QoS Concepts:

- Imagine driving on heavily congested road & you are in a rush to meet a friend for dinner.
- You hear the siren and see the lights of an ambulance behind you on the congested road.
- You need to move off the road to let the ambulance through on heavily congested road.
- Ambulance getting to hospital takes priority over you getting to the restaurant on time.
- Like ambulance taking priority, some forms of network traffic need priority over others.

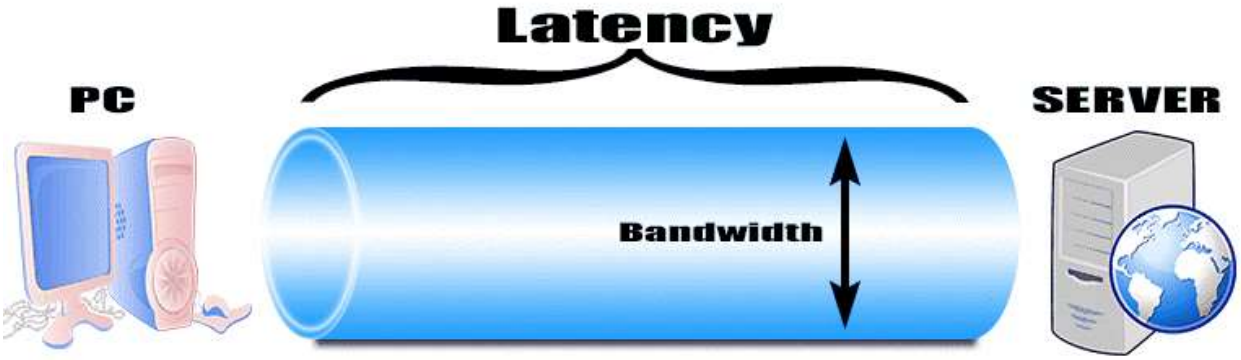
Quality of Service (QoS):

- Quality of Service (QoS) basically assists in the management of packet loss, delay and jitter.
- QoS refers to any technology manages data traffic to reduce packet loss, latency & jitter.
- QoS controls & manages network resources by setting priorities for specific types of data.
- QoS is a feature of router & switch prioritizes traffic that more important traffic pass first.
- Quality of Service (QoS), result is a performance improvement for critical network traffic.
- QoS equipment is useful with VoIP phones or in LANs with high volumes of the local traffic.
- QoS is managing network resources to reduce packet loss & lower network jitter & latency.
- Technology manage resources by assigning various types of network data different priority.
- Organizations use QoS to meet traffic requirements of sensitive applications, such as voice.
- From the perspective of the OSI Reference model, QoS solutions focus primarily on layer 7.
- Quality of service is the description or measurement of the overall performance of a service.
- QoS is about using tools to change how the router or switch deals with the different packets

QoS Terminologies:

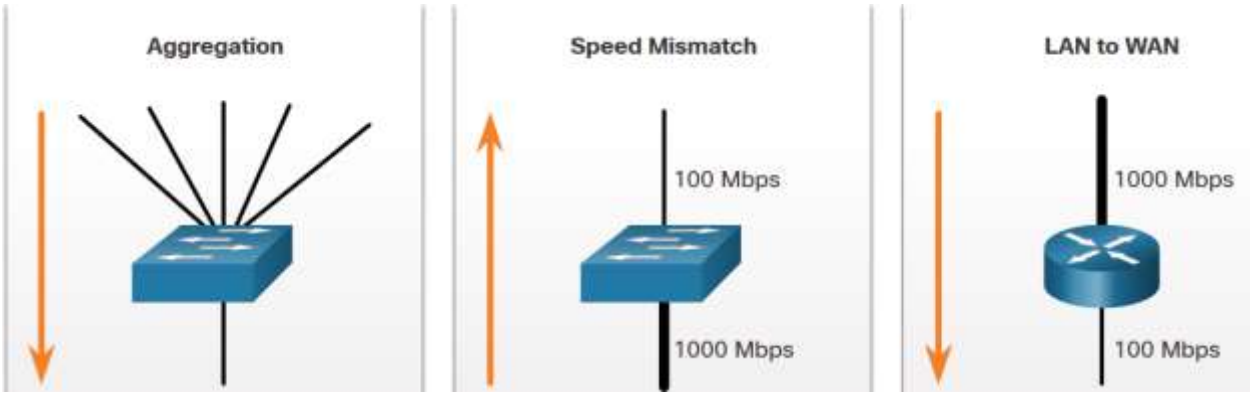
Bandwidth:

- Network bandwidth refers to the speed of a link or interface, in bits per second (bps).
- Think bandwidth as speed, it helps to also think of bandwidth as the capacity of the link.
- Network bandwidth is capacity of link how many bits can be sent over the link per second.
- Network bandwidth is measured in number of bits that can be transmitted in single second.
- Network bandwidth is measured in number of bits transmitted int bits per second (bps).
- Network device may be described as having capability to perform at 10 gigabits per second.



Congestion:

- In the world of Computer Networking, Network congestion causes the Delay or Latency.
- Interface experiences congestion when it is presented with more traffic than it can handle.
- Network congestion points are ideal candidates for Quality of Service (QoS) mechanisms.



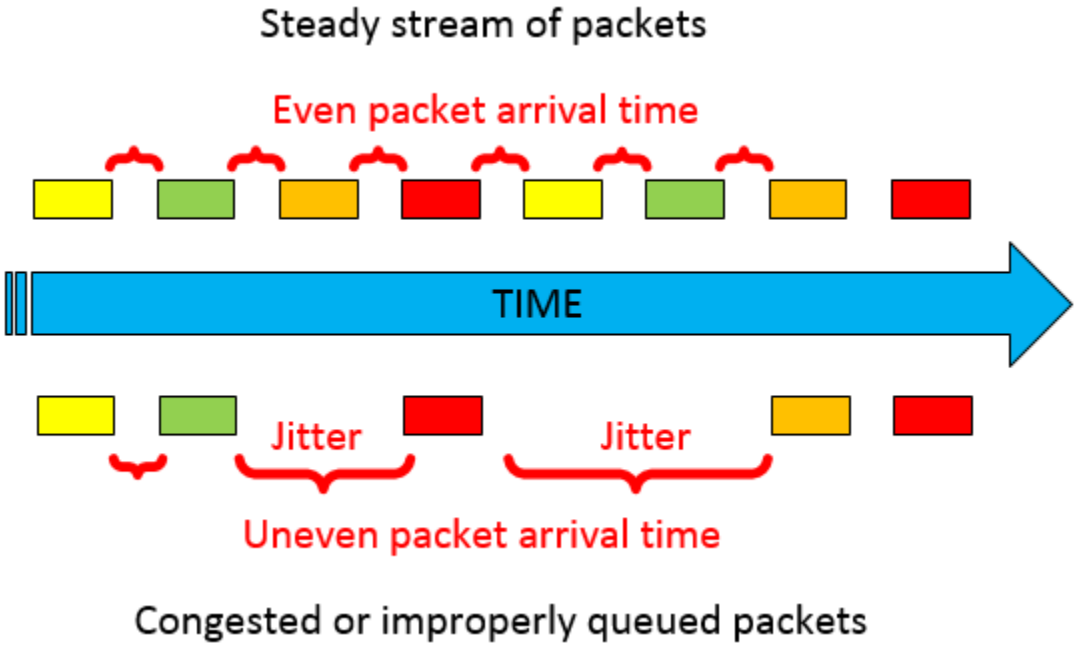
Delay:

- Delay or latency refers to time it takes for packet to travel from source to the destination.
- In Quality of Service there are two types of delays one is fixed delay & other is variable.
- Quality of Service (QoS), a fixed delay is a specific amount of time a specific process takes.
- Example of Fixed Delay, such as how long it takes to place a bit on the transmission media.
- Quality of Service Variable delay takes an unspecified amount of time & affected by factors.
- Example of Variable Delay is affected by factors such how much traffic is being processed.

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	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.

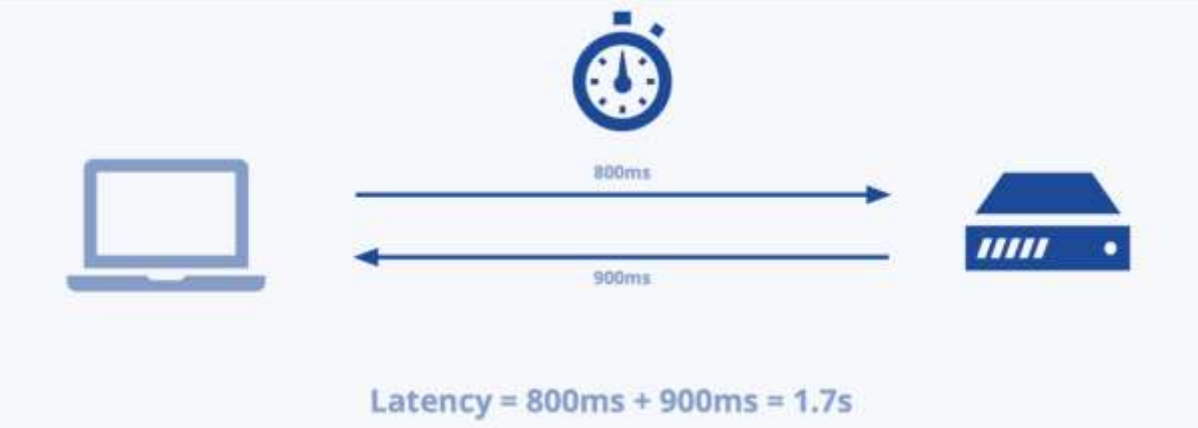
Jitter:

- In Quality of Service Jitter is defined as a variation in the delay of received packets.
- In Quality of Service Jitter is the difference between the delays of the IP packets.
- In Quality of Service Jitter is the variation in delay of a streaming application.
- When the delay of latency of the variate through the network it causes jitter.
- Jitter is the irregular time delay in the sending of the data packets over the network.
- At sending side, packets are sent in continuous stream with packets spaced evenly apart.
- Due to the network congestion, the improper queuing, or the configuration errors.
- In Quality of Service delay between each packet can vary instead of remaining constant.
- Delay & jitter need to be controlled minimized to support real-time & interactive traffic.



Latency:

- Latency is the time required by a data packet to reach the destination from the source.
- Broadly, the latency is the time elapsed between the execution of the two events.
- Latency is simply time required to process messages at both source and destination ends.
- Latency is imply time required to process and the delays generated in the given network.
- Latency is amount of time it takes for packet to travel from point A to point B round-trip.
- In Quality of Service (QoS), one-way end-to-end delay, also known as the network latency.



Packet Loss:

- Without any QoS mechanisms, packets are processed in order in which they are received.
- When congestion occurs, network devices such as routers and switches can drop packets.
- This means that time-sensitive packets, such as real-time video and voice, will be dropped.
- With same frequency as data that is not time-sensitive, such as email and web browsing.
- The Packet loss is a very common cause of the voice quality problems on the IP network.
- With Qualify of Service, in a properly designed network, packet loss should be near zero.

Voice:

- In the world of Computer Network voice is very sensitive to delays and dropped packets.
- It makes no sense and not possible to re-transmit or re-send the voice if packets are lost.
- Therefore, voice packets must receive a higher priority than other types of network traffic.
- Voice can tolerate certain amount of latency, jitter, and loss without any noticeable effects.

Videos:

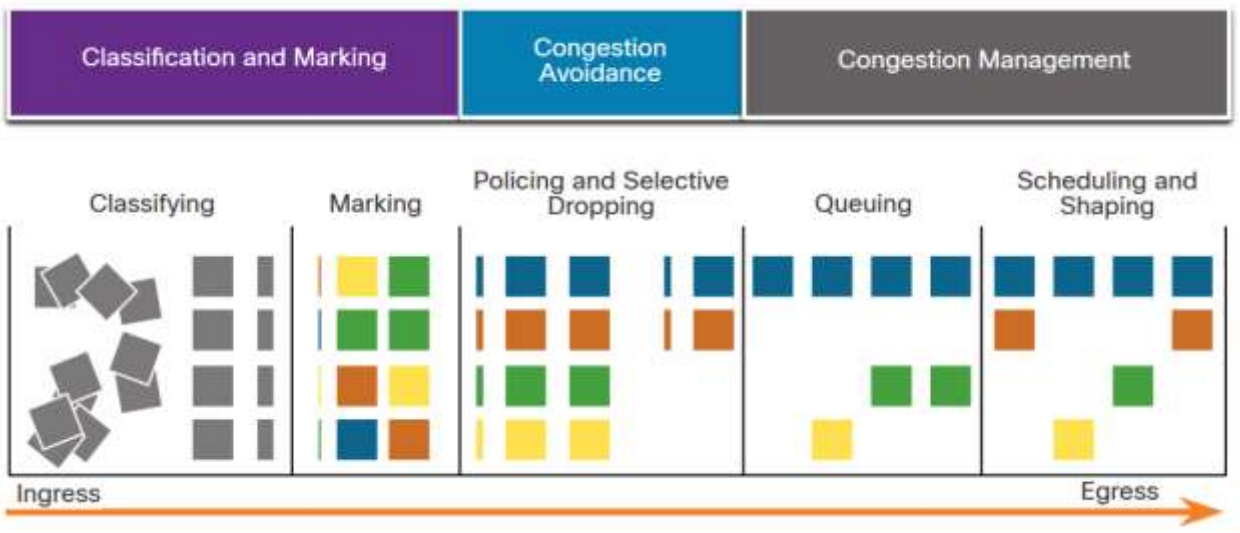
- Without QoS & significant amount of extra bandwidth capacity, video quality degrades.
- Without Quality of Service (QoS) The picture appears blurry, jagged, or in slow motion.
- Also, the audio portion of the feed may become unsynchronized with the video picture.
- Video traffic tends to be unpredictable, inconsistent, and bursty compared to voice traffic.
- Compared to voice, video is less resilient to loss and has a higher volume of data per packet.

Data:

- 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 that, if the data packets are lost in the transit, they will be resent.
- Data traffic can be smooth, Network control traffic is usually smooth & predictable.
- However, some TCP applications can consume a large portion of network capacity.
- FTP will consume as much bandwidth as it can get when you download a large file.
- Although data traffic is relatively insensitive to drops & delays compared to voice etc.

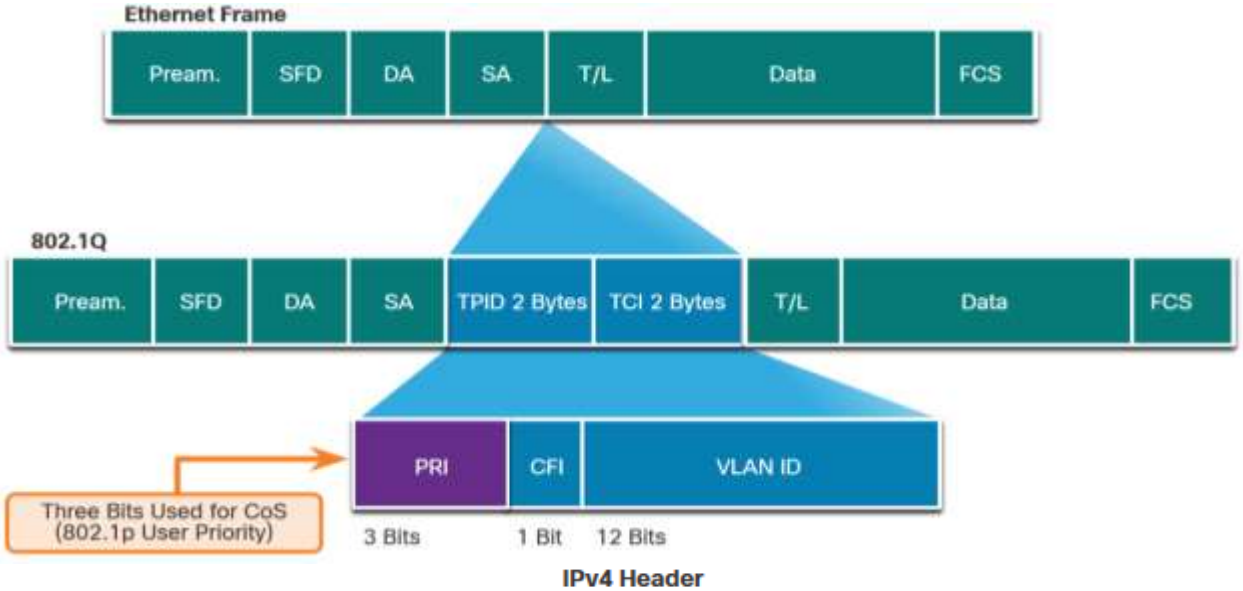
Quality of Service Tools:

QoS Tools	Description
Classification and marking tools	Sessions, or flows, are analyzed to determine what traffic class they belong to. When the traffic class is 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.

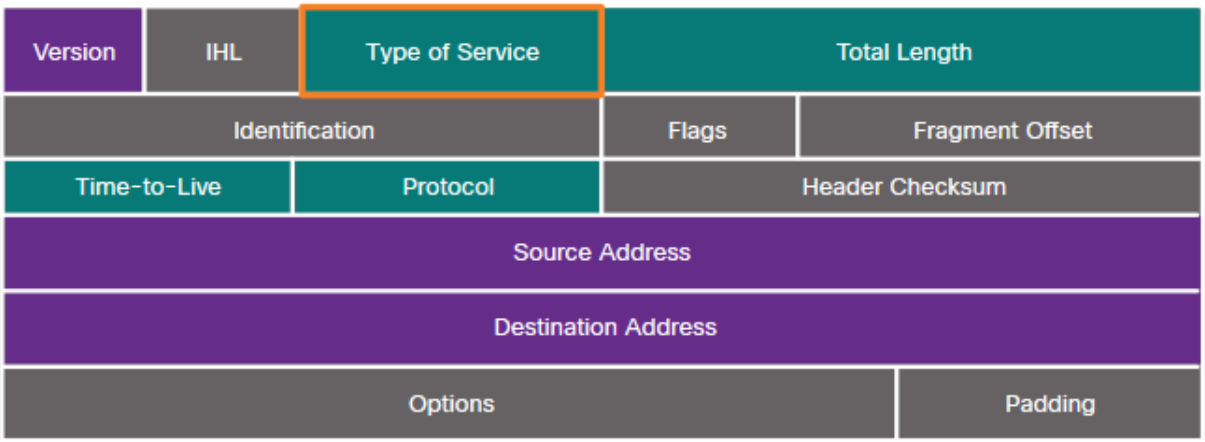


Classification and Marking:

- Before a packet can have a QoS policy applied to it, the packet has to be classified.
- Classification and marking allows us to identify or to mark types of the packets.
- Classification determines the class of traffic to which packets or frames belong.
- For Quality of Service only after traffic is marked can policies to be applied to it.
- How a packet is classified depends on the Quality of Service QoS implementation.
- Methods of classifying traffic flows at L2 & 3 include using interfaces, ACLs & class maps.
- Traffic can be classified at L4 to 7 using Network Based Application Recognition (NBAR).
- NBAR is classification & protocol discovery feature of Cisco IOS that works with QoS.
- In Quality of Service, marking means that we are adding a value to the packet header.
- The Devices receiving the packet look at this field to see if it matches a defined policy.
- Marking should be close to the source device as possible to establishes trust boundary.
- 802.1Q is the IEEE standard that supports VLAN tagging at Layer 2 on Ethernet networks.
- The 802.1Q standard also includes the QoS prioritization scheme known as IEEE 802.1p.
- In the IPv4 and IPv6 specify an 8-bit field in their packet headers to mark the packets.

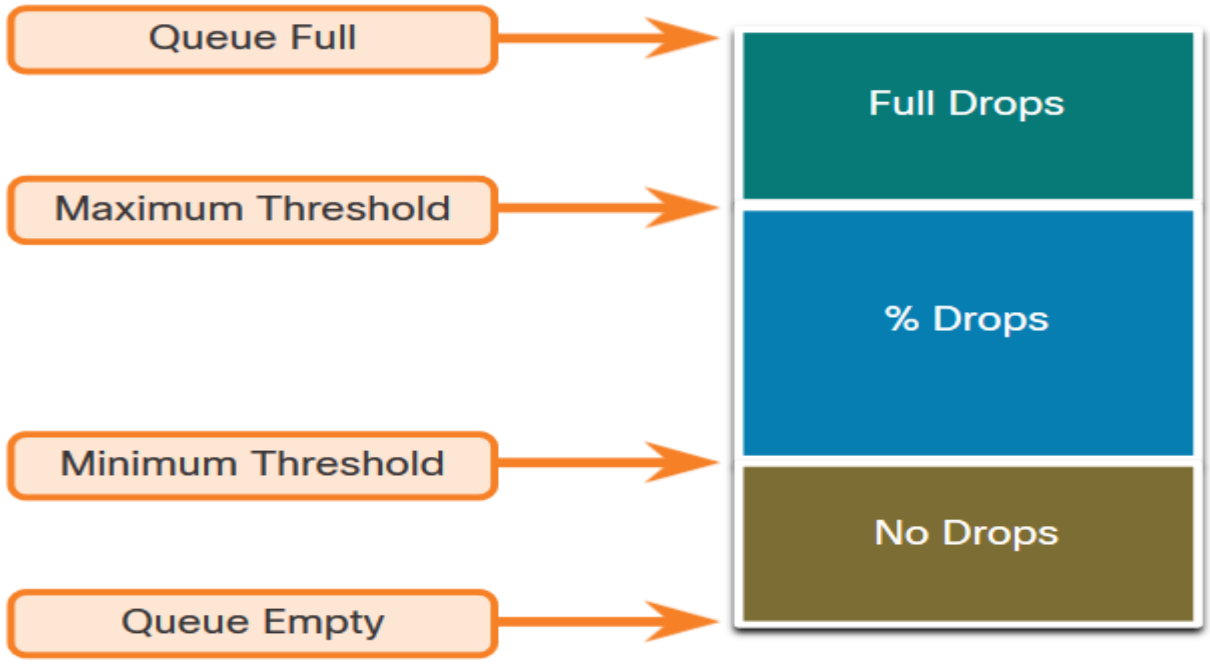


IPv4 Header



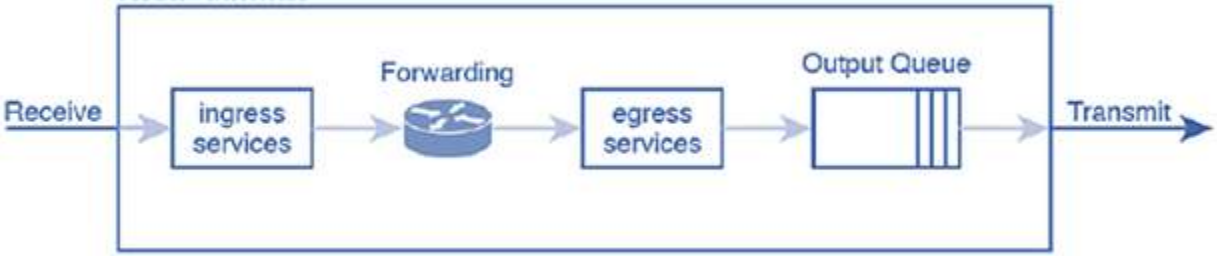
Congestion Avoidance:

- Addresses how to manage packet loss that occurs when network devices get too busy.
- Monitor network traffic loads in effort to anticipate & avoid congestion common network.
- Congestion Avoidance, internetwork bottlenecks before congestion becomes a problem.
- It prevents network from being overloaded using monitoring and packet discarding policy.
- In Congestion Avoidance when 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.
- In congestion avoidance when the maximum threshold is passed, all packets are dropped.
- IOS includes weighted random early detection as possible congestion avoidance solution.
- There is WRED and DWRED are the Cisco IOS QoS congestion avoidance features available.



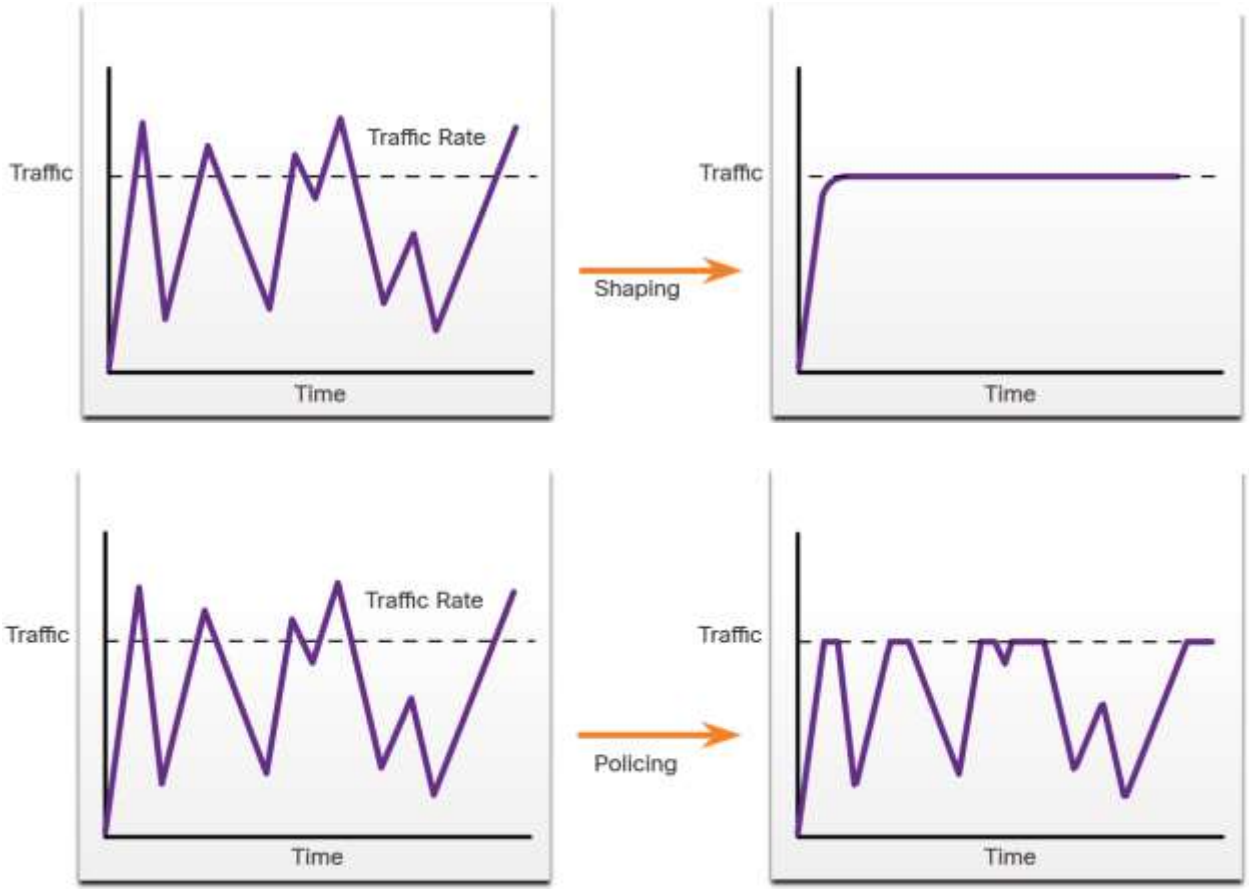
Congestion Management:

- Describes scheduling of packets to give one type of packet priority over another.
- Congestion management includes queuing and scheduling methods & techniques.
- Whenever, where excess traffic is buffered or queued and sometimes may dropped.
- In Congestion Management, while it waits to be sent out an egress interface or Link.
- Set of Quality of Service features that handle the queuing and scheduling of the traffic.
- Congestion management techniques provide some control of the order of transmission.



Traffic Shaping & Policing:

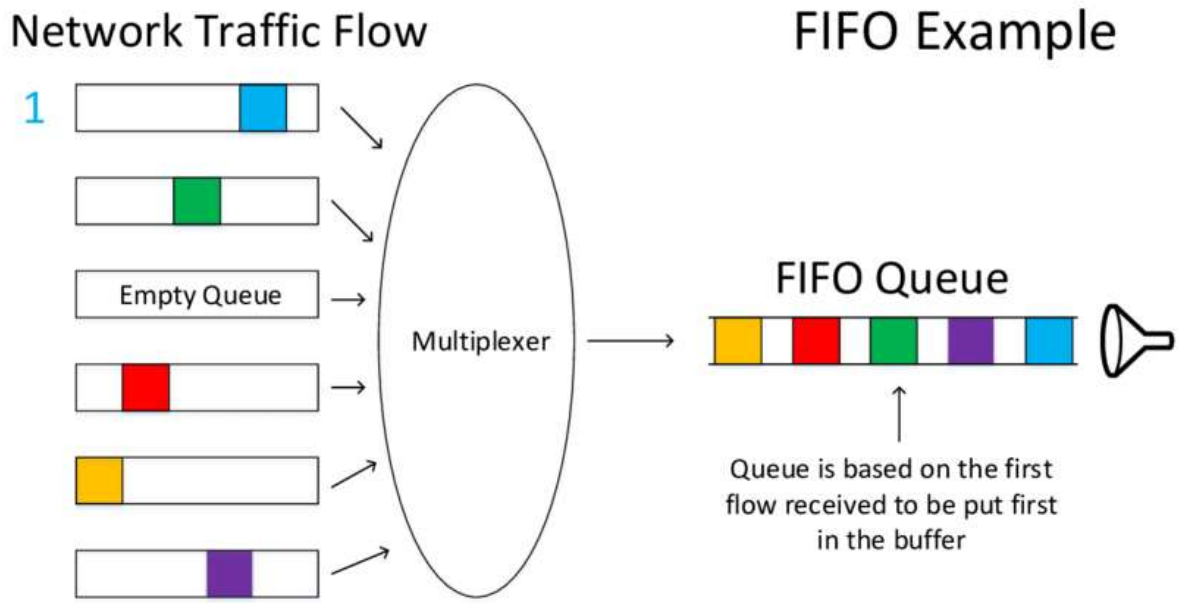
- Traffic shaping & traffic policing are two mechanisms provided by Cisco IOS QoS.
- Traffic shaping & traffic policing are Cisco IOS QoS software to prevent congestion.
- Shaping implies existence of queue & of sufficient memory to buffer delayed packets.
- In Quality of Service traffic Policing does not sufficient memory to buffer delay packet.
- In Quality of Service, ensure that you have sufficient memory when enabling shaping.
- Shaping requires a scheduling function for later transmission of any delayed packets.
- This scheduling function allows you to organize the shaping queue into different queues.
- Shaping is outbound concept packets going out an interface get queued & can be shaped.
- In Quality of Service (QoS), Traffic Policing is applied to inbound traffic on an interface.
- Quality of Service Policing is commonly implemented by service providers to enforce CIR.
- Between the ISP and the customer, there is a pre-defined contracted rate called the CIR.
- ISP will police ingress traffic non-conforming is caught by policer & dropped or marked.
- Customer typically doesn't want any traffic dropped so shaping is done on egress interface.
- Policing, anything is above drop it, shaping anything is above store it in memory & delay it.
- Shaping is QoS technique use to enforce lower bitrates what physical interface capable of.
- When we use shaping, we will buffer the traffic to a certain bitrate while policing will drop.



Queuing Features:

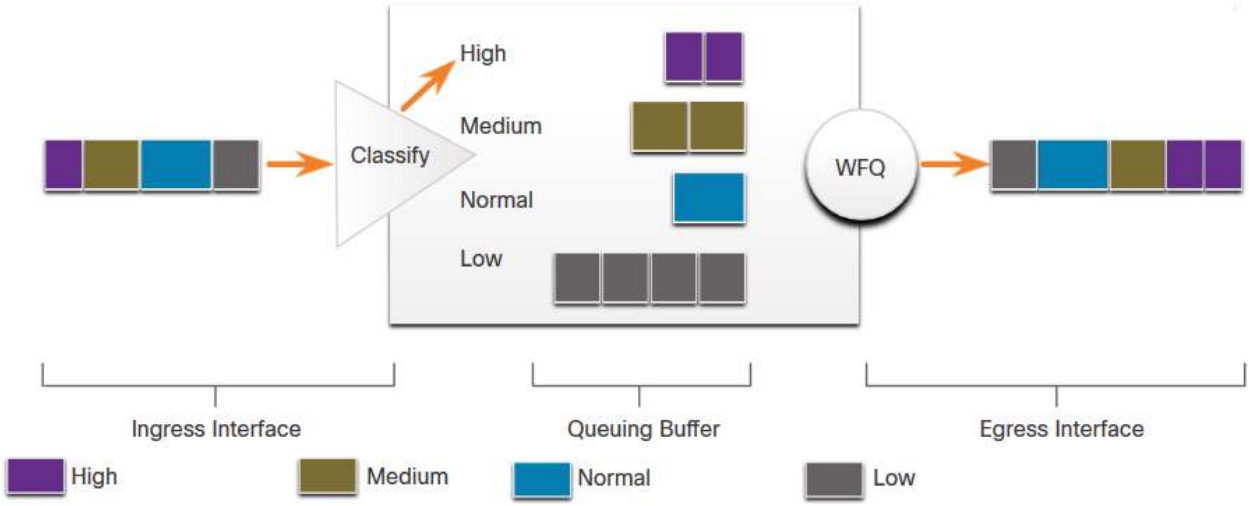
First in, First out (FIFO):

- In the world of Quality of Service (QoS), FIFO is an acronym for First in First Out.
- This expression describes the principle of a queue or first-come-first-serve behavior.
- What comes in first is handled first, what comes in next waits until the first is finished.
- Its simplest form, First in First Out (FIFO) queuing, also known as first-come, first-served.
- First in First Out (FIFO) queuing, buffers and forwards packets in the order of their arrival.
- It has no concept of priority or classes of traffic & makes no decision about packet priority.
- In First in First Out (FIFO), there is only one queue, and all the packets are treated equally.
- Packets are sent out an interface in the order in which they arrive When the FIFO is used.
- Important or time-sensitive traffic can be dropped when there is congestion on device.



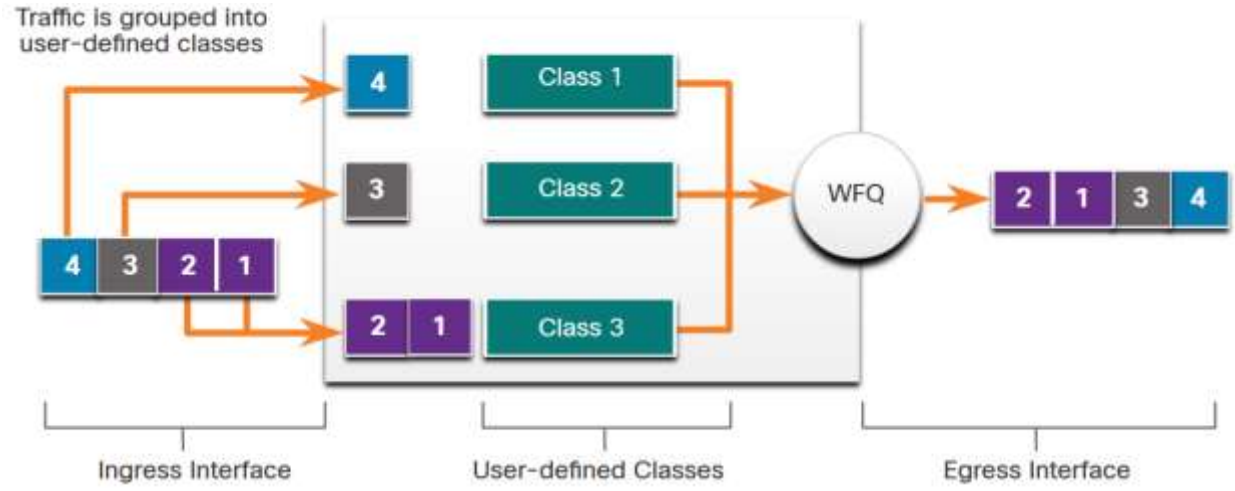
Weighted Fair Queuing (WFQ):

- WFQ is automated scheduling method that provides fair bandwidth allocation to all traffic.
- Weighted Fair Queuing (WFQ) does not allow the classification options to be configured.
- WFQ applies priority, or weights, to identified traffic and classifies it into conversations.
- WFQ then determines how much bandwidth each flow is allowed relative to other flows.
- WFQ allows you to give low-volume, interactive traffic, such as Telnet sessions and voice.
- WFQ give Priority over high-volume traffic, such as File Transfer Protocol (FTP) sessions.
- When multiple file transfers flows are occurring, transfers are given comparable bandwidth.
- Weighted Fair Queuing classifies traffic into different flows based on packet header address.
- Including such as source and destination IP addresses, MAC, port no, protocol & ToS value.
- The Type of Service (ToS) value in internet Protocol header can be used to classify traffic.



Class-Based Weighted Fair Queuing (CBWFQ):

- It extends standard WFQ functionality to provide support for user-defined traffic classes.
- With CBWFQ, define traffic classes based on match criteria including protocols and ACLs.
- The Packets satisfying the match criteria for a class constitute the traffic for that class.
- FIFO queue is reserved for each class & traffic belonging to a class is directed to the queue.
- When a class has been defined according to its match criteria, can assign it characteristics.
- To characterize a class, you assign it the bandwidth, weight, and maximum packet limit.
- Bandwidth assigned to class is guaranteed bandwidth delivered to class during congestion.
- Class-Based Weighted Fair Queuing To characterize class, also specify queue limit for class.
- Which is the maximum number of packets allowed to accumulate in the queue for the class.
- Packets belonging to class are subject to bandwidth & queue limits that characterize class.



Low Latency Queuing (LLQ):

- The Low Latency Queuing (LLQ) feature brings strict priority queuing (PQ) to CBWFQ.
- Strict PQ allows delay-sensitive packets voice to be sent before packets in other queues.
- The LLQ provides strict priority queuing for CBWFQ, reducing jitter in voice conversations.
- Without LLQ, CBWFQ provides WFQ based on defined classes with no strict priority queue.
- Weight for a packet belonging to a specific class is derived from bandwidth assigned to class
- Therefore, bandwidth assigned to packets of class determines order which packets are sent.
- All packets are serviced fairly based on weight; no class of packets granted strict priority.
- This scheme poses problems for voice traffic that is largely intolerant of delay and delay.
- For the voice traffic, variations in delay introduce irregularities of transmission as jitter.
- LLQ allows delay-sensitive packets voice to be sent first before packets in other queues.
- LLQ giving delay-sensitive packets preferential treatment over other any network traffic.
- Although it is possible to classify various types of real-time traffic to strict priority queue.
- In LLQ Cisco recommends that only the voice traffic be directed to the priority queue.