

Quality of Service

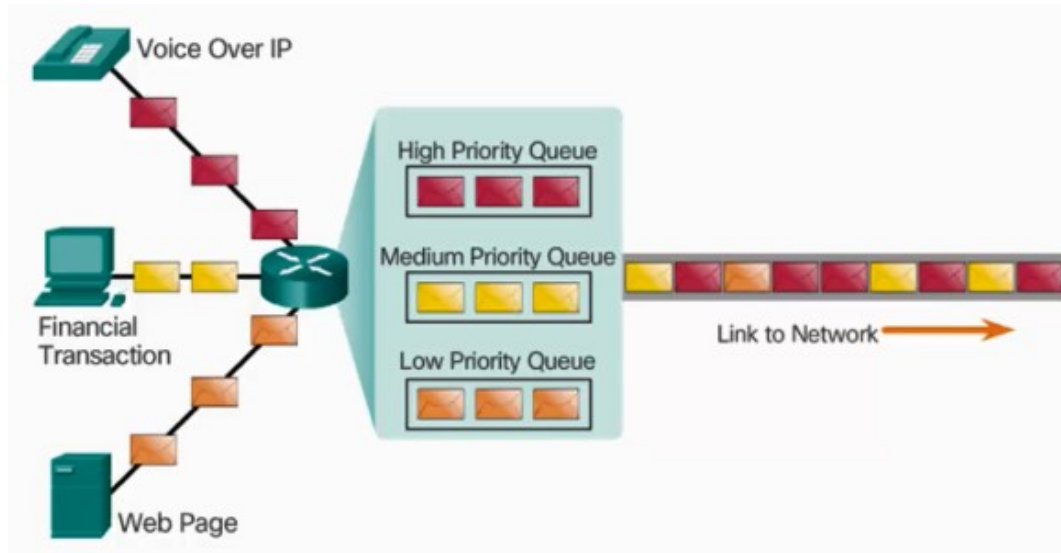
CCNA Routing and Switching

Connecting Networks v6.0



Network Transmission Quality

The Purpose of QoS

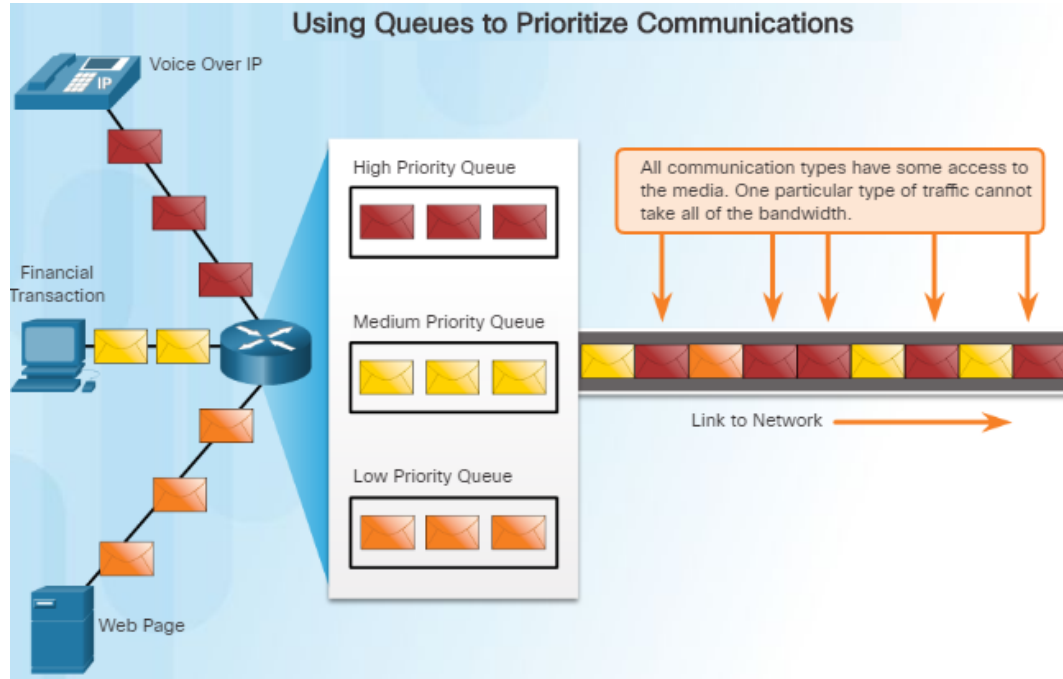


- When the volume of traffic is greater than what can be transported across the network, devices queue, or hold, the packets in memory until resources become available to transmit them.

- Quality of Service (QoS) allows network administrators to prioritize certain types of traffic over others.
- Video traffic and voice traffic require greater resources from the network than other types of traffic.
- Financial transactions are time sensitive and have greater needs than web traffic (HTTP).
- Congestion occurs when multiple communication lines aggregate onto a single device, such as a router, and then much of that data is placed on fewer outbound interfaces or onto a slower interface.

Network Transmission Quality

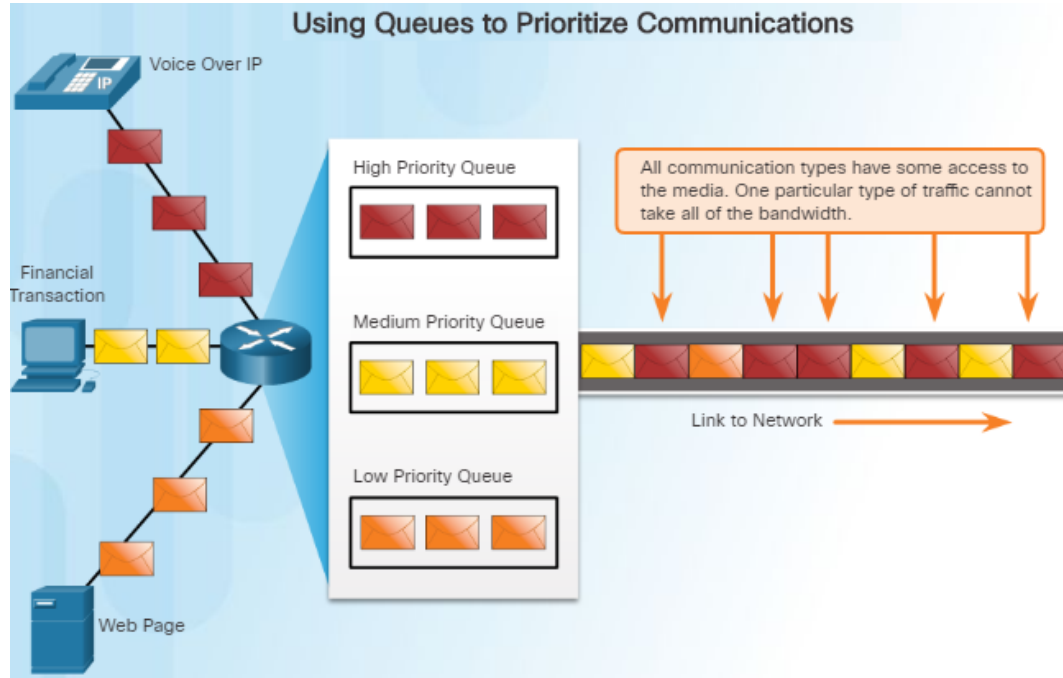
Prioritizing Traffic



- QoS is an ever increasing requirement of networks today thanks to new applications available to users such as voice and live video transmissions which create higher expectations for quality delivery.
- Congestion occurs when multiple communication lines aggregate onto a single device, such as a router, and then much of that data is placed on fewer outbound interfaces or onto a slower interface.
- When the volume of traffic is greater than what can be transported across the network, devices queue, or hold, the packets in memory until resources become available to transmit them.

Network Transmission Quality

Prioritizing Traffic

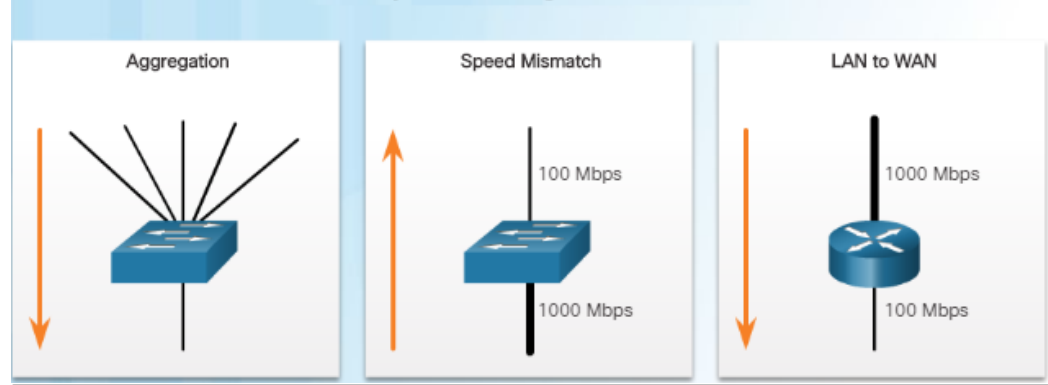


- Queuing packets causes delay because new packets cannot be transmitted until previous packets have been processed.
- Packets will be dropped when memory fills up.
- One QoS technique that can help with this problem is to classify data into multiple queues as shown in the figure to the left.
 - High Priority Queue
 - Medium Priority Queue
 - Low Priority Queue
- It is important to note that a device should implement QoS only when it is experiencing congestion.

Network Transmission Quality

Bandwidth, Congestion, Delay, and Jitter

Examples of Congestion Points

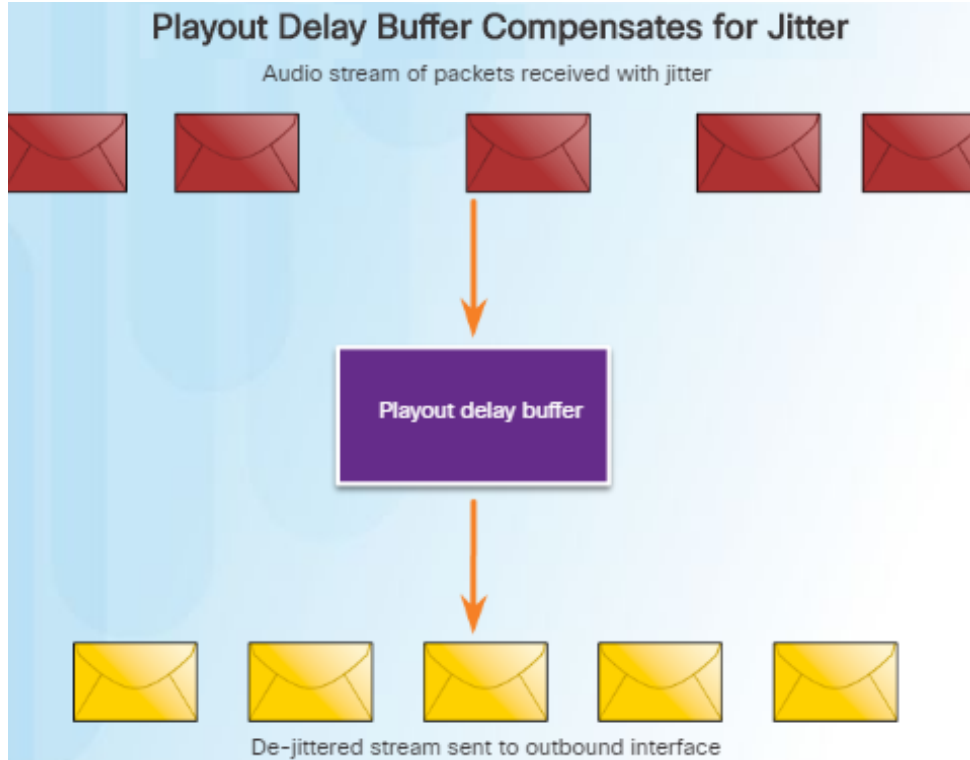


Delay	Description
Code delay	The fixed amount of time it takes to compress data at the source before transmitting to the first internetworking device, usually a switch.
Packetization delay	The fixed time it takes to encapsulate a packet with all the necessary header information.
Queuing delay	The variable amount of time a frame or packet waits to be transmitted on the link.
Serialization delay	The fixed amount of time it takes to transmit a frame onto the wire.
Propagation delay	The variable amount of time it takes for the frame to travel between the source and destination.
De-jitter delay	The fixed amount of time it takes to buffer a flow of packets and then send them out in evenly spaced intervals.

- Network bandwidth is measured in the number of bits that can be transmitted in one second (bps).
- Network congestion causes delay. An interface experiences congestion when it is presented with more traffic than it can handle.
- Delay or latency refers to the time it takes for a packet to travel from the source to the destination.
 - Fixed delay
 - Variable delay
- Jitter is the variation in delay of received packets.

Network Transmission Quality

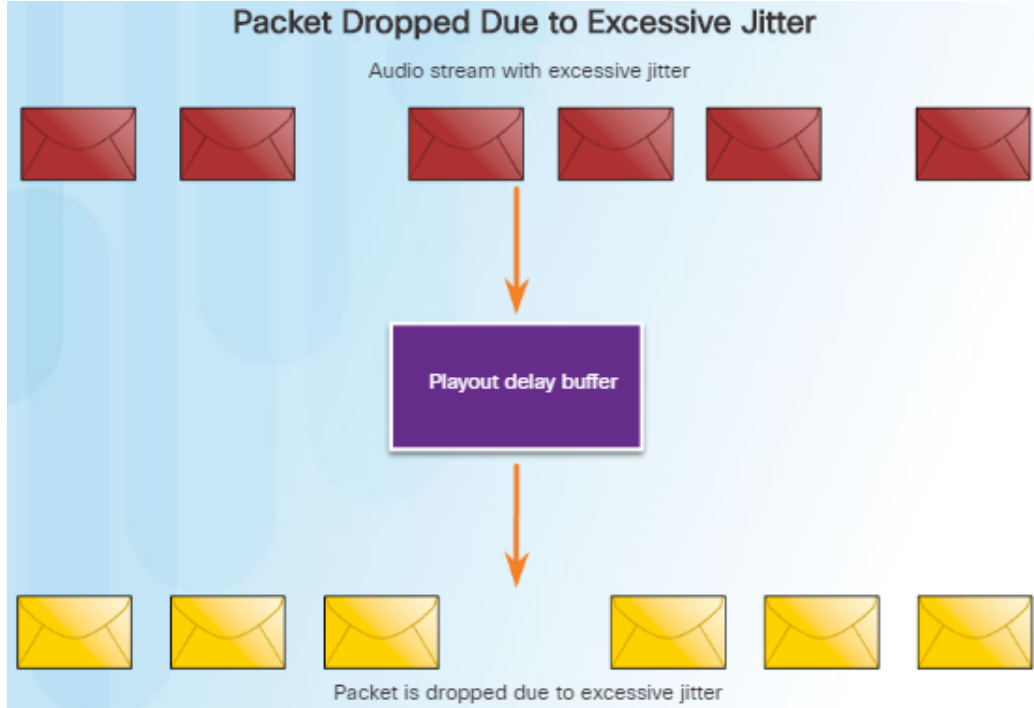
Packet Loss



- Without any QoS mechanisms in place, packets are processed in the order in which they are received.
 - When congestion occurs, network devices will drop packets.
 - This includes time-sensitive video and audio packets.
- For example, when a router receives a digital audio stream for VoIP, it must compensate for the jitter that is encountered.
 - The mechanism that handles this function is the playout delay buffer.
 - The playout delay buffer must buffer these packets and then play them out in a steady stream.
 - The digital packets are later converted back to an analog audio stream.

Network Transmission Quality

Packet Loss (Cont.)



- If the jitter is so large that it causes packets to be received out of the range of this buffer, the out-of-range packets are discarded and dropouts are heard in the audio.
- For losses as small as one packet, the digital signal processor (DSP) interpolates what it thinks the audio should be and no problem is audible to the user.
- However, when jitter exceeds what the DSP can handle, audio problems are heard.
- In a properly designed network, voice packet loss should be zero
- Network engineers use QoS mechanisms to classify voice packets for zero packet loss.

Traffic Characteristics

Network Traffic Trends



- In the early 2000s, the predominant types of IP traffic were voice and data.
- Voice traffic has a predictable bandwidth need and known packet arrival times.
- Data traffic is not real-time and has an unpredictable bandwidth need.
- More recently, video traffic has become increasingly important to business communications and operations.
- According to the Cisco Visual Networking Index (VNI), video traffic represented 67% of all traffic in 2014. By 2019, video will represent 80% of all traffic.
- The type of demands that voice, video, and data traffic place on the network are very different.


Traffic Characteristics

Voice

Voice Traffic Characteristics

Voice

- Smooth
- Benign
- Drop sensitive
- Delay sensitive
- UDP priority



The illustration shows a blue IP phone with 'IP' on its side. To its right is a simple bar chart with a vertical y-axis and a horizontal x-axis, featuring a single blue bar.

One-Way Requirements

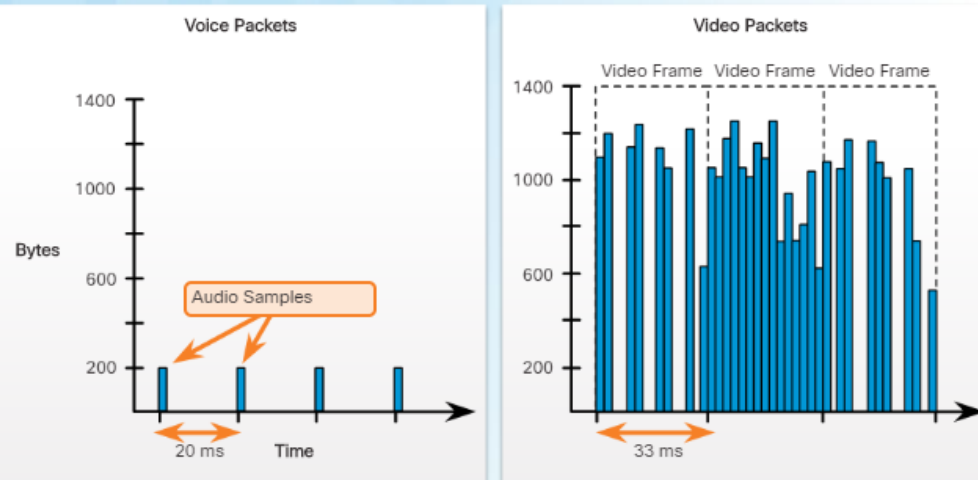
- Latency ≤ 150 ms
- Jitter ≤ 30 ms
- Loss $\leq 1\%$
- Bandwidth (30 - 128 Kb/s)

- Voice traffic is predictable and smooth.
- However, voice traffic is very sensitive to delay and dropped packets; there is no reason to retransmit voice if packets are lost.
- Voice packets must receive a higher priority than other types of traffic.
- Cisco products use the RTP port range 16384 to 32767 to prioritize voice traffic.
- Voice can tolerate a certain amount of latency, jitter, and loss without any noticeable effects.
- Latency should be no more than 150 ms.
- Jitter should be no more than 30 ms.
- Voice packet loss should not exceed 1%.

Traffic Characteristics

Video

Voice and Video Sampling Comparison



- Compared to voice, video is less resilient to loss and has a higher volume of data per packet as shown above.
 - Notice how voice packets arrive every 20 ms and are 200 bytes.
 - In contrast, the number and size of video packets varies every 33 ms based on the content of the video.

- Without QoS and a significant amount of extra bandwidth capacity, video quality typically degrades.
- The picture appears blurry, jagged, or in slow motion. The audio portion may become unsynchronized with the video.
- Video Traffic Characteristics:
 - Video – Bursty, greedy, drop sensitive, delay sensitive, UDP priority
 - One-Way Requirements:
 - Latency $\leq 200 - 400$ ms
 - Jitter $\leq 30 - 50$ ms
 - Loss $\leq 0.1 - 1\%$
 - Bandwidth (384 Kb/s – 20+ Mb/s)

Traffic Characteristics

Data

Data Traffic Characteristics

Data

- Smooth/bursty
- Benign/greedy
- Drop insensitive
- Delay insensitive
- TCP retransmits



- Most applications use either TCP or UDP. Unlike UDP, TCP performs error recovery.
- Data applications that have no tolerance for data loss, such as email and web pages, use TCP to ensure packets will be resent in the event they are lost.
- Some TCP applications, such as FTP, can be very greedy, consuming a large portion of network capacity.
- Although data traffic is relatively insensitive to drops and delays compared to voice and video, a network administrator still needs to consider the quality of the user experience.

Factors to Consider for Data Delay

Factor	Mission Critical	Not Mission Critical
Interactive	Prioritize for the lowest delay of all data traffic and strive for a 1 to 2 seconds response time.	Applications could benefit from lower delay.
Not interactive	Delay can vary greatly as long as the necessary minimum bandwidth is supplied.	Gets any leftover bandwidth after all voice, video, and other data application needs are met.

- Two factors that need to be determined:
 - Does the data come from an interactive application?
 - Is the data mission critical?

Queuing Algorithms

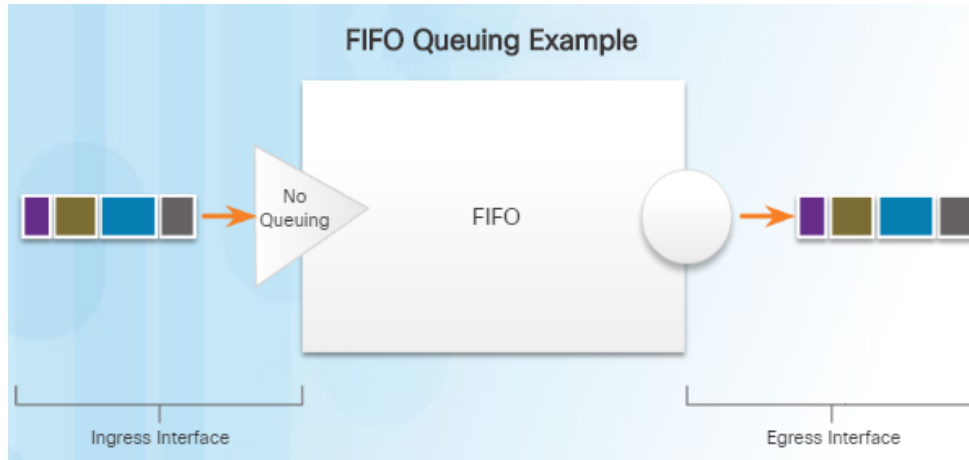
Queuing Overview



- The QoS policy implemented by the network administrator becomes active when congestion occurs on the link.
- Queuing is a congestion management tool that can buffer, prioritize, and if required, reorder packets before being transmitted to the destination.
- This course will focus on the following queuing algorithms:
 - First-In, First-Out (FIFO)
 - Weighted Fair Queuing (WFQ)
 - Class-Based Weighted Fair Queuing (CBWFQ)
 - Low Latency Queuing (LLQ)

Queuing Algorithms

First In First Out (FIFO)

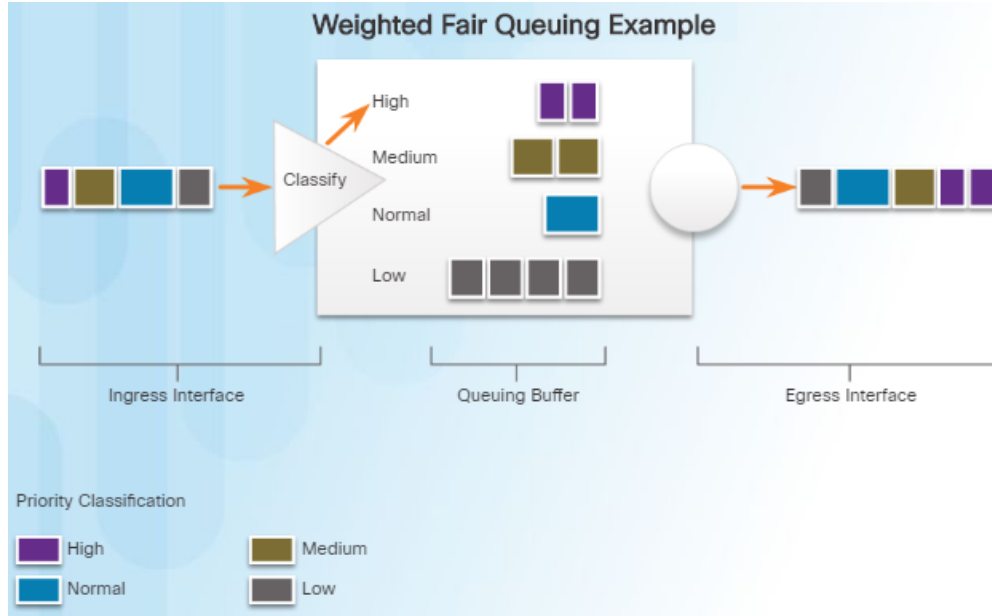


- FIFO is effective for large links that have little delay and minimal congestion
- If your link has very little congestion, FIFO queuing may be the only queuing you need to use.

- FIFO queuing, also known as first-come, first-served queuing, involves buffering and forwarding of packets in the order of arrival.
- FIFO has no concept of priority or classes of traffic and consequently, makes no decision about packet priority.
- There is one queue and all packets are treated equally.
- When FIFO is used, important or time-sensitive traffic can be dropped when congestion occurs on the router or switch interface.
- When no other queuing strategies are configured, FIFO is used on serial interfaces at E1 (2.048 Mbps) and below.

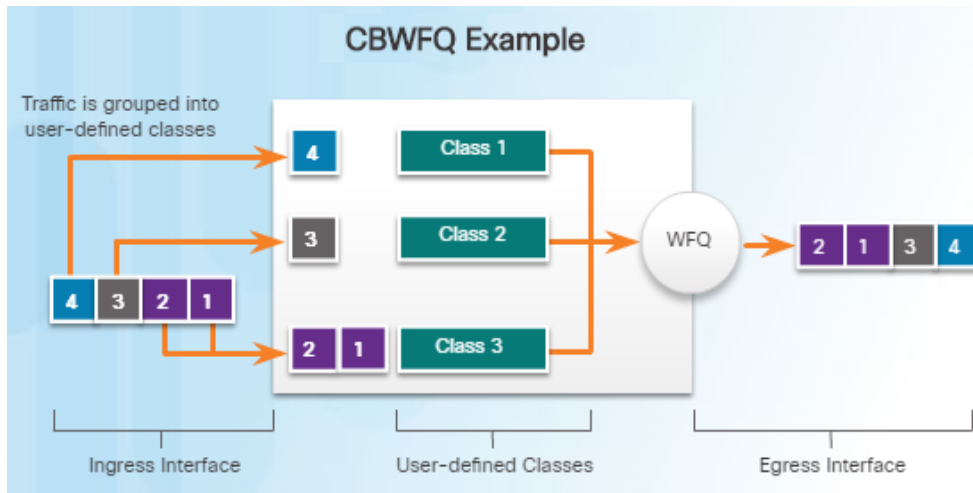
Queuing Algorithms

Weighted Fair Queuing (WFQ)



- WFQ is an automated scheduling method that provides fair bandwidth allocation to all network traffic.
- WFQ applies priority, or weights, to identified traffic and classifies it into conversations or flows.
- WFQ then determines how much bandwidth each flow is allowed relative to other flows.
- WFQ schedules interactive traffic to the front of a queue to reduce response time. It then shares the remaining bandwidth among high-bandwidth flows.
- WFQ classifies traffic into different flows based on packet header addressing, including source/destination IP addresses, MAC addresses, port numbers, protocols, and type of service (ToS) values.

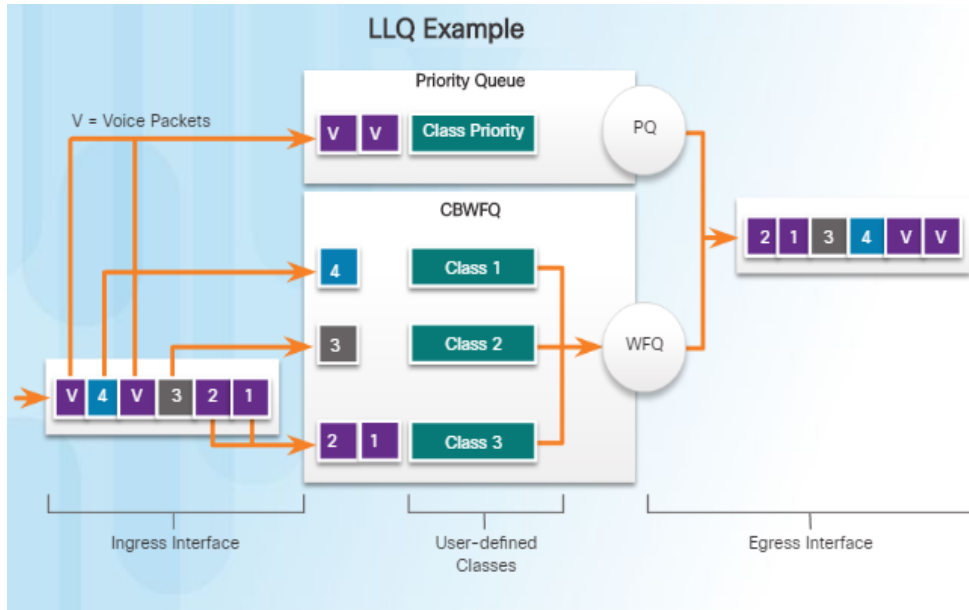
Class-Based Weighted Fair Queuing (WFQ)



- Packets that match the criteria for a class constitute the traffic for that class. A FIFO queue is reserved for each class, and traffic belonging to a class is directed to the queue.
- CBWFQ extends the standard WFQ functionality to provide support for user-defined traffic classes.
- You define traffic classes based on match criteria including protocols, ACLs, and input interfaces.
- When a class has been defined according to its match criteria, you can assign it characteristics.
 - To characterize a class, you assign it bandwidth, weight, and maximum packet limit.
 - The bandwidth assigned to a class is the guaranteed bandwidth delivered to the class during congestion.

Queuing Algorithms

Low Latency Queuing (LLQ)



- The LLQ feature brings strict priority queuing (PQ) to CBWFQ which reduces jitter in voice conversations. See the figure to the left.
- Strict PQ allows delay-sensitive data such as voice to be sent before packets in other queues.
- Without LLQ, CBWFQ provides WFQ based on defined classes with no strict priority queue available for real-time traffic.
 - All packets are serviced fairly based on weight.
 - This scheme poses problems for voice traffic that is largely intolerant of delay.
- With LLQ, delay-sensitive data is sent first, before packets in other queues are treated.
- LLQ allows delay-sensitive data such as voice to be sent first giving it preferential treatment.

Selecting an Appropriate QoS Policy Model

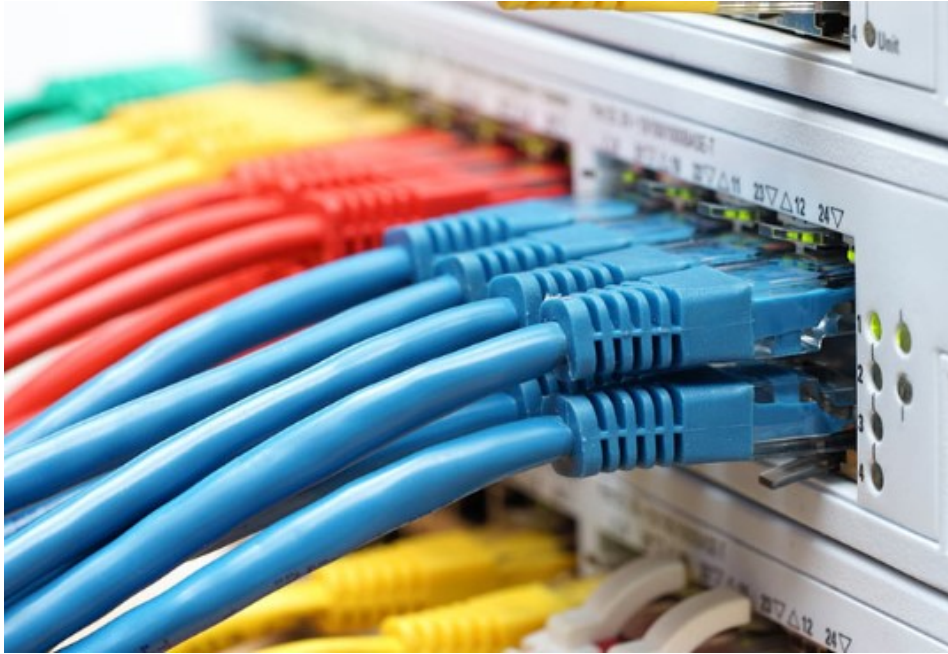
Models for Implementing QoS

Model	Description
Best-effort model	<ul style="list-style-type: none">Not really an implementation as QoS is not explicitly configured.Use when QoS is not required.
Integrated services (IntServ)	<ul style="list-style-type: none">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)	<ul style="list-style-type: none">Provides high scalability and flexibility in implementing QoS.Network devices recognize traffic classes and provide different levels of QoS to different traffic classes.

- How can QoS be implemented in a network? The three models for implementing QoS are these:
 - Best-effort model
 - Integrated services (IntServ)
 - Differentiated Services (DiffServ)
- The table in the figure to the left summarizes these three models.
- QoS is implemented in a network using either or both of these:
 - IntServ – provides the highest guarantee of QoS, but is resource-intensive
 - DiffServ – less resource intensive and more scalable

QoS Implementation Techniques

Avoiding Packet Loss



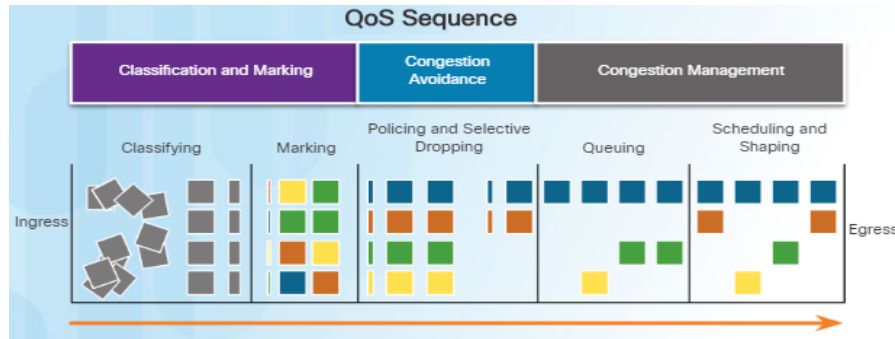
- Packet loss is usually the result of congestion on an interface.
- Most TCP applications experience slowdown because TCP automatically adjusts to network congestion.
 - Some applications do not use TCP and cannot handle drops (fragile flows).
- The following approaches can prevent drops in sensitive applications:
 - Increase link capacity to ease or prevent congestion.
 - Guarantee enough bandwidth and increase buffer space to accommodate bursts of traffic from fragile flows – WFQ, CBWFQ and LLQ.
 - Prevent congestion by dropping lower-priority packets before congestion occurs – weighted random early detection (WRED).

QoS Implementation Techniques

QoS Tools

Tools for Implementing QoS

QoS Tools	Description
Classification and marking tools	<ul style="list-style-type: none">Sessions, or flows, are analyzed to determine what traffic class they belong to.Once determined, the packets are marked.
Congestion avoidance tools	<ul style="list-style-type: none">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	<ul style="list-style-type: none">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.



- There are three categories of QoS tools:
 - Classification and marking tools
 - Congestion avoidance tools
 - Congestion management tools
- Ingress packets (gray squares) are classified and their respective IP header is marked (colored squares). To avoid congestion, packets are then allocated resources based on defined policies.
- Packets are then queued and forwarded out the egress interface based on their defined QoS shaping and policing policy.
- Classification and marking can be done on ingress or egress, whereas other QoS actions such as queuing and shaping are usually done on egress.

QoS Implementation Techniques

Classification and Marking

Traffic Marking for QoS

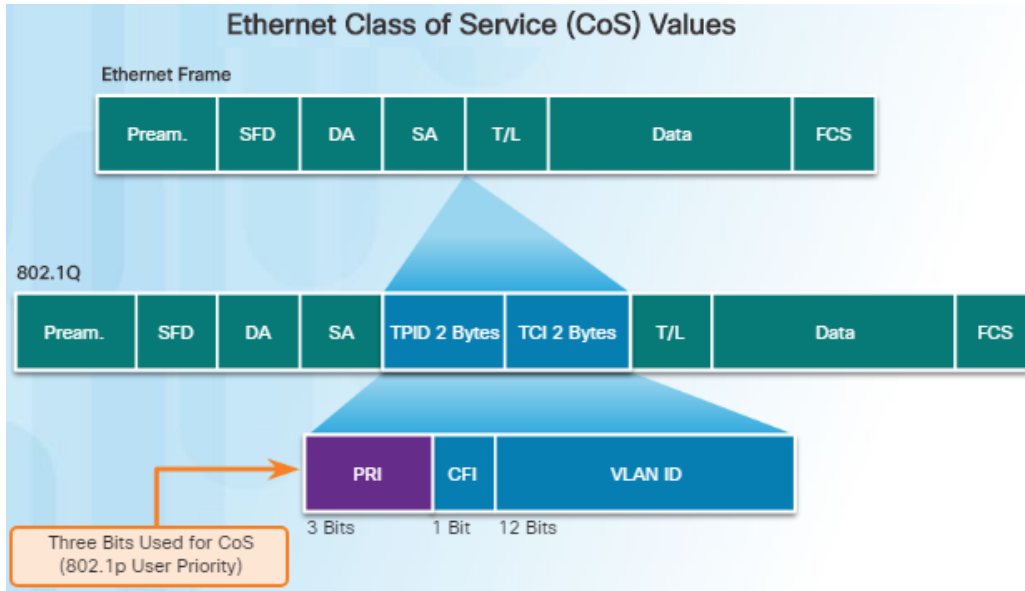
QoS Tools	Layer	Marking Field	Width in Bits
Ethernet (802.1Q, 802.1p)	2	Class of Service (CoS)	3
802.11 (Wi-Fi)	2	Wi-Fi Traffic Identifier (TID)	3
MPLS	2	Experimental (EXP)	3
IPv4 and IPv6	3	IP Precedence (IPP)	3
IPv4 and IPv6	3	Differentiated Services Code Point (DSCP)	6

- The table in the figure describes some of the marking fields used in various technologies. Consider the following points when deciding to mark traffic at Layers 2 or 3:
 - Layer 2 marking of frames can be performed for non-IP traffic.
 - Layer 2 marking of frames is the only QoS option available for switches that are not “IP aware”.
 - Layer 3 marking will carry the QoS information end-to-end.

- A packet has to be classified before it can have a QoS policy applied to it.
- Classification and marking allows us to identify, or “mark” types of packets.
- Classification determines the class of traffic to which packets or frames belong. Policies can not be applied unless the traffic is marked.
- Methods of classifying traffic flows at Layer 2 and 3 include using interfaces, ACLs, and class maps.
- Marking requires the addition of a value to the packet header and devices that receive the packet look at this field to see if it matches a defined policy.
- Marking should be done as close to the source as possible and this establishes the trust boundary.

QoS Implementation Techniques

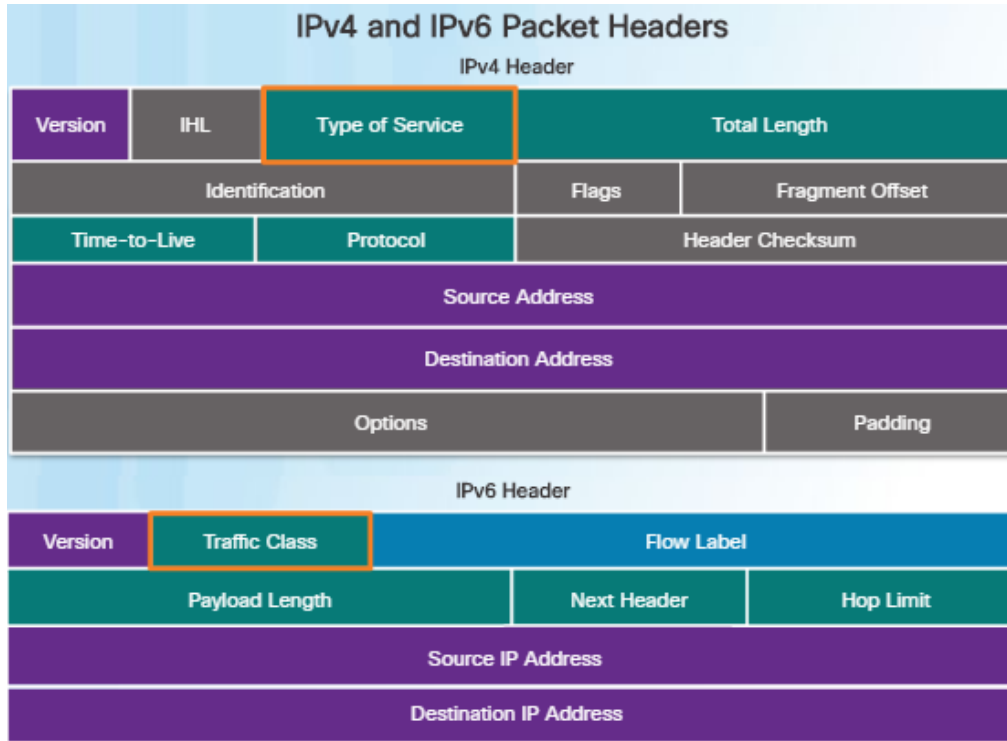
Marking at Layer 2



- 802.1Q is the IEEE standard that supports VLAN tagging at Layer 2 on Ethernet networks.
- When 802.1Q is implemented, two fields are added to the Ethernet Frame and are inserted following the source MAC address field as shown in the figure to the left.
- The 802.1Q standard includes the QoS prioritization scheme known as IEEE 802.1p. The standard uses the first three bits in the Tag Control Information (TCI) field and identifies the CoS markings.
- These three bits allow eight levels of priority (0-7).

QoS Implementation Techniques

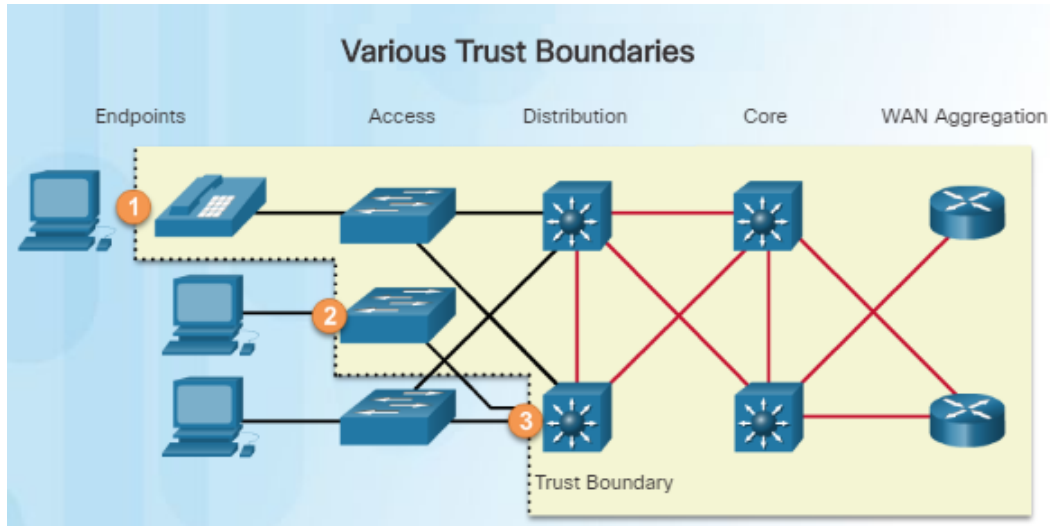
Marking at Layer 3



- IPv4 and IPv6 specify an 8-bit field in their packet headers to mark packets.
 - IPv4 – Type of Service (ToS) field
 - IPv6 – Traffic Class field
- These fields are used to carry the packet marking assigned by the QoS classification tools. Forwarding devices refer to this field and forward the packets based on the QoS policy.
- RFC 2474 redefines the ToS field by renaming and extending the IPP field. The new field has 6-bits allocated for QoS called the differentiated services code point (DSCP) field.
- These six bits offer a maximum of 64 possible classes of service.

QoS Implementation Techniques

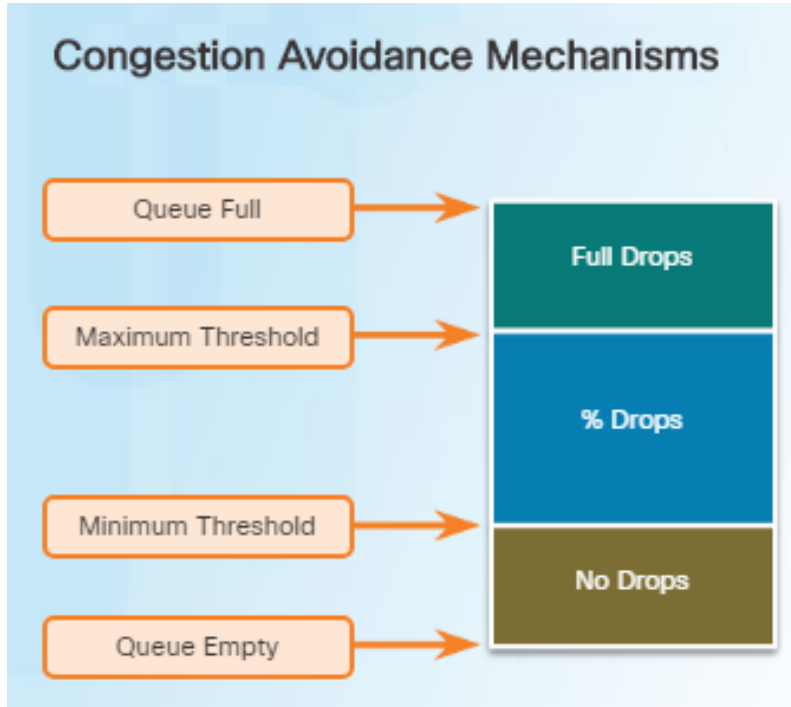
Trust Boundaries



- Where should markings occur?
- Traffic should be classified and marked as close to its source as possible.
- This defines the trust boundary as shown in the figure.
 - Trusted endpoints have the capabilities and intelligence to mark application traffic to the appropriate Layer 2 CoS or Layer 3 DSCP values. Examples of trust endpoints include IP phones, wireless access points, and videoconferencing systems.
 - Secure endpoints can have traffic marked at the Layer 2 switch.
 - Traffic can also be marked at Layer 3 switches and routers.
- Re-marking of traffic is typically necessary.

QoS Implementation Techniques

Congestion Avoidance



- Congestion avoidance tools monitor network traffic loads in an effort to anticipate and avoid congestion at common network bottlenecks before congestion becomes a problem.
- Congestion avoidance is achieved through packet dropping.
- These tools monitor the average depth of the queue.
 - For example, when 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.

QoS Implementation Techniques

Shaping and Policing

Shaping Traffic Example



Policing Traffic Example



- Traffic shaping and policing are two mechanisms provided by the Cisco IOS QoS software to prevent congestion.
- Traffic shaping retains excess packets in a queue and then schedules the excess for later transmission over increments of time.
 - The result of traffic shaping is a smoothed packet output rate as shown in the figure.
 - Shaping requires sufficient memory.
- Shaping is used on outbound traffic.
- Policing is commonly implemented by service providers to enforce a contracted customer information rate (CIR).
- Policing either drops or remarks excess traffic.
- Policing is often applied to inbound traffic.