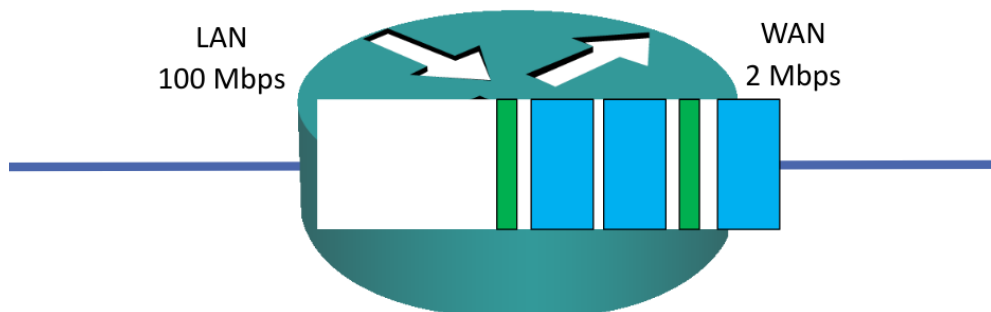


Quality Requirements for Voice and Video

- Voice and traditional standard definition video packets must meet these recommended requirements to be an acceptable quality call:
 - Latency (delay) ≤ 150 ms
 - Jitter (variation in delay) ≤ 30 ms
 - Loss $\leq 1\%$
- These are one way requirements, meaning a packet sent from a phone in HQ has 150ms to reach the phone in the branch, and vice versa
- HD video has stricter requirements

Congestion

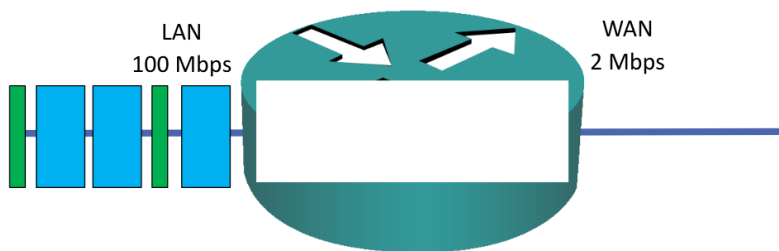
- Packets are arriving faster than they can be sent out
- Packets wait in the queue to go out
- Packets are sent out FIFO in the order they were received



- Congestion causes delay to packets as they wait in the queue
- As the size of the queue changes it causes jitter
- There is a limit to the size of the queue. If a packet arrives when the queue is full the router will drop it
- Voice and video calls (and applications) will be unacceptable quality if they do not meet their delay, jitter and loss requirements

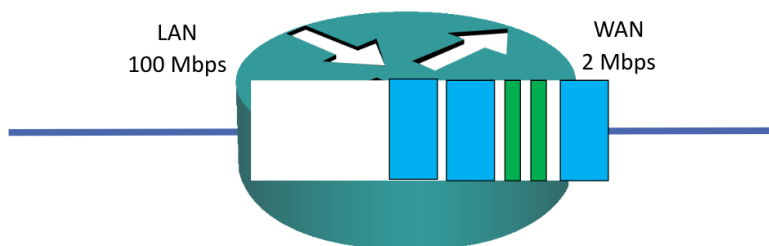
Congestion Example with QoS Queuing

- Traffic destined for the branch comes in on the LAN interface at a rate higher than 2 Mbps



Congestion Example

- Packets are arriving faster than they can be sent out
- Packets wait in the queue to go out
- The router recognises the voice packets and moves them to the front of the queue to minimise their delay



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 you're giving one type of traffic better service on the same link you started with, the other traffic types must get worse service
- The point is to give each type of traffic the service it requires
- QoS queuing is not a magic bullet and is designed to mitigate temporary periods of congestion. If a link is permanently congested the bandwidth should be increased

Classification and Marking

- For a router or switch to give a particular level of service to a type of traffic, it has to recognise that traffic first
- Common ways to recognise the traffic are by COS (Class of Service) marking, DSCP (Differentiated Service Code Point) marking, an Access Control List, or NBAR (Network Based Application Recognition)

Layer 2 Marking - CoS Class of Service

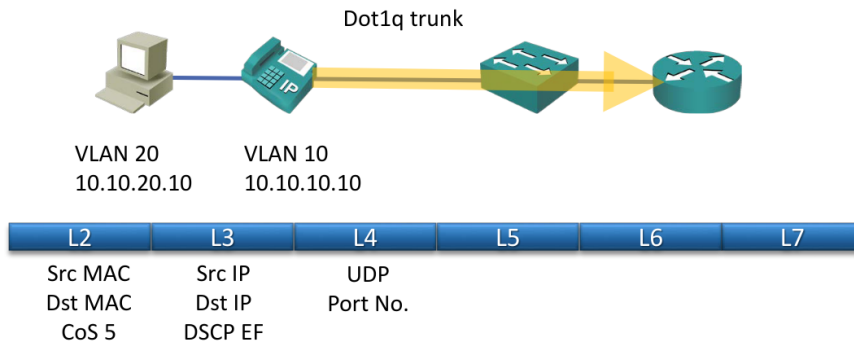
- There is a 3 bit field in the Layer 2 802.1q frame header which is used to carry the CoS QoS marking
- A value of 0 – 7 can be set. The default value is 0 which is designated as Best Effort traffic
- CoS 6 and 7 are reserved for network use
- IP phones mark their call signalling traffic as CoS 3 and their voice payload as CoS 5

Layer 3 Marking - DSCP

- The ToS Type of Service byte in the Layer 3 IP header is used to carry the DSCP QoS marking
- 6 bits are used which gives 64 possible values. The default value is 0 which is designated as Best Effort traffic
- IP phones mark their call signalling traffic as 24 (CS3) and their voice payload as 46 (EF)
- There are standard markings for other traffic types, such as 26 (AF31) for mission critical data, and 34 (AF41) for SD video

The Trust Boundary

- The switch should be configured to trust markings from the IP phone and pass them on unchanged, but mark traffic from the PC down to CoS 0 and DSCP 0



Recognising Traffic with an ACL

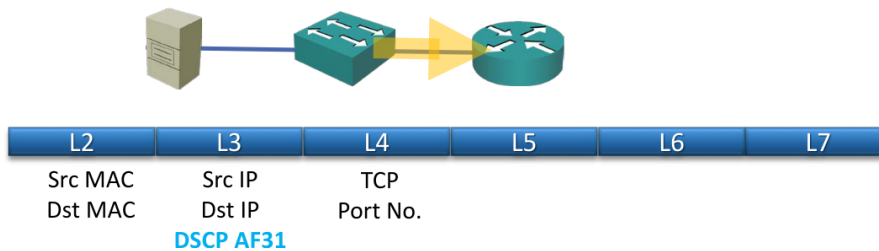
- An Access Control List can be used to recognise traffic based on its Layer 3 and Layer 4 information
- For example SSH traffic going to and from the router 10.10.100.10 on TCP port number 22

Recognising Traffic with NBAR

- NBAR (Network Based Application Recognition) can be used to recognise traffic based on its Layer 3 to Layer 7 information
- Signatures can be downloaded from Cisco and loaded on your router which recognise well known applications

Classification and marking

- 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
- If using another method such as an ACL or NBAR is being used, this should be done as close to the source as possible and then a DSCP value added



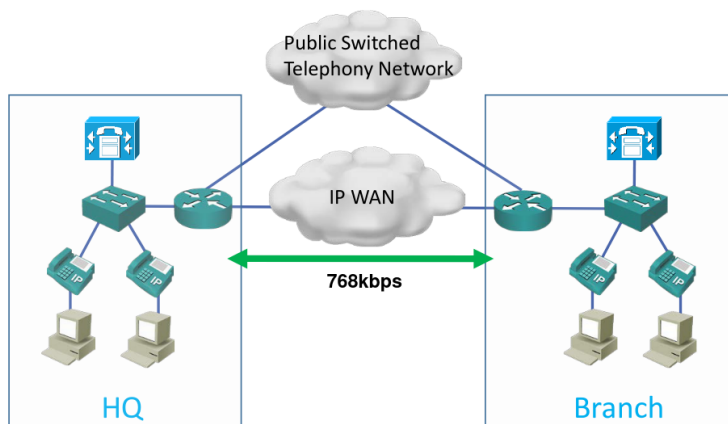
Congestion Management

- Queuing can be used to manage congestion on routers and switches
- CBWFQ (Class Based Weighted Fair Queuing) gives bandwidth guarantees to specified traffic types
- LLQ (Low Latency Queuing) is CBWFQ with a priority queue
- Traffic in the priority queue is sent 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

Congestion Management Example



- 768kbps WAN Link between offices
- Need to support 10 concurrent voice calls over the WAN
- Each call = 25.6kbps
- 256kbps provisioned for voice calls
- 512kbps provisioned for data
- Data will sometimes burst above 512kbps creating congestion

(This config isn't required on the CCNA)

Congestion Management Example - LLQ

- Configure the same LLQ policy on the routers in HQ and the branch
- Apply to the WAN interfaces

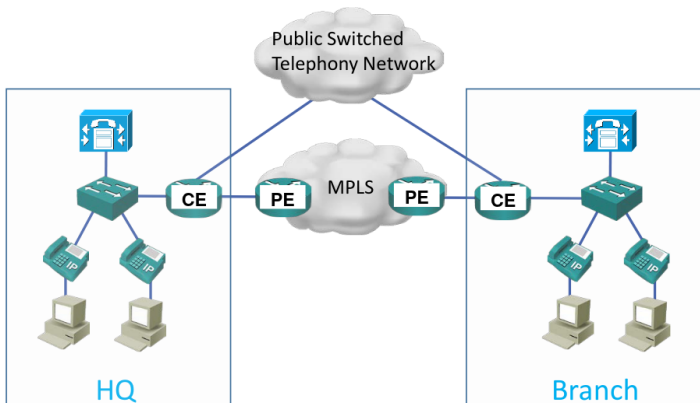
```
class-map VOICE-PAYLOAD
match ip dscp ef
class-map CALL-SIGNALING
match ip dscp cs3
!
policy-map WAN-EDGE
class VOICE-PAYLOAD
priority percent 33
class CALL-SIGNALING
bandwidth percent 5
class class-default
fair-queue
!
interface Serial0/0/0
bandwidth 768
service-policy out WAN-EDGE
```

Shaping and Policing

- Traffic Shaping and Policing can be used to control traffic rate.
- They both measure the rate of traffic through an interface and take an action if the rate is above a configured limit.

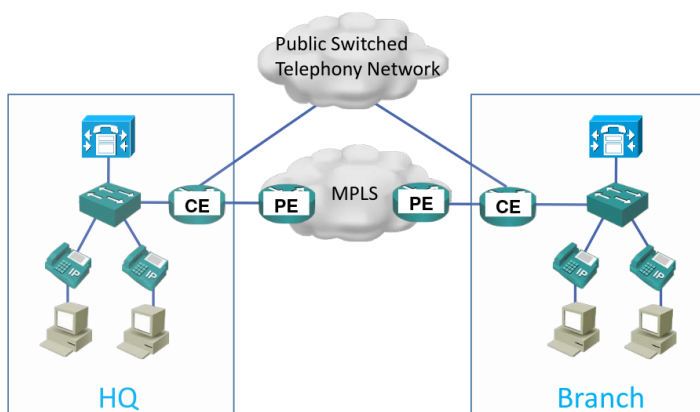
- Traffic shaping buffers any excess traffic so the overall traffic stays within the desired rate limit.
- Traffic policing drops or re-marks excess traffic to enforce the specified rate limit.
- Classification can be used to allow different rates for different traffic types.

Policing Scenario – Service Provider PE



- In this example the customer has provisioned an MPLS VPN with a service provider
- The physical links from the CE to PE routers are 100 Mbps, but the customer has paid for 10 Mbps in their SLA
- The service provider enables policing inbound on the PE routers to limit the customer to 10 Mbps bandwidth. Excess traffic is dropped

Shaping Scenario – Customer CE



- The CE to PE link is 100 Mbps. If the customer sends at a rate above 50 Mbps, excess traffic will be dropped by the provider
- Some traffic would reach the destination, some would not
- The dropped traffic would be random, it could be data or voice
- When voice packets are dropped call quality is unacceptable
- The customer configures shaping outbound on the CE WAN interfaces, with nested LLQ

Shaping Example



- 10 Mbps SLA on WAN outside interface
- 100Mbps LAN inside interface
- 1 Mbps provisioned for voice
- 3 Mbps provisioned for video
- 6 Mbps provisioned for data
- Data will sometimes burst above 6 Mbps creating congestion

```
class-map VOICE
  match ip dscp ef
class-map VIDEO
  match ip dscp af41
class-map SIGNALLING
  match ip dscp cs3

policy-map NESTED
  class VOICE
    priority 1024
  class VIDEO
    priority 3072
  class SIGNALING
    bandwidth 128
  class class-default
    fair-queue
```

```
policy-map WAN-EDGE
  class class-default
    shape average 10000000
    service-policy NESTED

Interface FastEthernet0/0
  service-policy out WAN-EDGE
```