**Quality of Service or QOS** allows for certain types of traffic to prioritized over other types. This is to help alleviate bandwidth limitations and combat latency and jitter. The ITY recommends that network latency in general should not exceed 400 ms and any real-time applications not exceed 150 ms. Though both the ITU and Cisco have demonstrated that real-time applications quality does not start to degrade until 200 ms.

# Policing and Shapers

Traffic policers and shapers are traffic-conditioning QoS mechanisms used to classify traffic and enforce other QoS mechanisms such as rate limiting. They classify traffic in identical manners but differ in their implementation.

- **Policers** – Drop or re-mark incoming or outgoing traffic that goes beyond a desired traffic rate.
- **Shapers** – Buffer and delay egress traffic rates that momentarily peak above the desired rate until egress rate drops below the defined traffic rate. If the egress traffic rate is below the desired rate, the traffic is sent immediately.

# Serialization Delay

Serialization delay is the time it takes to put every bit of a packet onto a link. This is a fixed value that depends on the speed of a link (100mb, 1gb, 10gb, etc...). For example the serialization delay of a packet on a 1 gigabit link is 12 $\mu$s. The equation is below:

*s = packet size in bits ÷ line speed in bits per second*

# Processing Delay

Processing delay is the time it takes for the device to take a packet from the input interface and into the output queue of the output interface. This is a fixed value but depends on certain hardware and software configurations such as these:

- CPU Speed
- CPU Utilization
- IP packet switching mode
- Router architecture
- Input and output interface configurations

# QoS Models

- Best Effort: nothing gets any special treatment.
- Integrated Services (IntServ): An application signals the network to make a bandwidth reservation and indicates that it required QoS treatment.
- Differentiated Services (DiffServ): The network identifies classes that require special QoS treatment.

---

## IntServ

This model is intended for real-time applications such as voice and video. The applications signal their requirements to the network to reserve end-to-end resources such as bandwidth that would provide an acceptable experience. Resource Reservation Protocol (RSVP) is used to signal the network and communication requirements. In addition it provides call admission control (CAC) which guarantees that nothing else can reserve that bandwidth.

For IntServ to function every node in the network including the endpoints need to both support IntServ and be configured. This is the big drawback of IntServ as this solution does not scale well.

The device that needs to send data sends an RSVP PATH message to the reliever which communicates the destination and source IP address, port, protocol, previous hop, and bandwidth reservation. Then the receiver sends an RSVP RESV message in the reverse path as before with the same information and bandwidth is reserved at each link. If the receiver at the other end needs to also send data then it must also follow the same RSVP PATH and RSVP RESV process to reserve bandwidth in the opposite direction.

---

## DiffServ

DiffServ seek to address the limitations of both best effort and IntServ models. Without a signaling protocol and RSVP flow state to maintain at each link makes DiffServ highly scalable as it manages QoS characteristics on a hop by hop basis independently of each device on the network. DiffServ does not provide end-to-end QoS. DiffServ divides IP traffic into classes and marks traffic to apply a different level of service to each class.

First IP traffic must be identified and categorized and into different classes. After it is classified it can be marked this allows other devices to apply QoS mechanisms to those packets as they go through the network. There are marking options for both L2 frames and L3 packets.

# Classification

Classification is responsible for identifying different traffic streams. Traffic descriptors are used to categorize packets within a specific class. Classification should take place at the network edge as class to the source of the traffic as possible. Once a packet is marked it can be marked, remarked, queued, policed, or shaped. These are some common descriptors used for classification:

- Internal: QoS groups (locally significant)
- Layer 1: Physical interface, sub interface, or port
- Layer 2: MAC address and 802.11 Q/p CoS bits
- Layer 2.5: MPLS Experimental (EXP) bits
- Layer 3: Differentiated Services Code Points (DSCP), IP Precedence (IPP), and source/destination IP address
- Layer 4: TCP and UDP ports
- Layer 7: Next Generation Network-Based Applications Recognition (NBAR2)

The most commonly used of these in an enterprise environment are Layer 2, Layer 3, Layer 4, and Layer 7.

# Layer 7 Classification

NBAR2 is a DPI engine that can classify and identify a wide variety of protocols and applications from layer 3 all the way up to layer 7 even applications that use dynamic TCP/UDP port numbers.
NBAR2 has two modes of operation:

- Protocol Discovery: Allows NBAR2 to discover and get real-time stats on applications currently running on the network. These stats can be used to define QoS classes and policies using MQC
- Modular QoS CLI (MQC): Traffic matching different protocols such as Webex and YouTube can be placed into different traffic classes and allowing policies to be applied.

# DSCP

| Application | DSCP | Decimal Value |
|---|---|---|
| Network Control | CS7 | 56 |
| Internetwork Control | CS6 | 48 |
| VoIP | EF | 46 |
| Broadcast Video | CS5 | 40 |
| Multimedia Conferencing | AF4 | 34-38 |
| Real-Time Interaction | CS4 | 32 |
| Multimedia Streaming | AF3 | 26-30 |
| Signaling | CS3 | 24 |
| Transactional Data | AF2 | 18-22 |
| Network Management | CS2 | 16 |
| Bulk Data | AF1 | 10-14 |
| Scavenger | CS1 | 8 |
| Best-Effort | Default | 0 |

# Marking

Packet marking is a QoS process that marks or colors a packet by changing a field in the header of a frame or packet with a traffic descriptor allowing it to be distinguished from other packets during the application of other QoS processes.
These are the descriptors used for marking traffic:

- Internal: QoS groups
- Layer 2: 802.11 Q/p CoS bits
- Layer 2.5: MPLS Experimental (EXP) bits
- Layer 3: Differentiated Services Code Points (DSCP) and IP Precedence (IPP)

The IEEE spec for 802.1Q has two 2-bit fields one of which is the Tag Protocol Identifier (TPID) field, and the Tag Control Information (TCI) field. The TPID field is 16-bits and has a value of 0x8100 which identifies it as an 802.1Q tagged frame. The TCI field is also 16-bits and is comprised of three fields. The Priority Code Point (PCP) 3-bits. The Drop Eligible Indicator (DEI) 1 bit. And the VLAN ID which is the other 12 bits.

The PCP field is defined in the IEEE 802.1p spec. This field is used to mark packets that belong to a specific CoS. This allows a L2 frame to be marked 8 ways, 0 - 7. O being the lowest and 7 being the highest.

# PCP Chart

| PCP Value/Priority | Acronym | Traffic Type |
|---|---|---|
| 0 (Lowest) | BK | Background |
| 1 (Default) | BE | Best Effort |
| 2 | EE | Excellent Effort |
| 3 | CA | Critical Applications |
| 4 | VI | Video with < 100ms of latency and jitter |
| 5 | VO | Video with < 10ms of latency and jitter |
| 6 | IC | Internetwork Control |
| 7 (Highest) | NC | Network Control |

## Configuration

Here is some Cisco documentation with some example QoS configurations.

https://www.cisco.com/c/en/us/td/docs/ios-xml/ios/qos_dfsrv/configuration/15-mt/qos-dfsrv-15-mt-book/qos-dfsrv.html#GUID-9F267640-3F4C-48D1-A57D-0D9B552A5DD1

*Router(config)# class-map match-any EMAIL*
*Router(config-cmap)# match protocol pop3*
*Router(config-cmap)# match protocol smtp*
*Router(config-cmap)# match protocol imap*
*Router(config-cmap)# match protocol exchange*
*Router(config-cmap)# exit*
*Router(config)# class-map VOICE*
*Router(config-cmap)# match protocol rtp audio*
*Router(config-cmap)# exit*
*Router(config)# class-map OTHER*
*Router(config-cmap)# match protocol bittorrent*

Policy maps are where QoS policies are actually applied and policing, shaping, etc... are configured.

*Router(config)# policy-map DEMO*
*Router(config-pmap)# class EMAIL*

This guarantees at least 512 Kbps of bandwidth.
*Router(config-pmap-c)# bandwidth 512*
*Router(config-pmap-c)# random-detect dscp-based*
*Router(config-pmap-c)# exit*
*Router(config-pmap)# class VOICE*

This allows upto 256 Kbps and no more.
*Router(config-pmap-c)# priority 256*
*Router(config-pmap-c)# exit*
*Router(config-pmap)# class OTHER*

Any traffic higher than 128 Kbps will be dropped (this one is input in Bps).
*Router(config-pmap-c)# police 128000*
*Router(config-pmap-c-police)# exit*
*Router(config-pmap-c)# exit*
*Router(config-pmap)# class class-default*

Enabled weighted fair queuing on the default class-map.
*Router(config-pmap-c)# fair-queue*
*!*
*Router(config)# interface gig 0/2*
*Router(config-if)# service-policy output DEMO*
*!*
*Router# show policy-map interface gig 0/2*