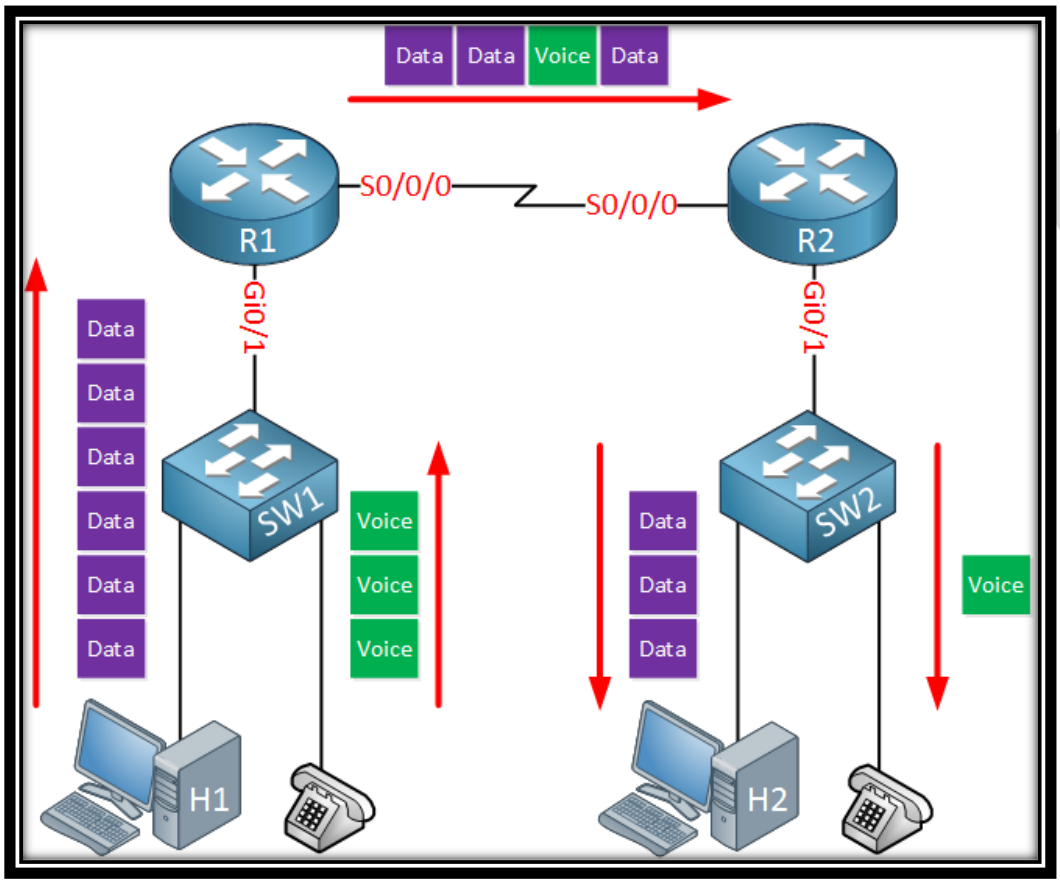
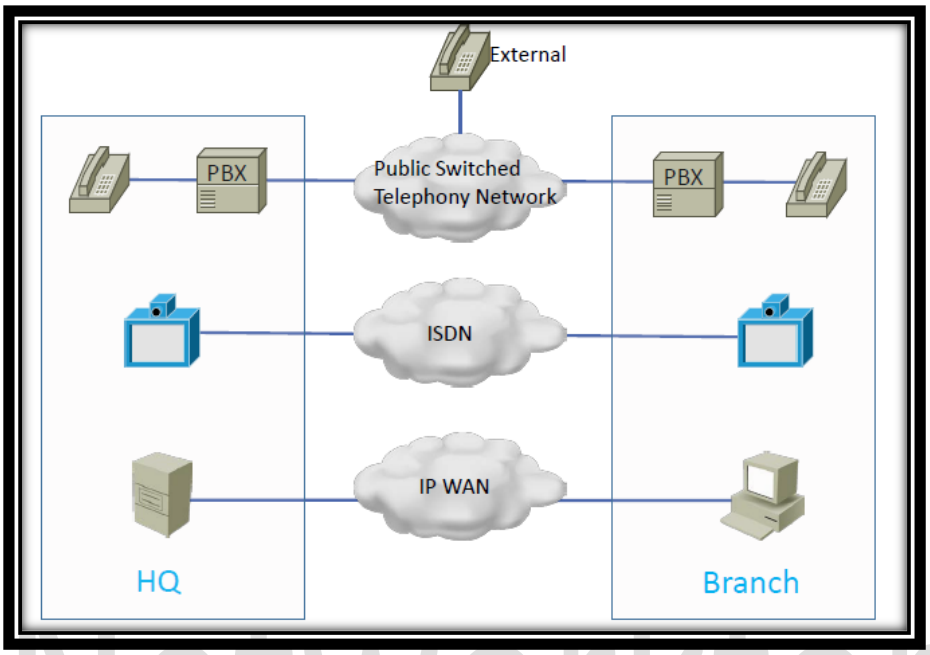
QOS (Quality of Service):

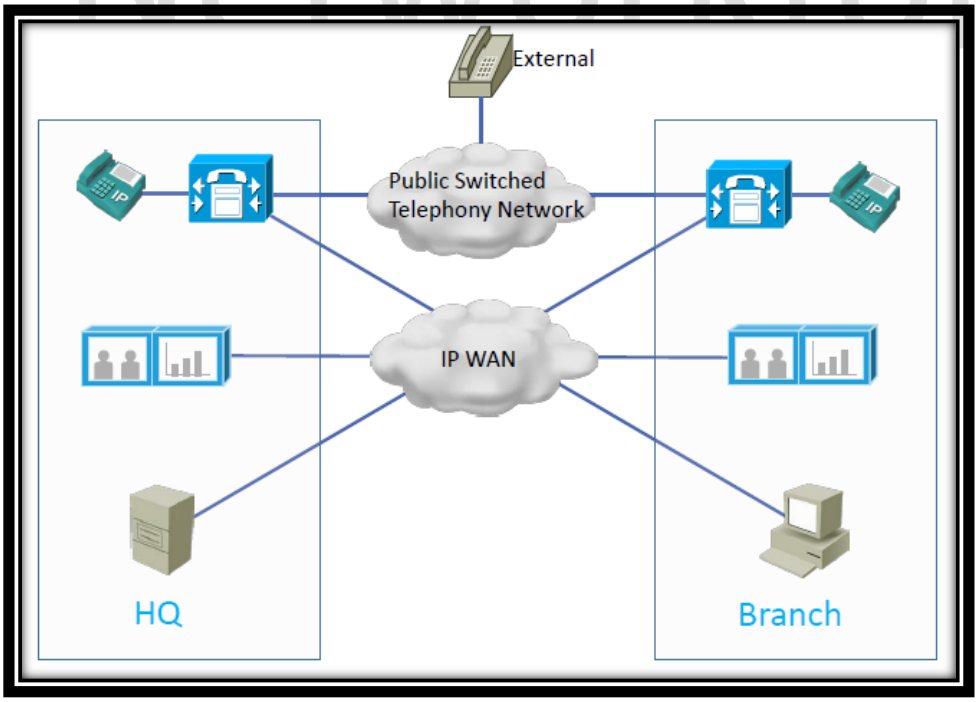
- Quality of Service (QOS) is a feature of routers and switches which prioritizes traffic so that more important traffic can pass first.
- Quality of Service allows us to queue, limit, Filter and make reservations for our network traffic.
- Network devices don't really care about the type of traffic they have to forward.
- Example our switch receives an Ethernet frame, looks for the destination MAC address and forwards the frame towards the destination.
- And for Router also applies same thing like our router receives an IP packet, looks for the destination in the routing table and it forwards packet towards the destination.
- QoS is about using tools to change how the router or switch deals with the different packets.
- Quality of service (QoS) refers to any technology that manages data traffic to reduce packet loss, latency and jitter on the network.
- QoS controls and manages network resources by setting priorities for specific types of data on the network.



Dedicated Voice, Video and Data Networks:



Converged Networks:



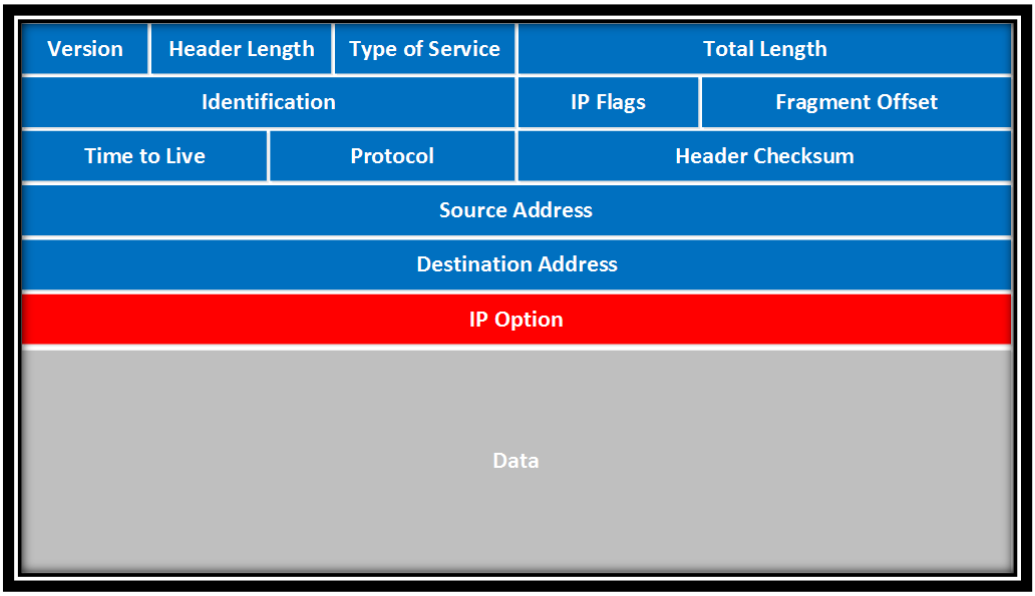
- On old traditional network , data, voice and video had their own separate network infrastructure and did not impact each other
- On modern networks, data, voice and video run over the same shared infrastructure.
- This enables cost savings and advanced features for voice and video.
- Data, voice and video are all fighting for all same shared bandwidth.
- FIFO (First in First Out) whenever congestion is experienced on router or switch, packets are sent out in a First in First out (FIFO) manner by default.
- Congestion can be experienced wherever it is possible for packets to come in quicker than they can be sent out.
- Example Traffic is going left to right from Head office to Branch office. The WAN edge router has a Fast Ethernet Interface on the inside LAN Interface and an E1 Interface on the outside WAN Interface.
- Packets are arriving faster than they can be sent out.
- Packets wait in the queue to go out. Packets are sent out FIFO in the order they were received.
- Congestion causes delay to packets as they wait in the queue.
- As the size of the queue changes it causes Jitter and there is a limit to the size of the queue. If packet arrives when the queue is full the router will drop it.
- Voice and Video calls and other applications will be unacceptable quality if they do not meet their delay, jitter and loss requirements.

How to Mitigate Congestion or How to Solve Congestion Problem?

- Add more bandwidth – this costs money.
- Use Quality of Service Techniques to give better service to the traffic which needs it.

Effects of Qos Queuing:

- QOS queuing can reduce latency, jitter and loss for particular traffic.
- The original driver for QoS was voice over IP but it can also be used to give better service to data applications.
- If we are giving one type of traffic better service on the same link we started with, the other traffic type must get worse service.
- The point is to give each type of traffic the service it requires.
- QoS queuing is designed to mitigate temporary periods of congestion. If a link is permanently congested the bandwidth should be increased.



Qos Terminologies:

Bandwidth:

- Network bandwidth refers to the speed of a link or interface, in bits per seconds.
- Think bandwidth as speed that mean capacity of the link
- Network bandwidth is measured in number of bits transmitted into bits per second

Congestion:

- Network congestions causes the delay or Latency
- Network congestions occur when interface experiences more traffic that it can handle

Delay:

- Delay or latency refers to time it takes for packet to travel from source to destination.
- In Quality of service there are two types of delays one is fixed delay and other is variable.

Jitter (Variation in delay):

- In QOS Jitter is defined as a variation in the delay of received packets. Or Difference between the delays of the IP Packets or we can say jitter is the irregular time delay in the sending of the data packet over network.

Latency (Delay):

- Latency is the time required by a data packet to reach the destination from the source

Packet Loss:

- We have loss of packet when we have congestion and don't do QOS mechanisms or we can say when congestion occurs network devices such as router or switches can drop packets if we don't do Qos.

Voice:

- In the Network voice is very sensitive to delay and dropped packets. It make no sense and not possible to retransmit or re send the voice if packets are lost.
- So we need to take care for voice packets must receive a higher priority then other types of network traffic.

Videos:

- Videos also important without Qos significant amount of extra bandwidth capacity video quality degrades.

Data:

- Data applications that have no tolerance for data loss, such as email and web pages. And it use TCP to ensure that if the data packets are lost in the transit they will be resent.

Classification and Marking:

- For a router or switch to give a particular level of service to a type of traffic it has to recognize that traffic first.
- Common way to recognize the traffic are by Class of Service (COS) for layer 2 marking and Differentiated service code point (DSCP) marking for L3 and Access Control List or Network Based Application Recognition (NBAR).

Layer 2 Marking - Class of Service:

- There is a 3 bit field in the layer 2 802.1q frame header which is used to carry the Cos Qos Marking.
- A value of 0 - 7 can be set. The default values are 0 which is designated as Best Effort traffic.
- Cos 6 and 7 are reserved for network use -- Like routing protocol and network control traffic.
- **Higher the number more important the traffic**
- IP Phones mark their **call signaling traffic as cos 3** and their voice payload (Spoking voice) **as cos 5**.
- **Spoking voice is most important then calls signaling.**

Layer 3 Marking - DSCP (Differentiated Service Code Point):

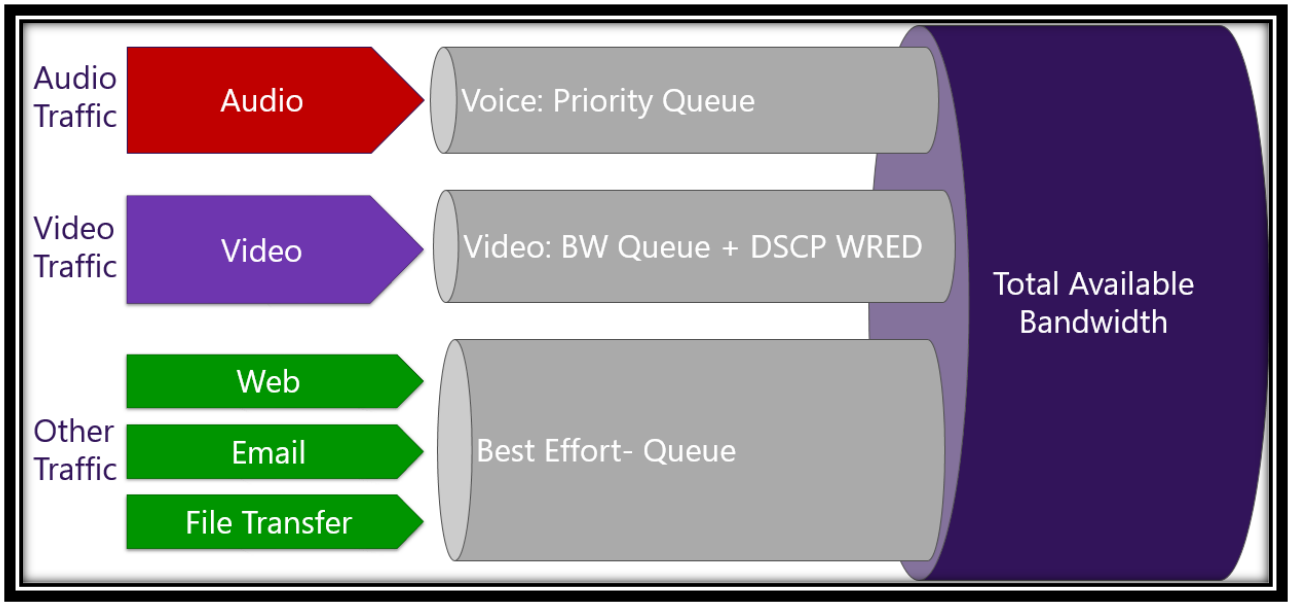
- DSCP Qos marking carry in Layer 3 IP Header.
- In IP Header DSCP Qos marking 6 bits are used which gives 64 possible values. The default value is 0 which is designated as Best effort traffic here we can mark DSCP zero that mean that traffic not gets any special services.
- IP phones mark their call signaling traffic as 24 (CS3) (**That is DSCP 24 is also known as CS3**) and their voice payload as 45 (EF) (DSCP value 45 is also known as EF)
- There are some standard markings for other traffic types such as 26 (AF31) for mission critical data and 34 (AF41) for SD video
- DSCP is the preferred classification and marking method because the router can very quickly gather the information from a single byte in the IP header.

Recognizing Traffic with an ACL:

- An Access control list can be used to recognize traffic based on its layer 3 and layer 4 information.
- For example SSH traffic going to and from the router on TCP port number 22. Need to give good bandwidth etc.
- So we can configure ACL on router with referring port # etc.

Recognizing Traffic with NBAR (Network Based Application Recognition):

- NBR can be used to recognize traffic based on its layer 3 to layer 7 information.
- We can downloaded signature from Cisco and loaded on your router which recognizes well known applications.



Congestion Management:

- So example in router or switch we are getting more traffic to send out with the outer interface bandwidth then we do queuing. Or in other words we can say Queuing can be used to manage congestion on routers and switches.

There are two type of Queuing policy commonly use

- CBWFQ (Class Based Weighted Fair Queuing) --- gives bandwidth guarantees to specified traffic types.
- LLQ (Low Latency Queuing) -- Priority queuing (Traffic into the priority queue is send before other traffic.

MQC (Modular Qos CLI):

- Cisco Qos Configuration uses the MQC modular Qos CLI
- It has 3 main sections
- **Class Maps** define the traffic to take an action on
- **Policy Maps** take the action on that traffic
- **Service Policies** apply the policy to an interface

Shaping and Policing:

- Traffic Shaping and Policing can be used to control traffic rate
- They both measure the rate of traffic through an interface an take an action if the rate is above a configured limited
- Traffic shaping buffers any excess traffic so the overall traffic stays within the desired rate limit.

- Traffic policing drops or remarks excess traffic to enforce the specified rate limit.