QOS

Tags	
Klaar	JA
Week	Week 5

What is QOS
Queuing Algorithms
- FIFO (First In First Out)
- WFQ (Weighted Fair Queuing)
- CBWFQ (Class-Based Weighted Fair Queuing)
- LLQ (Low Latency Queuing)

Implementing QOS
- Best-Effort
- Integrated Services
- Differentiated Services

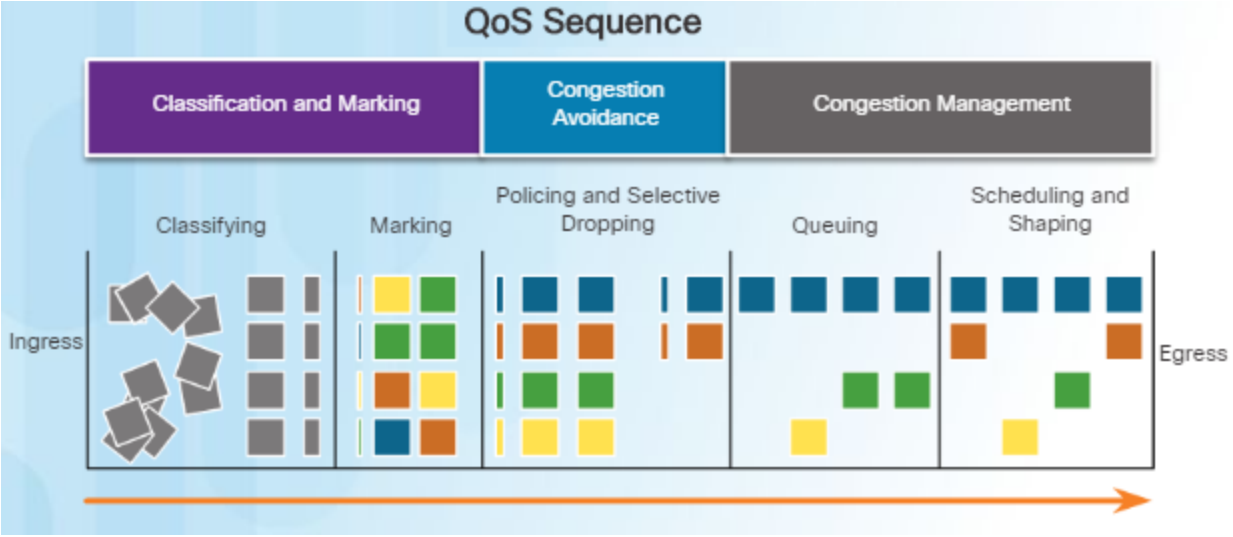
Classification and Marking
- Marking at Layer 2
- Marking at Layer 3

Congestion Avoidance
Shaping and Policing

What is QOS

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.



Bandwidth, Congestion, Delay and Jitter

- 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.

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.

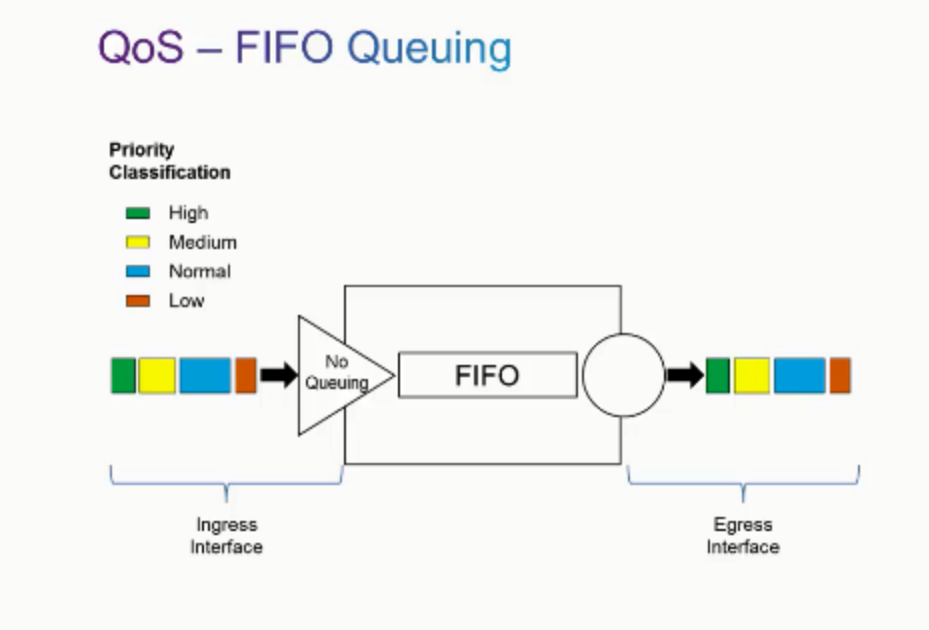
Queuing Algorithms

FIFO (First In First Out)

FIFO queuing, also known as first-come, first-served queuing, involves buffering and forwarding of packets in the order of arrival.

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.

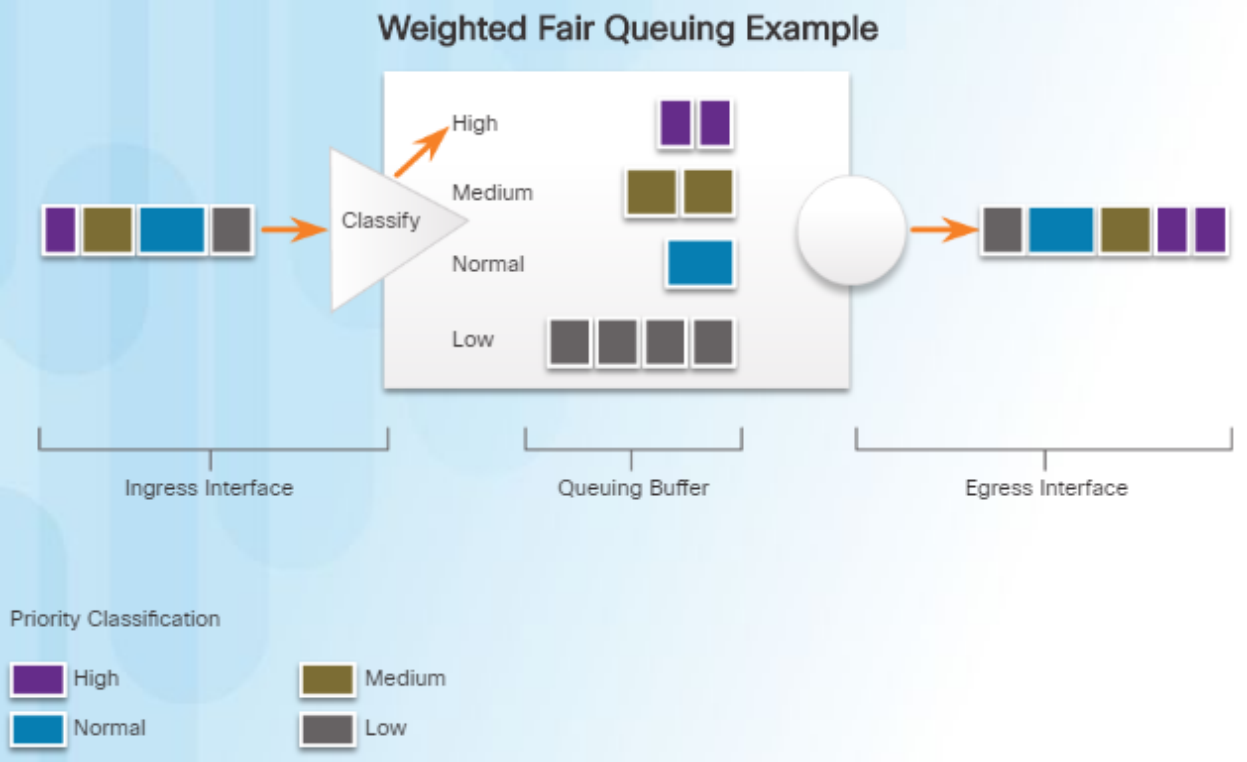


WFQ (Weighted Fair Queuing)

WFQ applies priority, or weights, to identified traffic and classifies it into conversations or 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.

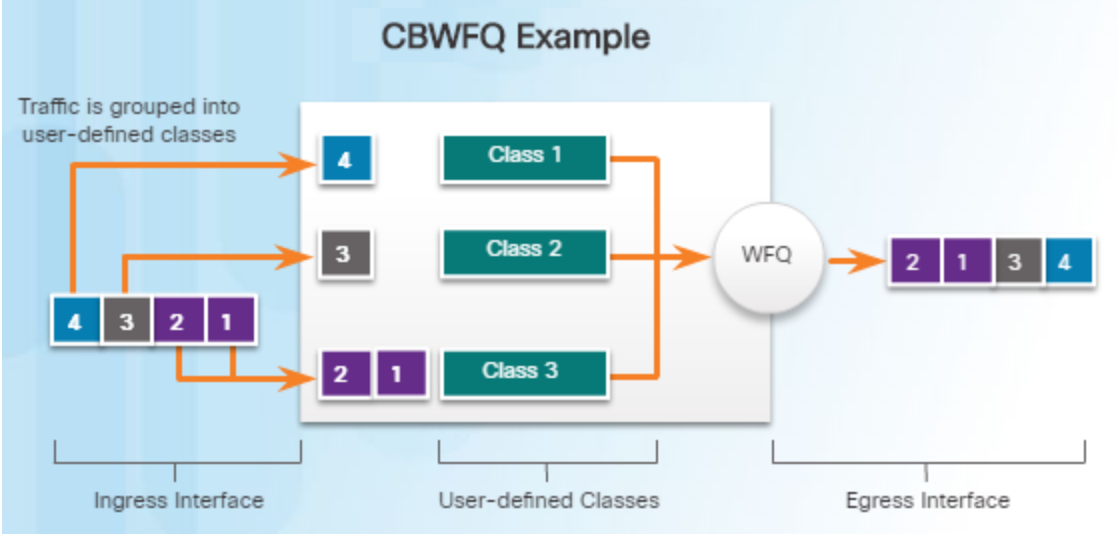
WFQ classifies traffic into different flows based on packet header addressing, including source/destination IP addresses, MAC addresses, port numbers, protocols, and type of service (ToS) values.



CBWFQ (Class-Based Weighted Fair Queuing)

CBWFQ extends the standard WFQ functionality to provide support for user-defined traffic classes.

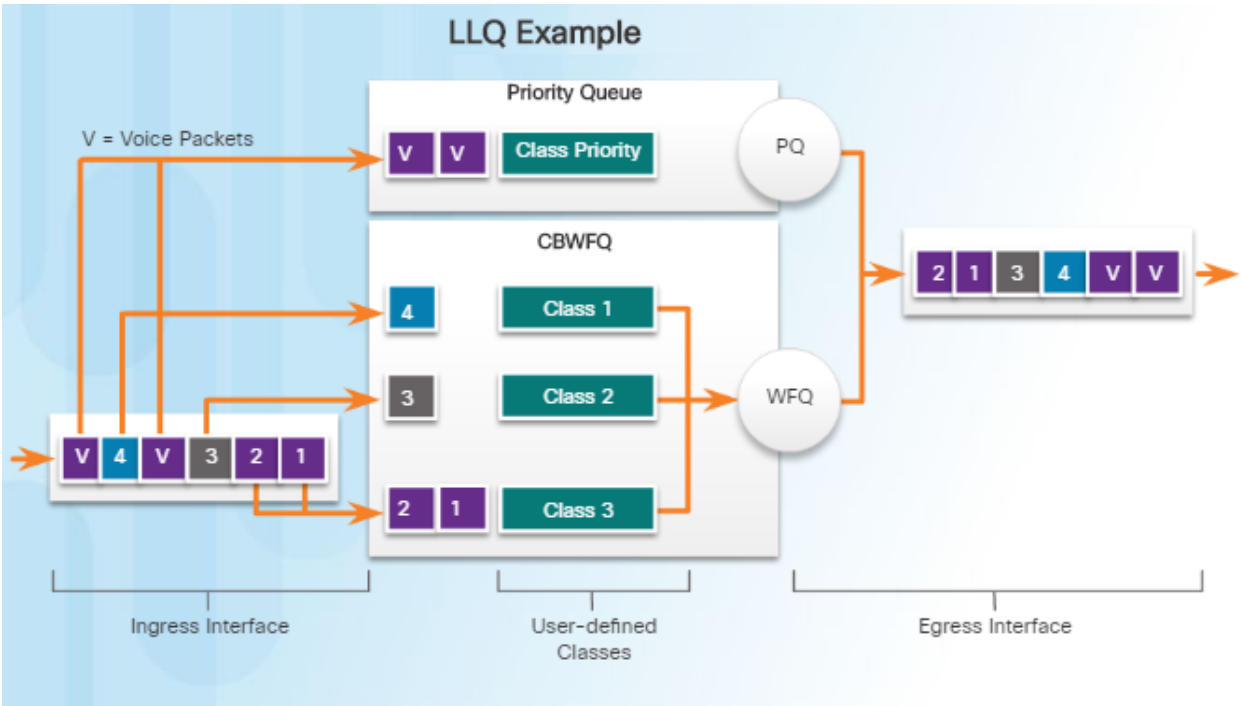
You define traffic classes based on match criteria including protocols, ACLs, and input interfaces



LLQ (Low Latency Queuing)

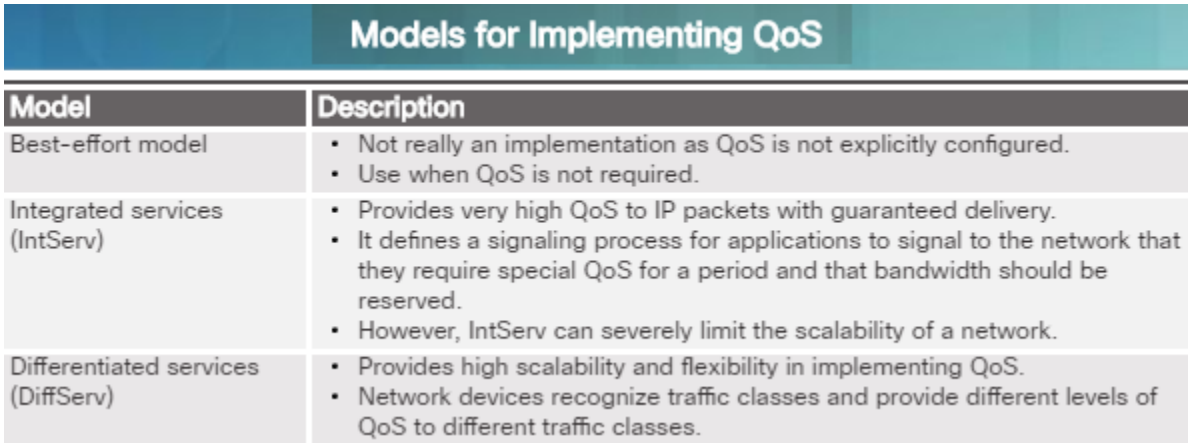
The LLQ feature brings strict priority queuing (PQ) to CBWFQ which reduces jitter in voice conversations.

With LLQ, delay-sensitive data is sent first, before packets in other queues are treated.



Implementing QOS

How can QoS be implemented in a network? The three models for implementing QoS are these:



Models for Implementing QoS

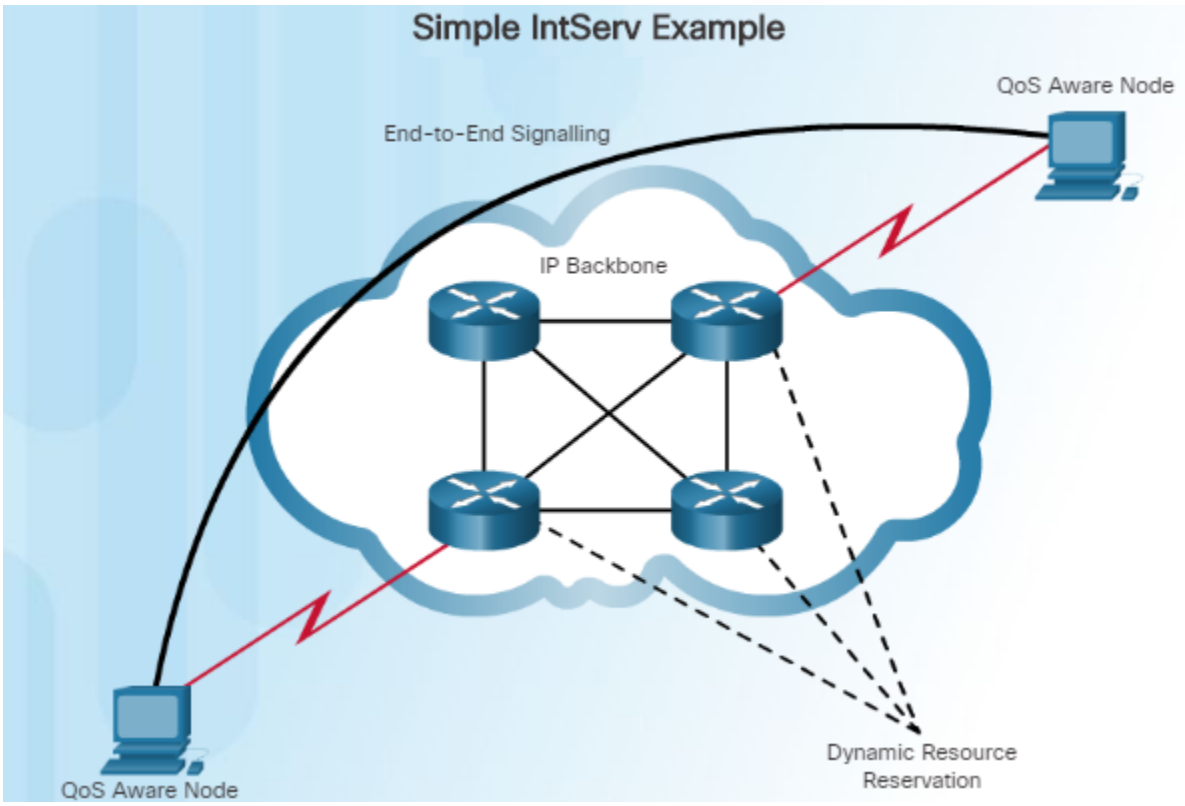
Model	Description
Best-effort model	• Not really an implementation as QoS is not explicitly configured. • Use when QoS is not required.
Integrated services (IntServ)	• Provides very high QoS to IP packets with guaranteed delivery. • It defines a signaling process for applications to signal to the network 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

The basic design of the Internet, which is still applicable today, provides for best-effort packet delivery and provides no guarantees.
The best-effort model treats all network packets the same way.
Without QoS, the network cannot tell the difference between packets. A voice call will be treated the same as an email with a digital photograph attached.

Integrated Services

IntServ provides a way to deliver end-to-end Qos that real-time applications require by explicitly managing network resources to provide QoS to specific user packet streams.

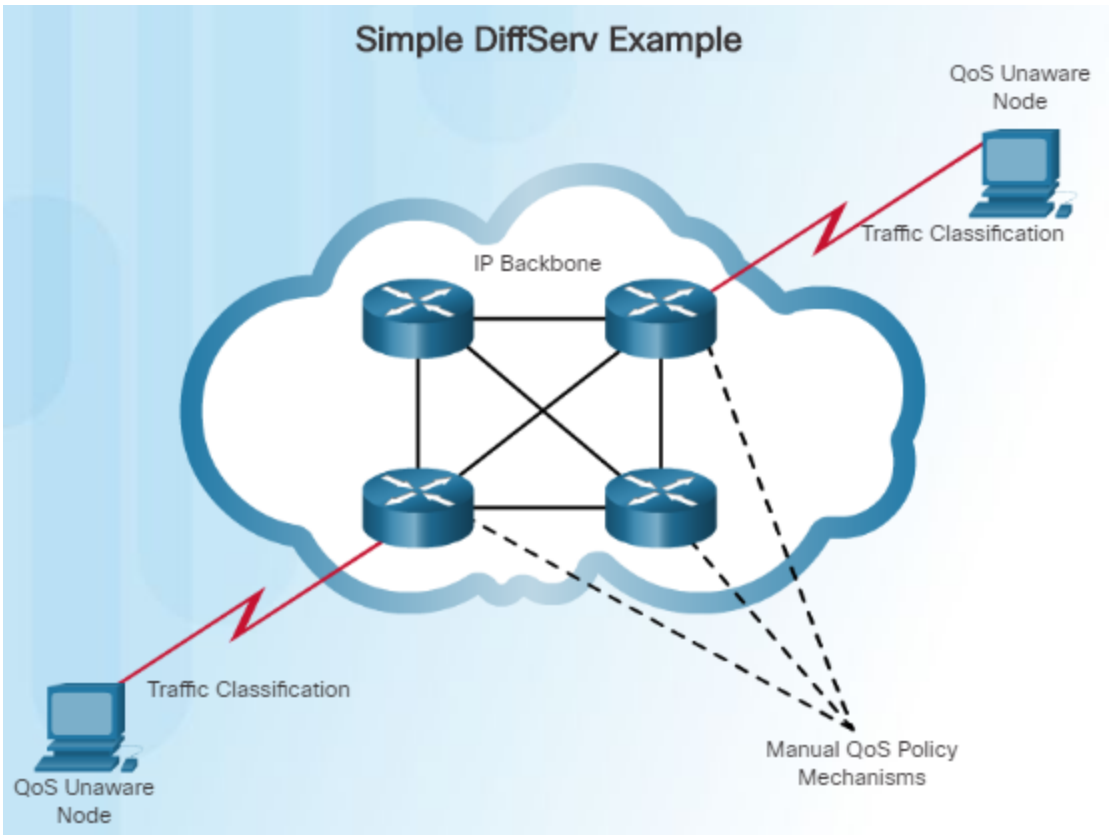


In the IntServ model, the application requests a specific kind of service from the network before sending the data.
The application informs the network of its traffic profile and requests a particular kind of service that can encompass its bandwidth and delay requirements.

Differentiated Services

The differentiated services (DiffServ) QoS model:

- Specifies a simple and scalable mechanism for classifying and managing network traffic.
- Provides QoS guarantees on modern IP networks.
- DiffServ can provide low-latency guaranteed service to critical network traffic such as voice or video.



The DiffServ design overcomes the limitations of both the best-effort and IntServ models.
DiffServ can provide an “almost guaranteed” QoS while still being cost-effective and scalable.

DiffServ is not an end-to-end QoS strategy because it cannot enforce end-to-end guarantees. However, it is a more scalable approach to implementing QoS.

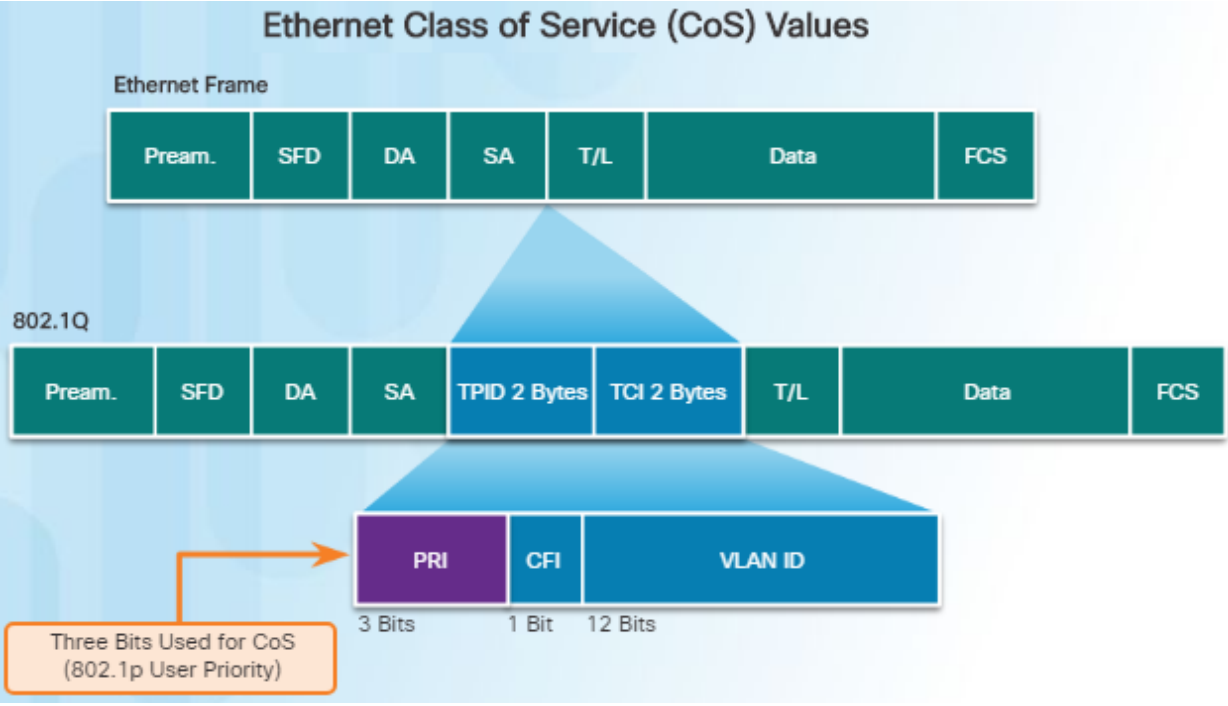
Classification and Marking

Traffic Marking for QoS

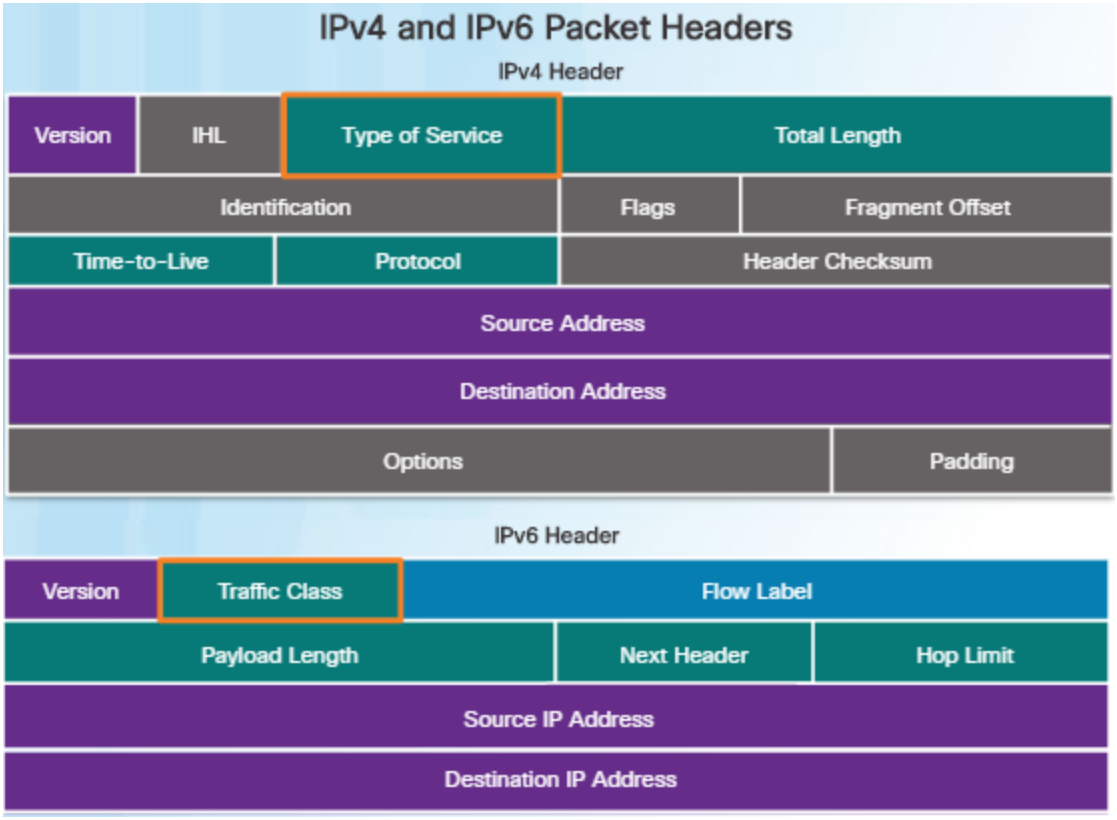
QoS Tools	Layer	Marking Field	Width in Bits
Ethernet (802.1Q, 802.1p)	2	Class of Service (CoS)	3
802.11 (Wi-Fi)	2	Wi-Fi Traffic Identifier (TID)	3
MPLS	2	Experimental (EXP)	3
IPv4 and IPv6	3	IP Precedence (IPP)	3
IPv4 and IPv6	3	Differentiated Services Code Point (DSCP)	6

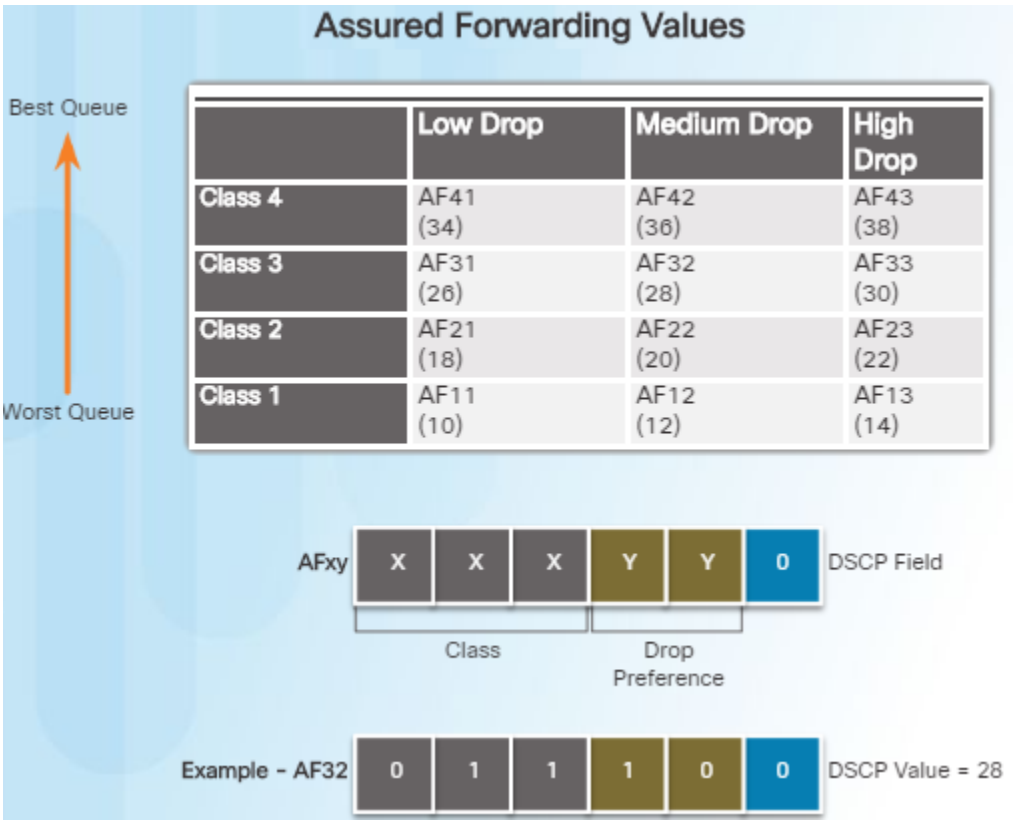
Classification determines the class of traffic to which packets or frames belong. Policies can not be applied unless the traffic is marked.
Methods of classifying traffic flows at Layer 2 and 3 include using interfaces, ACLs, and class maps.
Marking requires the addition of a value to the packet header and devices that receive the packet look at this field to see if it matches a defined policy.

Marking at Layer 2



Marking at Layer 3



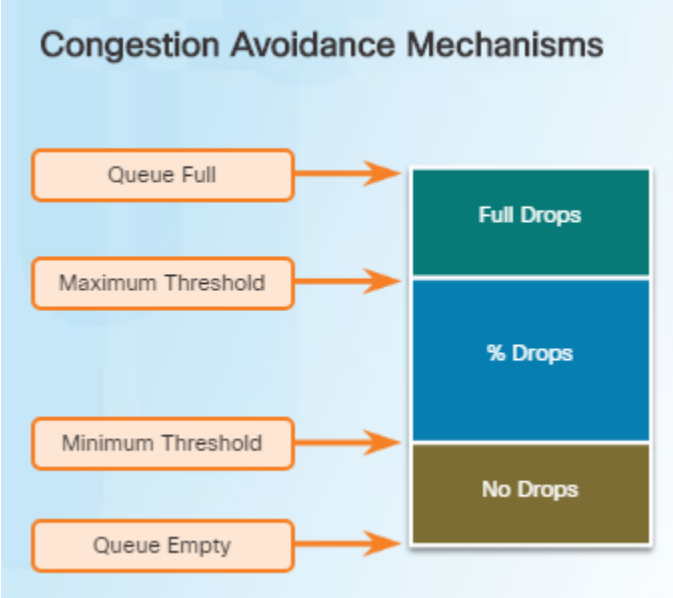


The 64 DSCP values are organized into three categories:

- Best-Effort (BE) – Default for all IP packets. The DSCP value is 0.
- Expedited Forwarding (EF) – The DSCP value is 46. At layer 3, Cisco recommends that EF should only be used to mark voice packets.
- Assured Forwarding (AF) – Uses the 5 most significant DSCP bits to indicate queues and drop preference. As shown in the figure, the first 3 most significant bits are used to designate the class.

Class 4 is the best queue and Class 1 is the worst queue.
The 4th and 5th most significant bits are used to designate the drop preference.
The 6th most significant bit is set to zero.
The AFxy formula shows how the AF values are calculated.

Congestion Avoidance



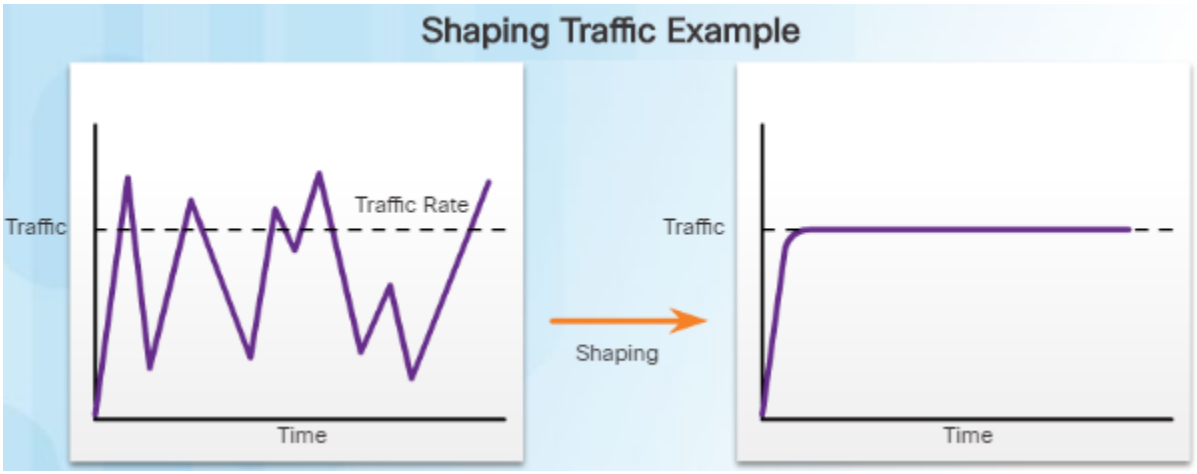
- Congestion avoidance tools monitor network traffic loads in an effort to anticipate and avoid congestion at common network bottlenecks before congestion becomes a problem.
- Congestion avoidance is achieved through packet dropping.
- These tools monitor the average depth of the queue.
    - For example,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

Shaping and Policing

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.