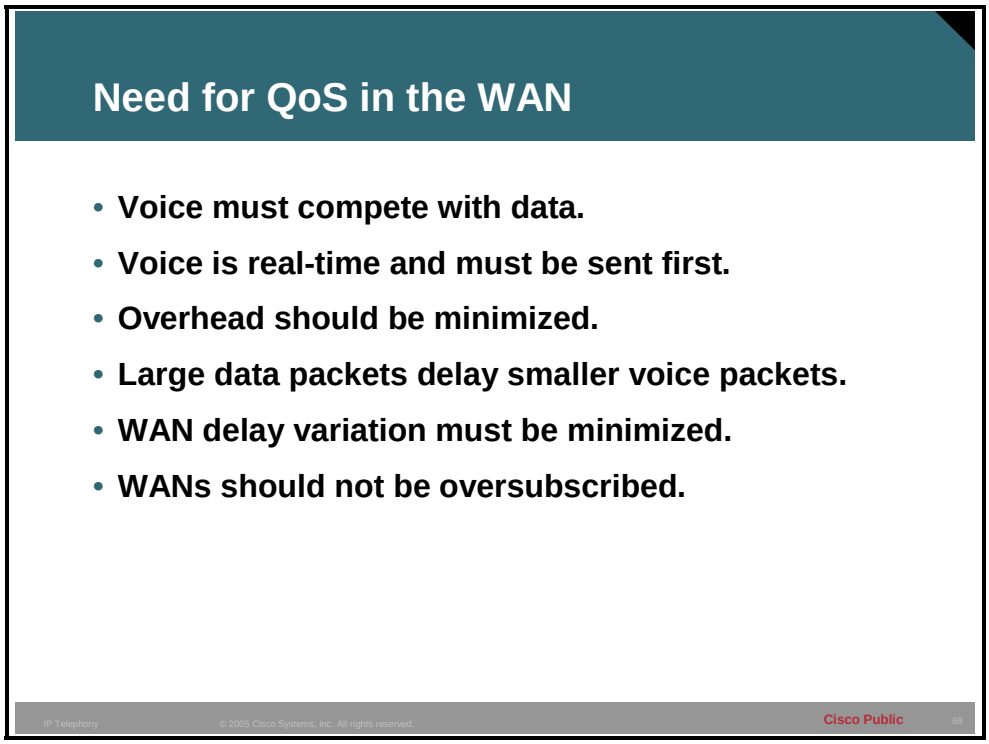
QoS Tools in the WAN

Need for QoS on WAN Links

This topic defines the need for QoS in a WAN.



You must consider and implement certain QoS measures before voice can be placed on a WAN to ensure the proper performance of voice and data applications. The following factors explain the need for QoS:

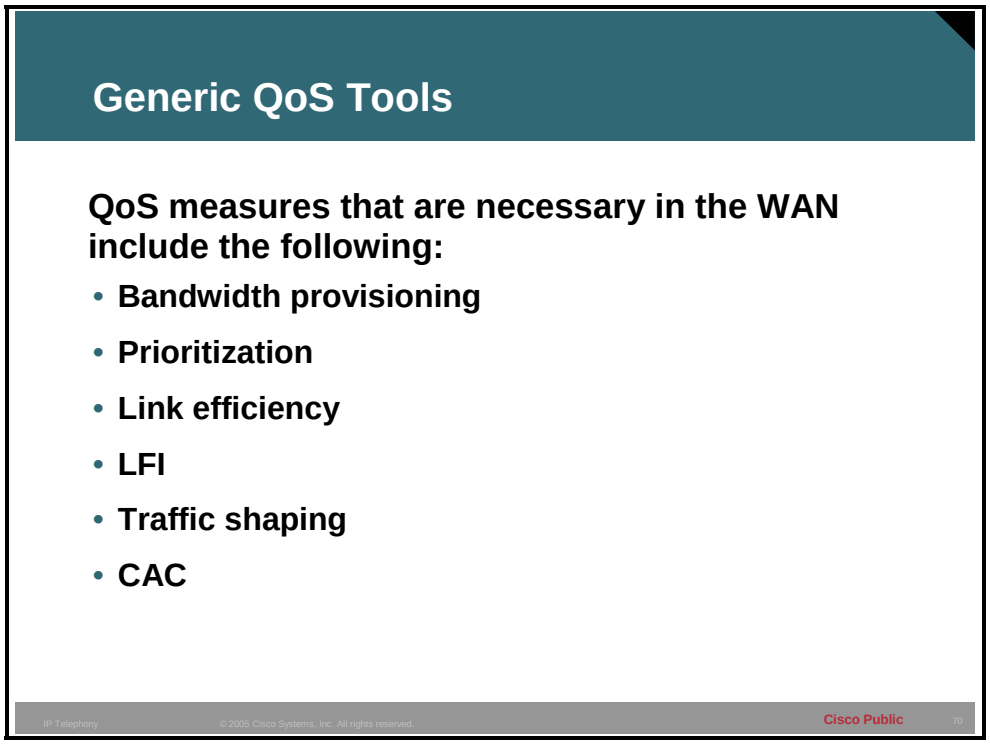
- **Voice must compete with data:** Voice traffic demand on the WAN is typically smooth; each voice call consumes a relatively fixed amount of bandwidth for the duration of the call. On the other hand, data traffic is bursty, peaking and leveling off based on user access and application type. The amount of total bandwidth that you provision must have the capacity to carry all of the expected traffic.
- **Voice is real-time and must be sent first:** Because voice is a real-time application and delayed packets have a severe impact on performance, you must prioritize voice ahead of data traffic that is not real-time traffic.
- **Overhead should be minimized:** Voice is sent using UDP. Unfortunately, this requirement adds significant overhead to the overall packet size. You must identify a means for numbering the packets that are needed, because minimizing the overhead with compression allows for smaller, more efficient packets.
- **Large data packets delay smaller voice packets:** With certain applications such as file transfers and web browsing, data packets that compete with voice can be relatively large. Voice packets are forced to wait until the larger data packets are sent. By fragmenting the

data packets into smaller, more manageable sizes, you can reduce the delay that is experienced by voice packets.

- **WAN delay variation must be minimized:** The natural behavior of certain WAN technologies, such as Frame Relay, ATM, and PPP can exacerbate delay variation and result in poor voice performance. By using tools to reduce the delay variation on WAN technologies, you can add significant performance to voice applications.
- **WANs should not be oversubscribed:** If you place too many voice calls on a WAN, the necessary bandwidth that is required by data applications will be choked. You must take care to ensure that data still gets its deserved bandwidth. Excess voice calls can severely impact all calls on the network.

Recommendations for Generic QoS in the WAN

This topic lists generic QoS tools.

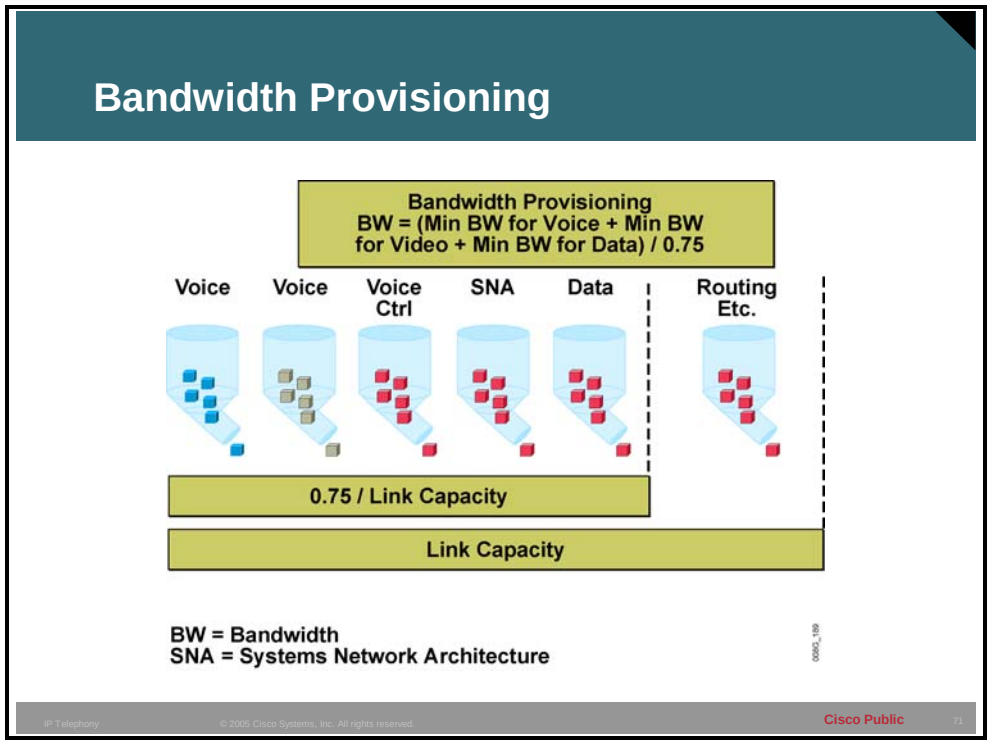


Depending on the Layer 2 technology that you use, the following QoS measures may be necessary in the WAN. Each of these QoS tools will be covered in this lesson:

- Bandwidth provisioning
- Prioritization
- Link efficiency
- LFI
- Traffic shaping
- CAC

Bandwidth Provisioning

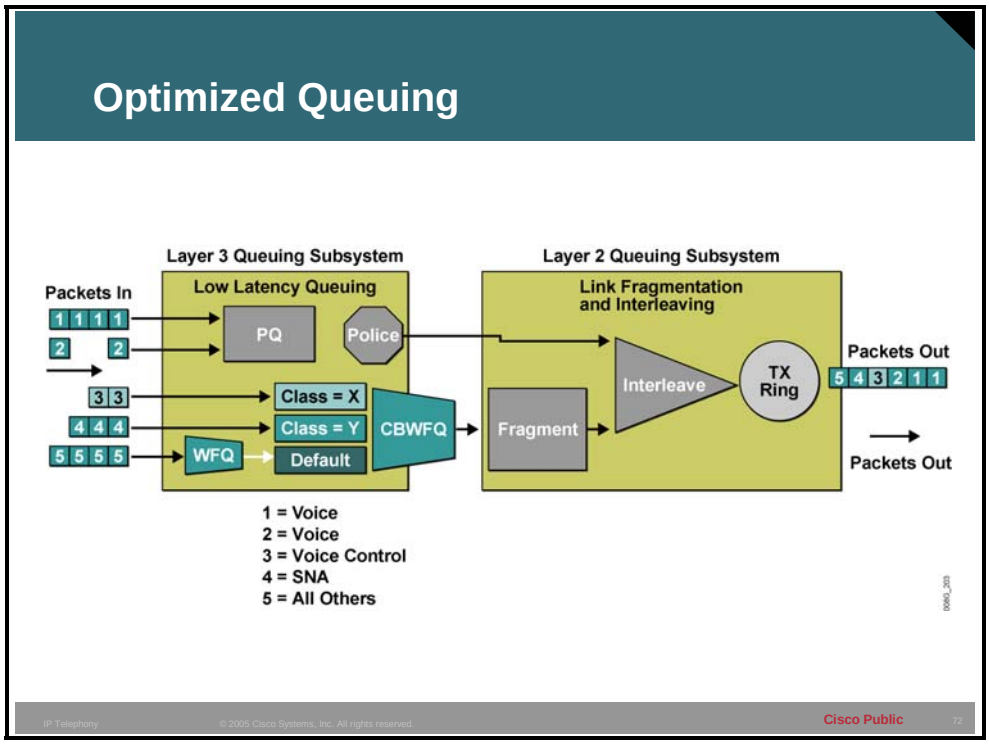
This topic describes bandwidth provisioning as a tool for implementing QoS in the WAN.



The figure represents a typical network. This network has a number of delay-sensitive traffic types: voice, video, and Systems Network Architecture (SNA). Before you place voice and video on a network, you must ensure that adequate bandwidth exists for all required applications; for example, SNA is also delay-sensitive but should not be queued *ahead* of voice and video. To begin, the minimum bandwidth requirements for each major application (for example, voice media streams, video streams, voice control protocols, and all data traffic) should be summed. This sum represents the minimum bandwidth requirement for any given link, and it should consume no more than 75 percent of the total bandwidth that is available on that link. This 75 percent rule assumes that some bandwidth is required for overhead traffic, such as routing and Layer 2 keepalives, as well as for additional applications such as e-mail, HTTP traffic, and other data traffic that is not easily measured.

Optimized Queuing

This topic describes optimized queuing as a tool for implementing QoS in the WAN.



When you are choosing from among the many available prioritization schemes, the major factors that you must consider include the type of traffic on the network and the wide-area media that is being traversed.

For multiservice traffic over an IP WAN, Cisco recommends LLQ for low-speed links. This approach allows up to 64 traffic classes with the ability to specify, for example, PQ behavior for voice and interactive video, a minimum bandwidth for SNA data and market data feeds, and, as illustrated in the figure, WFQ to other traffic types.

Example: Optimized Queuing

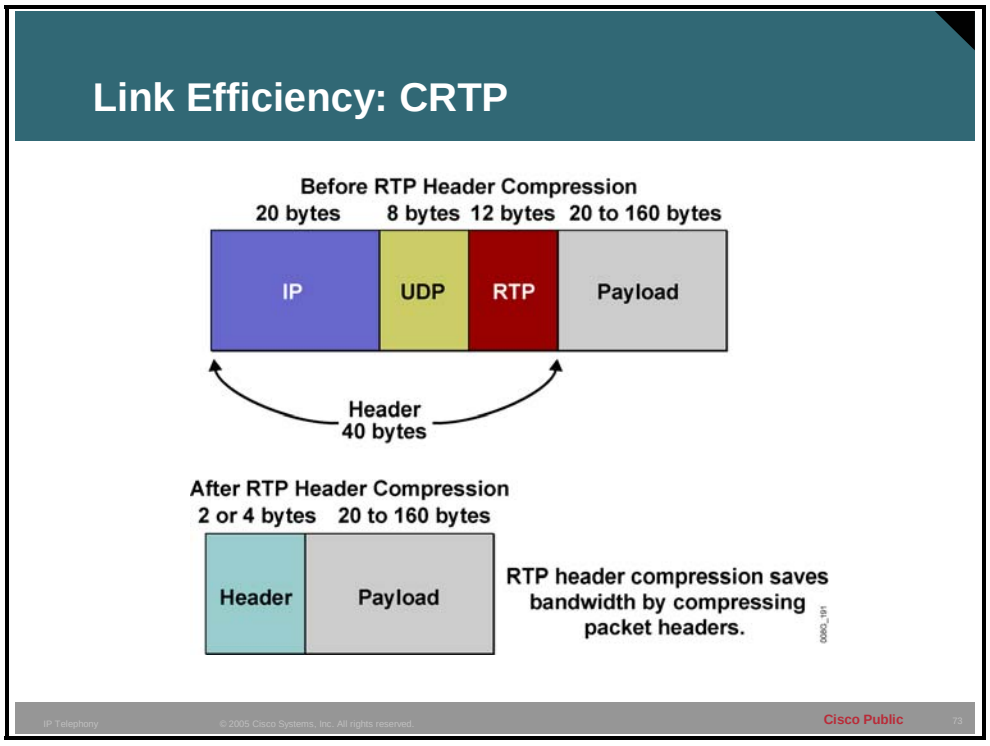
The figure illustrates LLQ, which works as follows:

- Voice is placed into a queue with PQ capabilities and allocated a bandwidth of 48 kbps, for example. This bandwidth is based on the total number of expected calls. The entrance criterion to this queue should be the DSCP value of Express Forwarding (EF), or IP precedence value of 5. Traffic in excess of 48 kbps is dropped if the interface becomes congested. Therefore, you must use an admission control mechanism to ensure that you do not exceed this value.
- As the WAN links become congested, it is possible to completely starve the voice control signaling protocols and therefore eliminate the capacity of the IP Phones to complete calls across the IP WAN. Voice control protocol traffic, such as H.323 and SCCP, requires its own CBWFQ with a minimum configurable bandwidth equal to a DSCP value of AF31, which correlates to an IP precedence value of 3.

- SNA traffic is placed into a queue that has a specified bandwidth of 56 kbps, for example. This bandwidth represents the expected SNA demand on the network. Queuing operation within this class is FIFO with a minimum allocated bandwidth of 56 kbps. Traffic in this class that exceeds 56 kbps is placed in the default queue. The entrance criterion to this queue could be TCP port numbers, a Layer 3 address, IP precedence, or a DSCP.
- You can place all remaining traffic in a default queue. If a bandwidth is specified, the queuing operation is FIFO. Alternatively, specifying the keyword **fair** assigns WFQ to the operation.

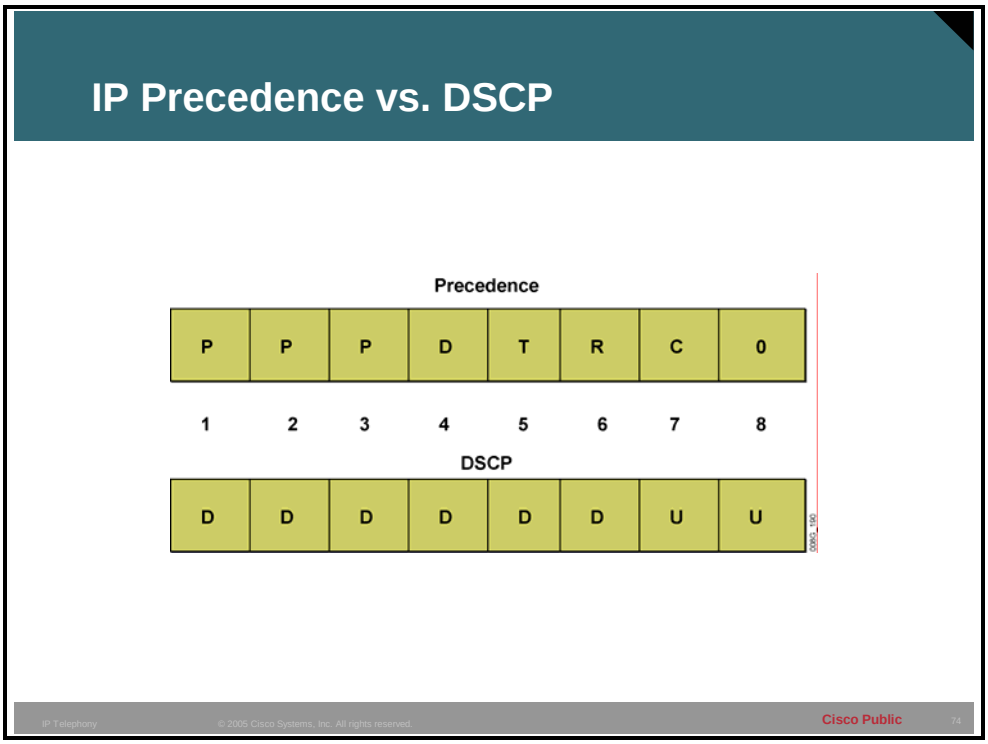
Link Efficiency

This topic describes link efficiency as a tool for implementing QoS in the WAN.



Because wide-area bandwidth is often prohibitively expensive, only low-speed circuits may be available or cost-effective when you are interconnecting remote sites. In this scenario, it is important to achieve the maximum savings by transmitting as many voice calls as possible over the low-speed link. Many compression schemes, such as G.729, can squeeze a 64-kbps call down to an 8-kbps payload. Cisco gateways and IP Phones support a range of codecs that enhance efficiency on these low-speed links.

The link efficiency is further increased by using CRTP, which compresses a 40-byte IP + UDP + RTP header to approximately 2 to 4 bytes. In addition, VAD takes advantage of the fact that in most conversations, only one party is talking at a time. VAD recovers this empty time and allows data to use the bandwidth.



IP Precedence

The 8-bit CoS field in an IP packet header was originally defined in RFC 791 (superseded by RFC 1122). It defined the most significant bits as shown in the table.

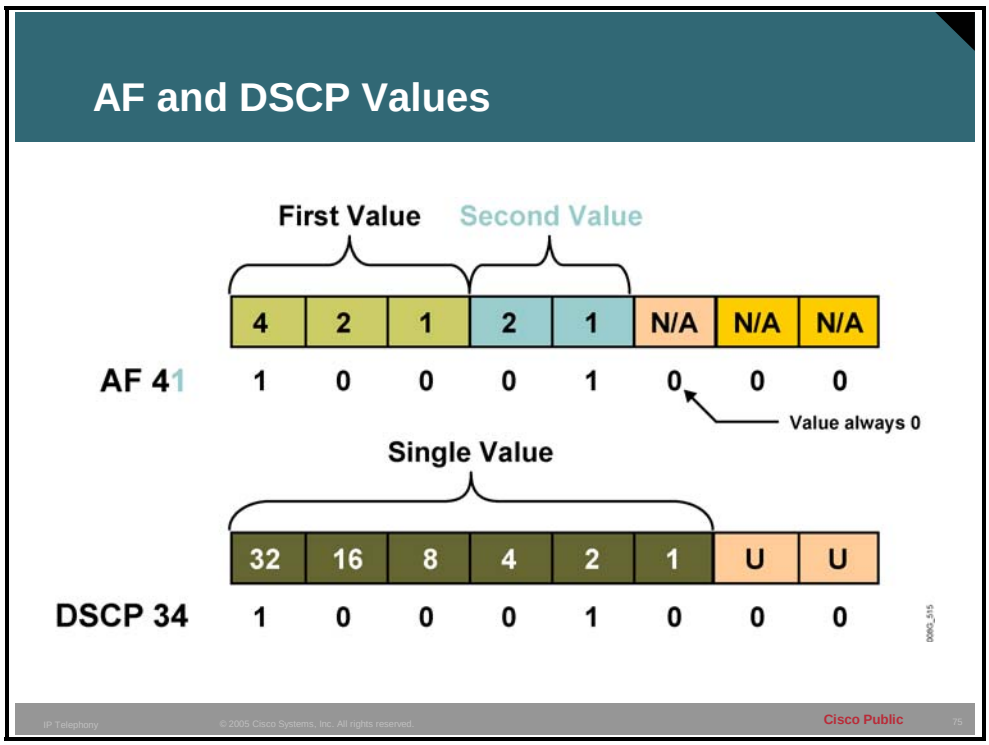
IP Precedence

Binary Value	Precedence Value	Description
111	7	Network Control
110	6	Internetwork Control
101	5	Critical
100	4	Flash Override
011	3	Flash
010	2	Immediate
001	1	Priority
000	0	Routine

In addition, the next 4 bits, when turned on, refer to the following:

- **Bit 4, D:** Instructs the network to minimize delay
- **Bit 5, T:** Instructs the network to maximize throughput
- **Bit 6, R:** Instructs the network to maximize reliability
- **Bit 7, C:** Instructs the network to minimize costs
- **Bit 8:** Reserved for future use

DSCP



The newest use of the eight CoS bits is commonly called the DiffServ standard. It uses the same precedence bits (the most significant bits: 1, 2, and 3) for priority setting, but further clarifies their functions and definitions and offers finer priority granularity through use of the next three bits in the CoS field. DiffServ reorganizes and renames the precedence levels (still defined by the three most significant bits of the CoS field) into the categories shown in the table.

DiffServ Standard

Precedence Bits	Description
Precedence 7	Stays the same (link layer and routing protocol keepalive)
Precedence 6	Stays the same (used for IP routing protocols)
Precedence 5	EF
Precedence 4	Class 4
Precedence 3	Class 3
Precedence 2	Class 2
Precedence 1	Class 1
Precedence 0	Best effort

Bits 3 and 4 of the CoS field (now called the DSCP in the DiffServ standard) allow further priority granularity through the specification of packet-drop probability for any of the defined classes. Collectively, classes 1 through 4 are referred to as Assured Forwarding (AF). The table illustrates the DSCP coding for specifying the priority level (class) plus the drop percentage.

(Bits 1, 2, and 3 define the class; bits 4 and 5 specify the drop percentage; bit 6 is always 0.) As an example, AF41 is expressed as 100010, where the first three bits represent class 4, the next two bits specify a low drop percentage, and the last bit is always 0. AF41 has a higher priority than any class 3, 2, or 1, and enjoys the lowest drop percentage. AF41 is also referred to as DSCP 34, as shown in the figure.

DSCP Coding

Drop Percentage	DSCP Coding			
	Class 1	Class 2	Class 3	Class 4
Low drop percentage	001010	010010	011010	100010
Medium drop percentage	001100	010100	011100	100100
High drop percentage	001110	010110	011110	100110

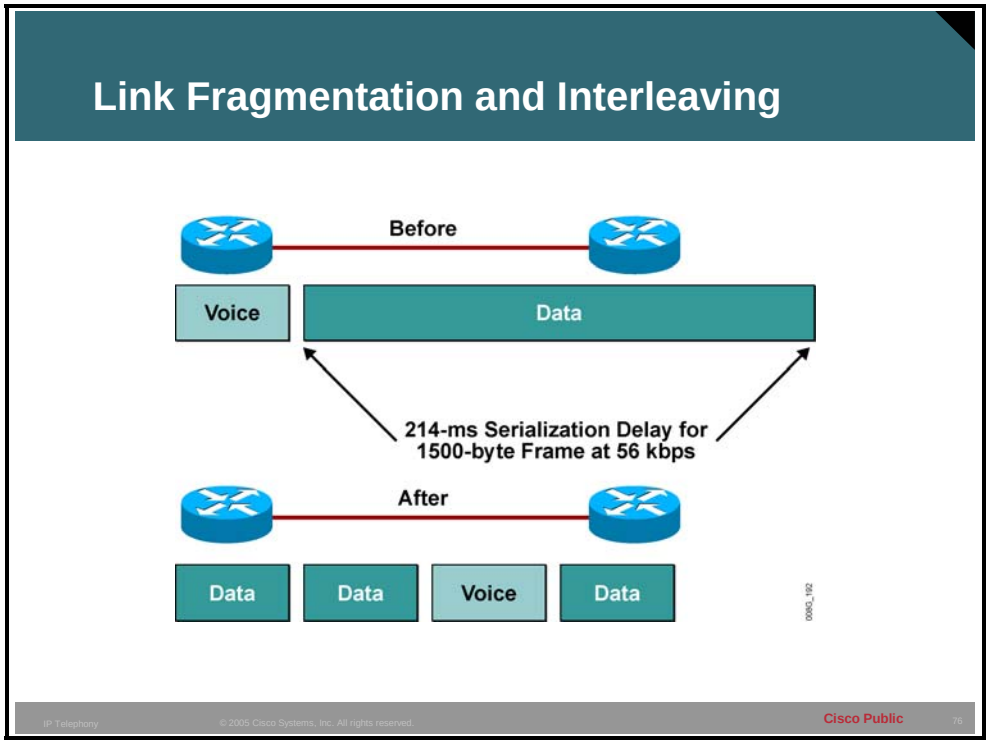
Using this system, a device first prioritizes traffic by class, then differentiates and prioritizes traffic that is in the same class by considering the drop percentage. It is important to note that this standard does not offer a precise definition of low, medium, and high drop percentages. Additionally, not all devices recognize the DiffServ bit 4 and 5 settings. In fact, even when the settings are recognized, they do not necessarily trigger the same forwarding action from each device on the network.

Each device implements its own response in relation to the packet priorities that it detects. The DiffServ proposal is meant to allow a finer granularity of priority setting for the applications and devices that can use it, but it does not specify interpretation (that is, action to be taken). In this application, bits 7 and 8 are unused.

The figure shows the relationship between AF values and DSCP values. AF41 can also be referred to as DSCP 34.

Link Fragmentation and Interleaving

This topic describes link fragmentation and interleaving (LFI) as a tool for implementing QoS in the WAN.



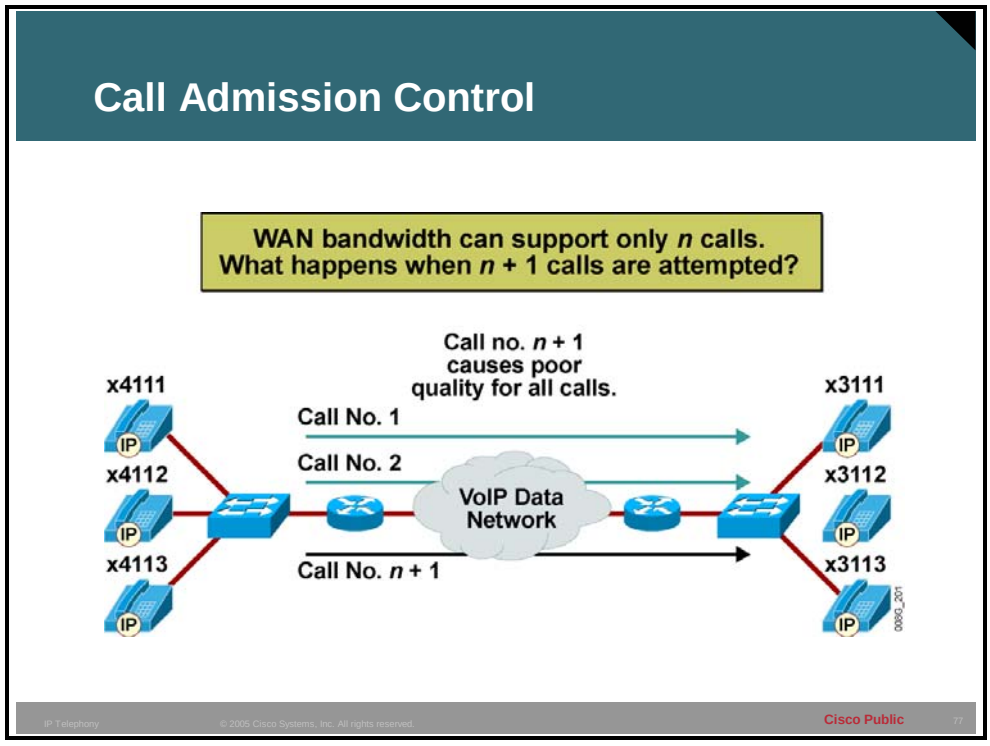
For low-speed links (less than 768 kbps), it is necessary to use techniques that provide LFI. This approach places bounds on jitter by preventing the delay of voice traffic behind large data frames. The following three techniques exist for this purpose:

- MLP for point-to-point serial links
- FRF.12 for Frame Relay
- MLP over ATM for ATM connections

The figure shows the necessity of fragmenting the large data packet into smaller segments. Data packets should be fragmented to the same size as voice packets. If the data packets are fragmented to a size smaller than voice packets, voice packets will also be fragmented. Determining the size of the voice packet will depend on the codec type selected. Also, remember that using CRTP can reduce the size of a voice packet significantly.

CAC

This topic describes CAC as a tool for implementing QoS in the WAN.



CAC is required to ensure that network resources are not oversubscribed. CAC could be described as a way to protect voice from voice. Calls that exceed the specified bandwidth are either rerouted using an alternative route such as the PSTN, or a fast busy tone is returned to the calling party. This way the next voice call does not degrade the quality of all the calls on the link. You can implement CAC on a gatekeeper or by using Cisco CallManager.

Reference For more information on VoIP CAC, see the following topic on the Cisco.com website: http://www.cisco.com/univercd/cc/td/doc/cisintwk/intsolns/voipsol/cac.htm

Example: CAC

The figure illustrates the need for CAC. If the network segment was designed to carry two VoIP calls, then when the third call arrives at either the gatekeeper or the CallManager, the call is either rerouted to the PSTN or the caller is sent a fast busy tone. This way, the third call does not impact the quality of the two existing calls.