Chapter 6: Quality of Service

Connecting Networks

Cisco | Networking Academy®
Mind Wide Open™

Chapter 6 - Sections & Objectives

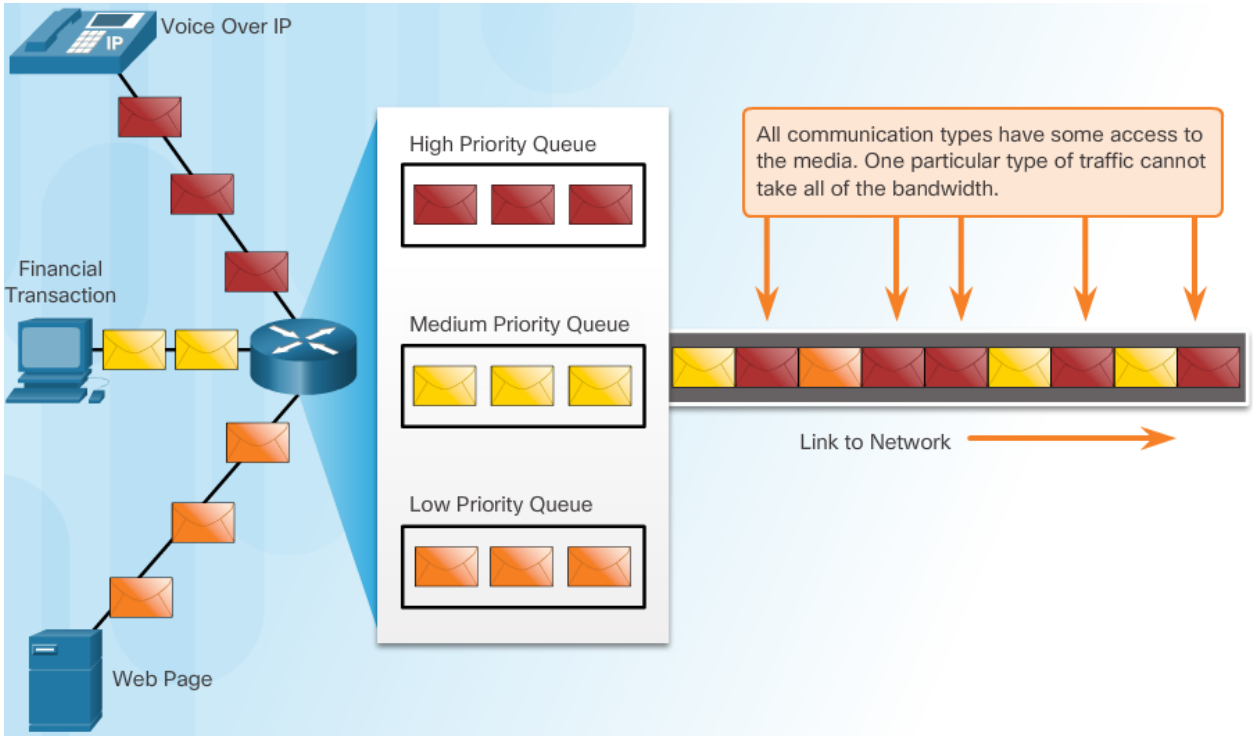
- 6.1 QoS Overview
  - Explain the purpose and characteristics of QoS.
- 6.2 QoS Mechanisms
  - Explain how networking devices implement QoS.

6.1 QoS Overview

Cisco | Networking Academy®
Mind Wide Open™

Network Transmission Quality

- Prioritizing Traffic
  - Queuing packets causes delay because new packets cannot be transmitted until previous packets have been processed.
  - If the number of packets to be queued continues to increase, the memory within the device fills up and packets are dropped.

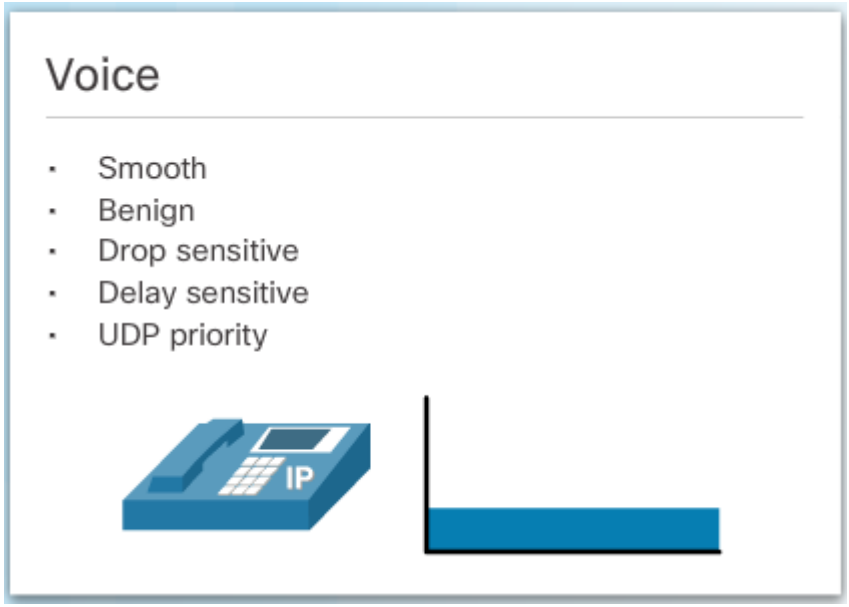


Network Transmission Quality

- Bandwidth, Congestion, Delay, and Jitter
  - Network congestion causes delay.
  - Delay is the time it takes for a packet to travel from the source to the destination.
  - Jitter is the variation in the delay of received packets.
- Packet Loss
  - When congestion occurs, network devices such as routers and switches can drop packets.
  - Packet loss is a very common cause of voice quality problems on an IP network.
  - In a properly designed network, packet loss should be near zero.
  - Network engineers use QoS mechanisms to classify voice packets for zero packet loss.

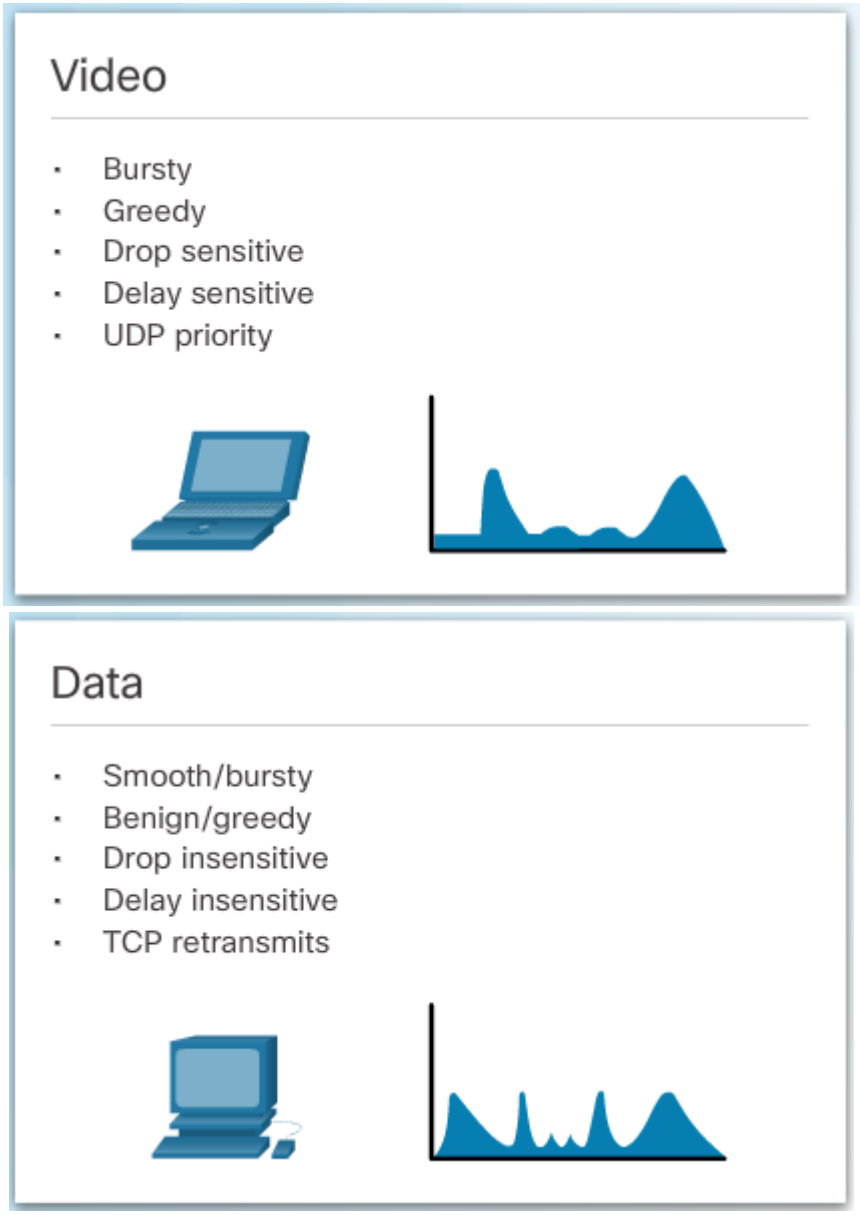
Traffic Characteristics

- Network Traffic Trends
  - The type of demands voice, video, and data traffic place on the network are very different.
- Voice
  - Voice is very sensitive to delays and dropped packets; there is no reason to re-transmit voice if packets are lost.
  - Voice packets must receive a higher priority than other types of traffic.
  - Voice can tolerate a certain amount of latency, jitter, and loss without any noticeable effects.



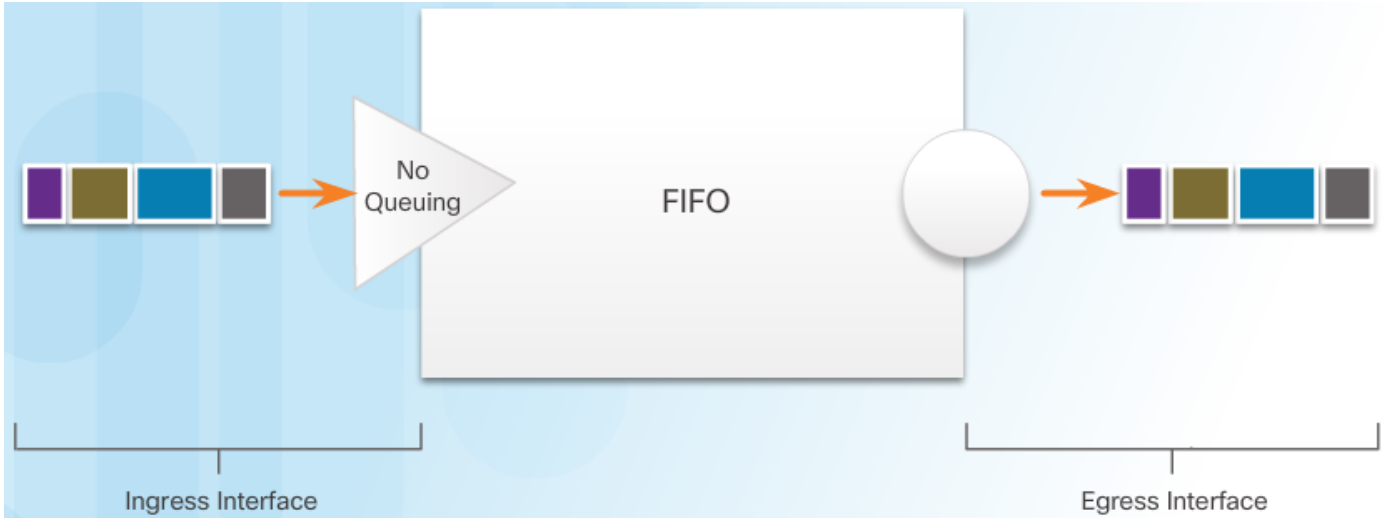
Traffic Characteristics

- Video
  - Compared to voice, video is less resilient to loss and has a higher volume of data per packet.
  - video can tolerate a certain amount of latency, jitter, and loss without any noticeable affects.
- Data
  - Data applications that have no tolerance for data loss, such as email and web pages, use TCP to ensure that, if packets are lost in transit, they will be resent.
  - Data traffic is relatively insensitive to drops and delays compared to voice and video.



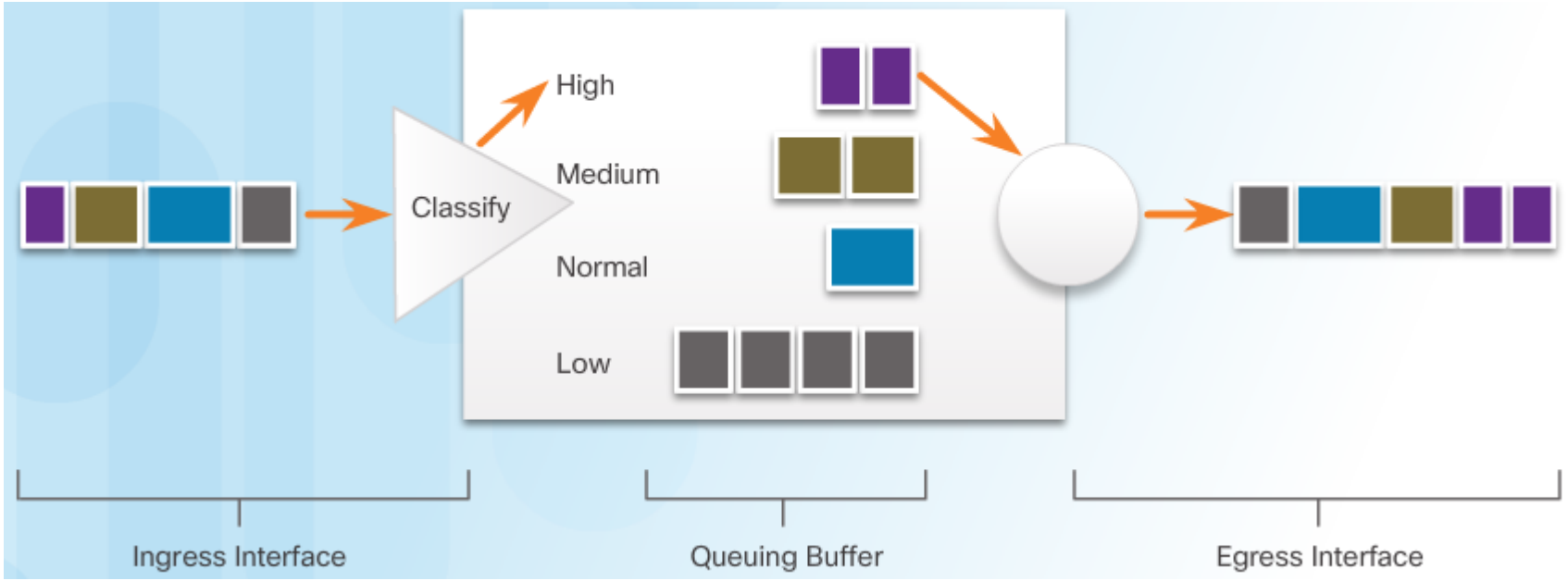
Queueing Algorithms

- First In First Out (FIFO)
  - FIFO has no concept of priority or classes of traffic and consequently, makes no decision about packet priority.
  - FIFO, which is the fastest method of queuing, is effective for large links that have little delay and minimal congestion.



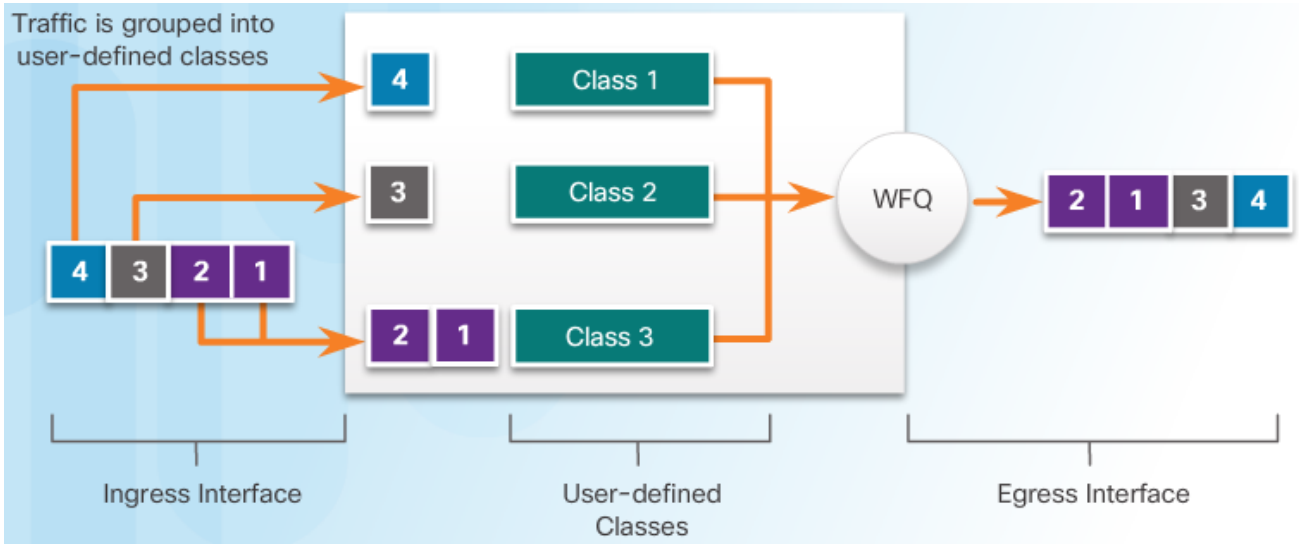
Queueing Algorithms

- Weighted Fair Queuing (WFQ)
  - An automated scheduling method that provides fair bandwidth allocation to all network traffic.
  - Applies priority, or weights, to identified traffic and classifies it into conversations or flows.
  - WFQ is not supported with tunneling and encryption because these features modify the packet content information required



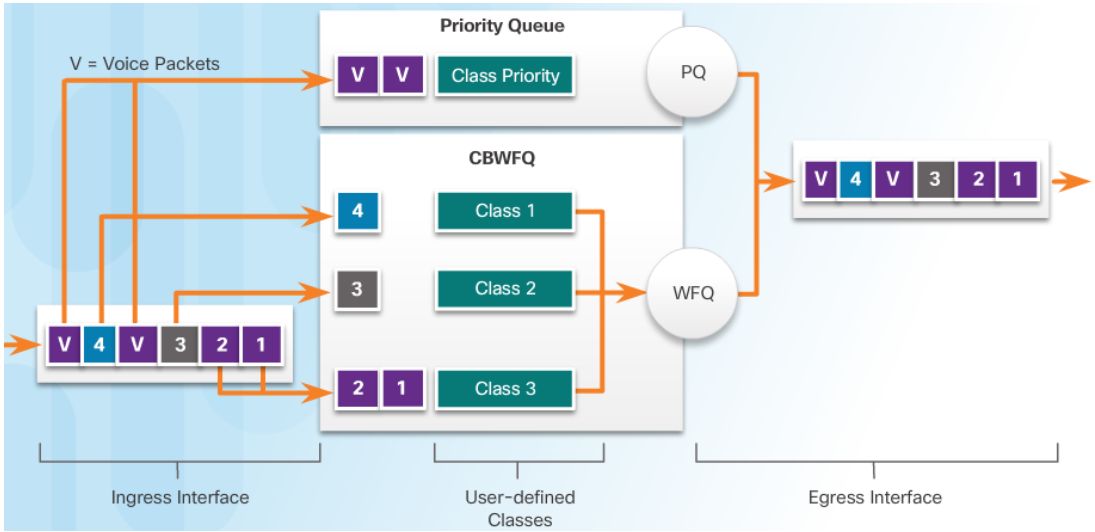
Queueing Algorithms

- Class-Based Weighted Fair Queuing (CBWFQ)
  - Extends the standard WFQ functionality to provide support for user-defined traffic classes.
  - To characterize a class, you assign it bandwidth, weight, and maximum packet limit.
  - You also specify the queue limit for that class, which is the maximum number of packets allowed to accumulate in the queue for the class.
  - Packets belonging to a class are subject to the bandwidth and queue limits that characterize the class.



Queueing Algorithms

- Low Latency Queuing (LLQ)
  - LLQ provides strict priority queuing for CBWFQ, reducing jitter in voice conversations.
  - The bandwidth assigned to the packets of a class determines the order in which packets are sent.
  - Without LLQ, all packets are serviced fairly based on weight; no class of packets may be granted strict priority.
  - LLQ allows delay-sensitive data such as voice to be sent first.



6.2 QoS Mechanisms

Cisco | Networking Academy®
Mind Wide Open™

QoS Mechanisms
QoS Models

- Selecting an Appropriate QoS Policy Model

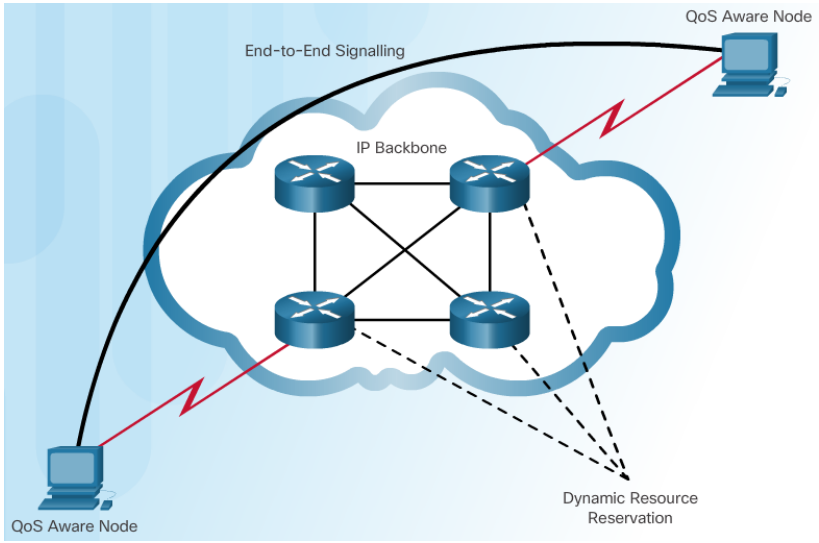
Model	Description
Best-effort model	• Not really an implementation as QoS is not explicitly configured. • Use when QoS is not required.
Integrated services (IntServ)	• Provides very high QoS to IP packets with guaranteed delivery. • It defines a signaling process for applications to signal to the network that they require special QoS for a period and that bandwidth should be reserved. • However, IntServ can severely limit the scalability of a network.
Differentiated services (DiffServ)	• Provides high scalability and flexibility in implementing QoS. • Network devices recognize traffic classes and provide different levels of QoS to different traffic classes.

- Best-Effort

Benefits	Drawbacks
The model is the most scalable.	There are no guarantees of delivery.
Scalability is only limited by bandwidth limits, in which case all traffic is equally affected.	Packets will arrive whenever they can and in any order possible, if they arrive at all.
No special QoS mechanisms are required.	No packets have preferential treatment.
It is the easiest and quickest model to deploy.	Critical data is treated the same as casual email is treated.

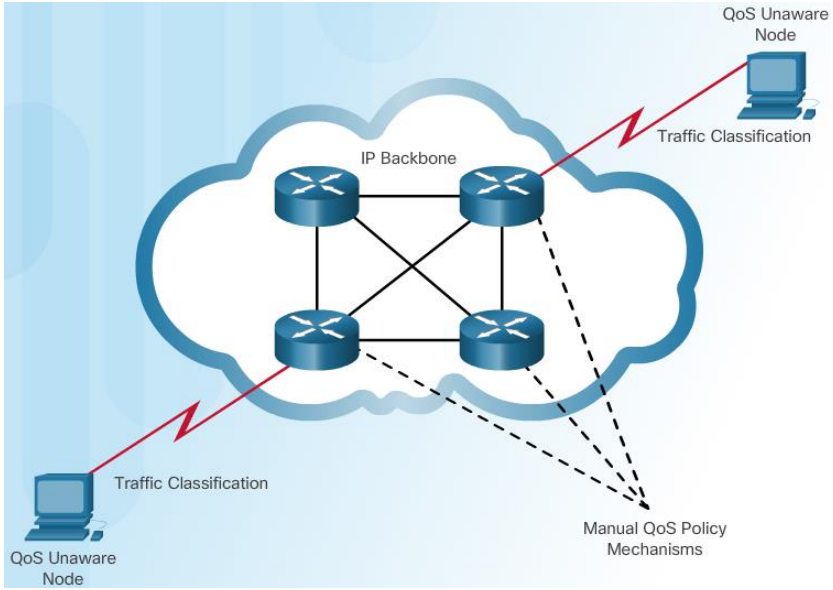
QoS Models

- Integrated Services (IntServ) (“hard QoS”)
  - Uses resource reservation and admission-control mechanisms as building blocks to establish and maintain QoS. The edge router performs admission control to ensure that available resources are sufficient in the network.
  - The IntServ standard assumes that routers along a path set and maintain the state for each individual communication.
  - If network devices along the path can reserve the necessary bandwidth, the originating application can begin transmitting.
  - If the requested reservation fails along the path, the originating application does not send any data.



QoS Models

- Differentiated Services (DiffServ) (“soft QoS”)
  - Specifies a simple and scalable mechanism for classifying and managing network traffic and providing QoS guarantees on modern IP networks.
  - DiffServ can provide an “almost guaranteed” QoS while still being cost-effective and scalable.
  - DiffServ uses a “soft QoS” approach. It works on the provisioned-QoS model, where network elements are set up to service multiple classes of traffic each with varying QoS requirements.
  - DiffServ divides network traffic into classes based on business requirements.
  - Each of the classes can then be assigned a different level of service.



QoS Implementation Techniques

- Avoiding Packet Loss
  - Dropped TCP segments cause TCP sessions to reduce their window sizes.
  - Some applications do not use TCP and cannot handle drops. (Voice)
- QoS Tools

QoS Tools	Description
Classification and marking tools	• Sessions, or flows, are analyzed to determine what traffic class they belong to. • Once determined, the packets are marked.
Congestion avoidance tools	• Traffic classes are allotted portions of network resources as defined by the QoS policy. • The QoS policy also identifies how some traffic may be selectively dropped, delayed, or re-marked to avoid congestion. • The primary congestion avoidance tool is WRED and is used to regulate TCP data traffic in a bandwidth-efficient manner before tail drops caused by queue overflows occur.
Congestion management tools	• When traffic exceeds available network resources, traffic is queued to await availability of resources. • Common Cisco IOS-based congestion management tools include CBWFQ and LLQ algorithms.

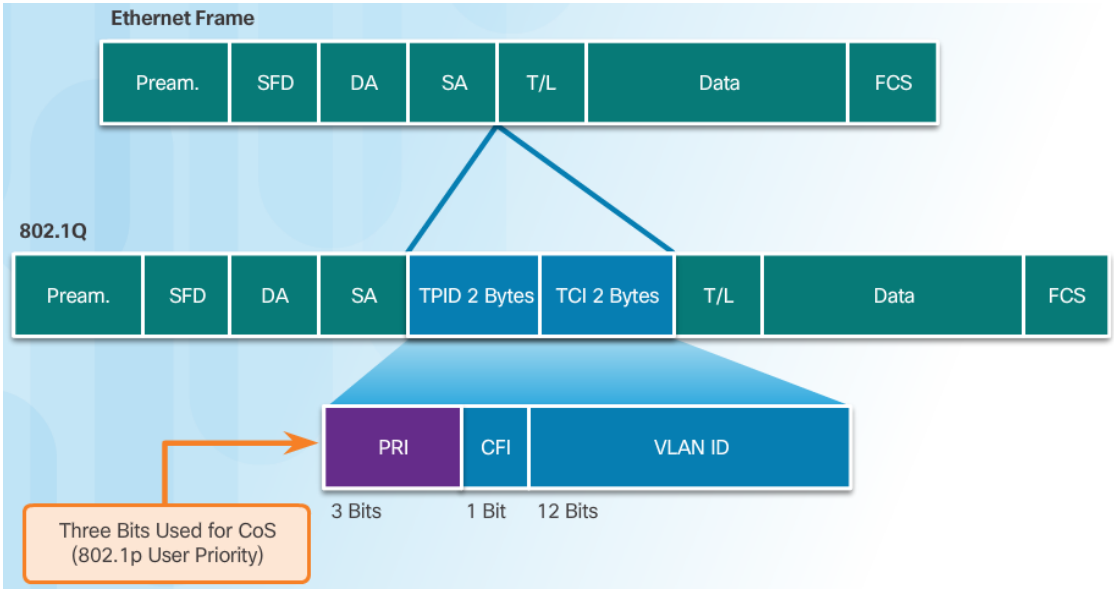
QoS Implementation Techniques

- Classification and Marking
  - Before a packet can have a QoS policy applied to it, the packet has to be classified.
  - Methods of classifying traffic flows at Layer 2 and 3 include using interfaces, ACLs, and class maps.

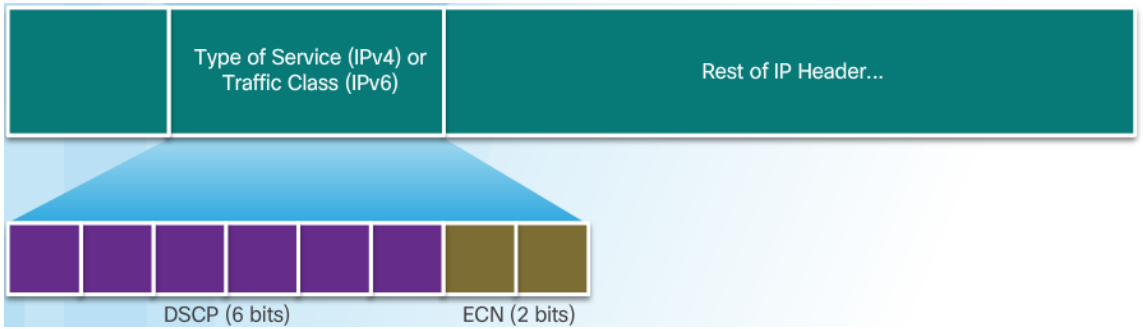
QoS Tools	Layer	Marking Field	Width in Bits
Ethernet (802.1Q, 802.1p)	2	Class of Service (CoS)	3
802.11 (Wi-Fi)	2	Wi-Fi Traffic Identifier (TID)	3
MPLS	2	Experimental (EXP)	3
IPv4 and IPv6	3	IP Precedence (IPP)	3
IPv4 and IPv6	3	Differentiated Services Code Point (DSCP)	6

QoS Implementation Techniques

- Marking at Layer 2
  (3 bits – 8 values)



- Marking at Layer 3
  (6 bits – 64 values)



QoS Implementation Techniques

- Trust Boundaries
  - Traffic should be classified and marked as close to its source as technically and administratively feasible.
- Congestion Avoidance
  - When the queue is below the minimum threshold, there are no drops.
  - As the queue fills up to the maximum threshold, a small percentage of packets are dropped.
  - When the maximum threshold is passed, all packets are dropped.
- Shaping and Policing
  - Traffic shaping retains excess packets in a queue and then schedules the excess for later transmission over increments of time.
  - Policing is applied to inbound traffic on an interface.
  - When the traffic rate reaches the configured maximum rate, excess traffic is dropped (or remarked).

6.3 Chapter Summary

Cisco | Networking Academy®
Mind Wide Open™

Summary

- The quality of network transmission is impacted by the bandwidth of the links between the source and destination, the sources of delay as packets are routed to the destination, and jitter or the variation in delay of the received packets. Without QoS mechanisms in place, packets are processed in the order in which they are received. When congestion occurs, time-sensitive packets will be dropped with the same frequency as packets that are not time-sensitive.
- Voice packets require latency of no more than 150 milliseconds (ms). Jitter should be no more than 30 ms, and voice packet loss should be no more than 1%. Voice traffic requires at least 30 Kb/s of bandwidth.
- Video packets require latency no more than 400 milliseconds (ms). Jitter should be no more than 50 ms, and video packet loss should be no more than 1%. Video traffic requires at least 384 Kb/s of bandwidth.
- For data packets, two factors impact the Quality of Experience (QoE) for end users:
  - Does the data come from an interactive application?
  - Is the data mission critical?

Summary Continued

- The four queuing algorithms discussed in this chapter are as follows:
  - **First in First Out (FIFO)** - Packets are forwarded in the order in which they are received.
  - **Weighted Fair Queuing (WFQ)** - Packets are classified into different flows based on header information including the ToS (Type of Service) value
  - **Class-Based Weighted Fair Queuing (CBWFQ)** - Packets are assigned to user-defined classes based on matches to criteria such as protocols, ACLs, and input interfaces. The network administrator can assign bandwidth, weight, and maximum packet limit to each class.
  - **Low Latency Queuing (LLQ)** - Delay-sensitive data such as voice is added to a priority queue so that it can be sent first (before packets in other queues).
- The three queuing models discussed in the chapter are as follows:
  - **Best-Effort** - This is the default queuing model for interfaces. All packets are treated in the same way. There is no QoS.
  - **Integrated Services (IntServ)** - IntServ provides a way to deliver the end-to-end QoS that real-time applications require by explicitly managing network resources to provide QoS to specific user packet streams, sometimes called microflows.
  - **Differentiated Services (DiffServ)** - DiffServ uses a soft QoS approach that depends on network devices that are set up to service multiple classes of traffic each with varying QoS requirements. Although there is no QoS guarantee, the DiffServ model is more cost-effective and scalable than IntServ.

Summary Continued

- QoS tools include the following:
  - **Classification and Marking** - Classification determines the class of traffic to which packets or frames belong. Marking means that we are adding a value to the packet header. Devices receiving the packet look at this field to see if it matches a defined policy.
  - **Congestion Avoidance** - Congestion avoidance tools monitor network traffic loads in an effort to anticipate and avoid congestion. As queues fill up to the maximum threshold, a small percentage of packets are dropped. Once the maximum threshold is passed, all packets are dropped.
  - **Shaping and Policing** - Shaping retains excess packets in a queue and then schedules the excess for later transmission over increments of time. Shaping is used on outbound traffic. Policing either drops or remarks excess traffic. Policing is often applied to inbound traffic.