QOS REFERENCE

Massimiliano Sbaraglia

IOS CISCO QOS CLASSIFICATION HEADER AND PAYLOAD INSPECTION

- Header Inspection
  - Layer 2: MAC Address
  - Layer 3: source and destination IP address
  - Layer 3: source and destination port number and protocol
- Payload Inspection (NBAR Network-based Application Recognition)
  - Enable NBAR on interface we will inspect all incoming packets and match things like:
    - URL
    - MIME type (zip file, image, etc..)
    - User-agent (mozilla, etc…)

CHARACTERISTICS NETWORK TRAFFIC

- **Bandwidth:** speed link in bit per second (bps)

- **Delay:**
  - **one-way delay:** measure the time of the traffic path between source and destination
  - **round-trip delay:** measure the time of the traffic path between source and destination more back
  - **processing delay:** measure the time for a device to perform all task required to forward the packet (lookup RIB, ARP table, outgoing access-list and more)
  - **queuing delay:** the amount of time a packet is waiting in a queue (regard the interface congestion)
  - **serialization delay:** measure the time to send all bits of a frame to the physical interface for trasmission
  - **propagation delay:** measure the time for bits to cross a physical medium

CHARACTERISTICS NETWORK TRAFFIC

- **Jitter:**
  - is the variation of one-way delay in a stream of packet (for example because of congestion in the network some packets are delayed; different time between interval packets on receiver)

- **Loss:**
  - the amount of lost data, show as a percentage of lost packets sent

VOICE AND VIDEO APPLICATION

- Sensitive to delay, jitter and loss
- with voip is necessary a codec process from analogic voice to digital signal (G711 = 20ms audio to 160 bytes data)
- packet voip = IP + UDP + RTP + Voice Data (200 bytes)

- Voip is sensitive to jitter and loss
  - One-way delay: < 150 ms
  - Jitter: < 30 ms
  - Loss: < 1%

- Video is sensitive to bandwidth, jitter and loss
  - One-way delay: 200 – 400 ms
  - Jitter: 30 – 50 ms
  - Loss: 0,1% - 1%

QOS TOOLS

- **Classification**
  - ACLs
  - NBAR (Network-Based Application Recognition): deep applications detect by looking the content IP packet
- **Marking**
  - Change one or more about header fields packet or frame
    - Packet: ToS Byte (IP Precedence; DSCP value)
    - Frame 802.1Q Tag: Priority
- **Congestion Management:**
  - Queueing traffic waiting for the interfaces: the router use classification to decide which packets serve into which queue

QOS CONGESTION MANAGEMENT SCHEDULING

- **FIFO:**
  - one queue where each traffic wait to pass in line (some devices offer multiple output queues)

- **Round Robin Scheduling:**
  - algorithm that cycles through the queues in order to pass traffic (starting with queue 1, queue 2, queue 3 and then goes back to queue 1)

- **Weighted Round Robin Scheduling:**
  - gives more preference to particular queue (ex: 4 packets to queue-1, 2 packets to queue-2 and so on and then goes back to queue-1)
  - Cisco use the scheduler called **CBWFQ** (Class-Based Weighted Fair Queueing) and guarantees a minumum bandwidth to each class on case of congestion; allows to configure the weight as a percentage of bandwidth interface
  - Low latency queueing **LLQ** work for data application ensuring a certain bandwidth to each queue and it is applicabile to traffic that has to be sent immediately without waiting (priority queue) and this kind of queue has precedence that all other queues
    - It is important to set a limit to the priority queue to avoid the queue starvation (if the scheduler is too busy to serve the LLQ queue, is possible that the other queue arenever served

QOS CONGESTION MANAGEMENT SCHEDULING

- **Policing and Shaping**
  - used to limit the bit rate: policing with discarding traffic and shapers will hold packets in a queue adding delay
  - Usually the policing is configured on ISP versus the customer to ensure only CIR (Committed Information rate) through incoming policing
  - Usually the shaping is configured on customer router to prevents dropped traffic at the ISP
- **Congestion Avoidance**
  - TCP windows size setting
  - Tail drop prevent
  - WRED: monitoring the output queue with a value queue (empty or full) and threshold value (minimum and maximum) and we can setting different treatment based on their marking to allow congestion avoidance

QOS POLICING ACTIONS

- Policer allow one of the following actions:
  - Pass traffic
  - Drop traffic
  - Remark packets with a different DSCP or IP-Precedence
- Policer categories:
  - Conforming: packets is compliant within the CIR
  - Exceeding: packets is using the excess burst capability
  - Violating: totally external to CIR
- Policer techniques:
  - Single rate, two-color: one token bucket
  - Single rate, three-color: two token buckets [first bucket is for Bc (committed burst); second bucket is for Be (excess burst)]
  - Dual rate [first rate is CIR (committed rate); second rate is PIR (Peak rate)], three-color: two token buckets