TEMA 4 QOS

1 - Qué es QoS?

La calidad de servicio (QoS) en una red se refiere a la habilidad de una red para dar servicios de alta calidad a un usuario final gestionando a la vez de manera efectiva el tráfico. Esto implica manipular el tráfico según las necesidades de las aplicaciones, permitiendo asignar prioridades, controlar el ancho de banda y emplear diferentes tecnologías y algoritmos para garantizar una experiencia óptima.

- QoS en arquitecturas de red

-> Conmutación de circuitos: se utiliza en el sistema telefónico para asignar recursos fijos cuando se inicia una conexión y estos recursos permanecen en uso hasta que se termina la conexión. La ventaja es el retraso bajo y predecible.
-> Conmutación de paquetes: permite que varios dispositivos se comuniquen simultáneamente y necesita asignarles recursos suficientes para comunicarse. Problemas que puede tener: la falta de previsibilidad para aplicaciones en tiempo real y la alta fluctuación que se produce cuando una gran cantidad de paquetes pasan de un enlace de red más rápido a uno más lento, o cuando varios enlaces de red se fusionan en un solo enlace. Cuando esto sucede búferes de memoria se llenan, aumentando el tiempo que tardan los paquetes en atravesar una red.

- Asignación de ancho de banda estática y dinámica

-> Estática: a los flujos se les asigna una cantidad fija de ancho de banda de la capacidad del enlace y, por lo tanto, otros flujos no pueden utilizar esa cantidad fija de ancho de banda del enlace.
-> Dinámica: a los flujos se les asigna una cantidad de ancho de banda, pero si el flujo no usa ese ancho de banda, otros flujos pueden usarlo.

Por definición, las redes de ancho de banda estática tienen que operar en el peor de los casos (picos de tráfico), mientras que las redes de ancho de banda dinámico operan en base al tráfico promedio.

2 - Parámetros de QoS

Son métricas utilizadas para determinar el rendimiento de diferentes aplicaciones en Internet.

-> Ancho de banda (rendimiento): es la cantidad de capacidad de enlace que requiere un flujo. Se mide en bits/s.
-> Volumen: es el número de bytes transmitidos en un intervalo de tiempo.
-> Pérdidas: es el número de paquetes perdidos en una transmisión. Las pérdidas suelen ocurrir cuando hay congestión, las causas pueden ser:

i) discrepancia de ancho de banda (enlace de alta capacidad con una baja).
ii) agregación (múltiples enlaces multiplexados en un solo enlace).
iii) cuando las colas están llenas y es necesario descartar paquetes.

3 -> Retraso: es el tiempo que tarda un paquete individual en atravesar la red. 2 tipos:

- Latencia (ms): tiempo que tarda un paquete en viajar desde el origen al destino (RTT). Se compone de varios retrasos:
  - procesamiento (tiempo que tarda un dispositivo en realizar todas las tareas necesarias para reenviar el paquete),
  - cola (cantidad de tiempo que un paquete espera en una cola),
  - serialización (tiempo que lleva enviar todos los bits de una trama a la interfaz física para su transmisión) y
  - propagación (tiempo que tardan los bits en cruzar un medio físico).
- Jitter: se define como una medida de la variación en el retardo de los paquetes. El alto jitter ocurre cuando hay congestión, es decir, cuando hay una discrepancia de ancho de banda o cuando hay agregación. El conmutador se ve obligado a almacenar en búfer los paquetes, lo que aumenta el tiempo que tardan los paquetes en atravesar la red.
  Supongamos que un teléfono IP envía un flujo constante de paquetes de voz con una separación de 10 ms. Debido a la congestión de la red, algunos paquetes se almacenan en el buffer y, por lo tanto, se retrasan y se produce un retardo entre paquetes. Se ha de lidiar con esto para que haya un retraso constante o experimentará una mala calidad de voz.

4 - Tipos de tráfico

-> Aplicaciones no interactivas: estas aplicaciones normalmente cargan o descargan archivos. Ejemplos son la transferencia web o la transferencia de archivos, entre otras.
• Entonces, el ancho de banda es importante, ya que minimiza los tiempos de transferencia de archivos. Las pérdidas de paquetes, el retraso de un extremo a otro y la jitter no son importantes, ya que no estamos interactuando con la descarga.

-> Aplicaciones interactivas: dichas aplicaciones a menudo se conectan a servidores.
• Estas aplicaciones no requieren ancho de banda, ya que la transferencia consta de sólo unos pocos bytes. El retraso no es importante. Lo mejor que podemos hacer con QoS es priorizar estas conexiones sobre las aplicaciones que consumen mucho ancho de banda.

-> Aplicaciones de voz y vídeo: son muy sensibles al retraso, la jitter y pérdida de paquetes.
• El ancho de banda no es un problema en VoIP, pero el retraso sí lo es, ya que no querrás esperar a escuchar a la otra persona.
• Una buena conexión VoIP debería garantizar un retraso de extremo a extremo ≤ 150 ms, una fluctuación ≤ 20 −30 ms y una pérdida de paquetes ≤ 1%.
• El tráfico de vídeo tiene requisitos similares a los del tráfico de voz, con un retraso de extremo a extremo entre 200 y 400 ms, una fluctuación entre 20 y 50 ms y una pérdida de paquetes entre 0,1% y 1%.

5 - Modelos de servicio de QoS

Es la capacidad de una red para proporcionar el servicio requerido por un tráfico de red específico de un extremo a otro o de un extremo a otro.

-> Best effort: no hay garantía de flujos en términos de ancho de banda, retardo, jitter y características de pérdida. Se caracteriza por ser tratado con baja prioridad en caso de que

los sistemas de buffer apliquen una política de prioridad de cola o simplemente, aplicar FIFO. Internet es un ejemplo.
-> Soft QoS (DiffServ): Los flujos se agrupan en clases y se da un tratamiento especial a las clases y no a los flujos, por lo que algunos tráficos se tratan mejor que otros en términos de mejor gestión, mayor ancho de banda, menor retardo (de extremo a extremo y jitter). La clasificación de clases se puede realizar mediante los bits de precedencia.
-> Hard QoS (IntServ): se garantiza una calidad de servicio robusta en términos de ancho de banda, retrasos y pérdidas. Esto se logra mediante el uso de un esquema de reserva, junto con colas de prioridad (PQ), configuración del tráfico y esquemas de monitoreo y programación. La clasificación de flows se puede realizar mediante ACL.

6 - Clasificación de paquetes en Internet

El flujo se identifica mediante una ACL, la clase se establece marcando los bits de precedencia y luego, desde ese enrutador, los paquetes se clasifican como pertenecientes a una clase identificando los bits de precedencia. Los bits de precedencia son parte de algún campo del paquete, que posteriormente se amplió a 6 bits en arquitecturas DiffServ (campo DSCP).

– Fórmulas aparte –

- Disciplinas de programación de colas

Las disciplinas de programación de colas definen cómo se almacenan en el búfer los paquetes mientras esperan ser transmitidos. Tienen dos parámetros principales: ancho de banda, que determina qué paquete se transmite a continuación, y espacio de búfer, que determina qué paquete se descarta a continuación (si es necesario). Las disciplinas de programación de colas influyen en el retraso (latencia y fluctuación), el rendimiento o las pérdidas de un flujo.

- First-in-first-out (FIFO) + drop-tail

El mecanismo de drop-tail significa que los paquetes que llegan se descartan cuando la cola está llena, independientemente del flujo. El principal problema con FIFO es que el servicio recibido por un flujo se ve afectado por las llegadas de paquetes de todos los demás flujos. Por otro lado, la gestión de buffers de colas de caída obliga a los enrutadores a tener un tamaño de buffer muy grande para mantener una alta utilización, lo que resulta en colas estacionarias y, posteriormente, largos retrasos. Las colas en estado estacionario (steady-state) significan que, en general, las colas tardan algún tiempo en alcanzar el régimen predicho por la teoría de colas (e.g el sistema de colas M/M/1). La razón es que las colas no aumentan instantáneamente a los altos niveles predichos por cargas elevadas, sino que aumentan (en el período transitorio) hasta que se alcanza el estado estacionario (el estado de equilibrio).
Estado transitorio significa que el estado de la cola depende del tiempo t, mientras que estado estable significa que el estado de la cola no depende del tiempo t. Se alcanza un estado estable pesado con retrasos prolongados después de algún tiempo con una carga pesada (ρ ~ 1).

7 - (lock-out)

Además, los flujos sufren de ráfagas, sincronización y lock-out. La transmisión en ráfagas ocurre cuando los hosts envían una transmisión de gran cantidad de paquetes en un corto

período de tiempo. Las ráfagas provocan grandes retrasos y pérdidas en las colas. Además, los hosts reaccionan de la misma manera cuando se producen pérdidas de paquetes en períodos de ráfagas. Esto provoca un efecto de sincronización que hace que suceda la misma situación en el mismo flujo cuando llegan nuevos paquetes a la cola. Entonces, un efecto secundario de la ráfaga y la sincronización es que unos pocos flujos pueden monopolizar el espacio de la cola (el efecto de lock-out).

Hay diferentes mecanismos de desconexión:

• Sincronización: se puede solucionar mediante la *random drop*, que consiste en eliminar aleatoriamente algunos paquetes de la cola;
• Lock-out: se puede solucionar utilizando *front drop* que consiste en dejar paquetes al principio de la cola;
• High steady-state queuing: se puede resolver usando la *early drop*, que consiste en descartar paquetes antes de que la cola se llene;
• Burstiness: se puede resolver usando la *early drop*, pero teniendo cuidado de no descartar paquetes demasiado pronto porque la cola puede reflejar solo ráfagas y no una sobrecarga real;
• Flujos frágiles: se puede resolver con *preference dropping*, identificando flujos sobre la marcha y marcando flujos críticos para un tratamiento específico (sin eliminarlos);
• Mal comportamiento del host: se puede solucionar descartando paquetes proporcionalmente.

- Detección temprana aleatoria (RED)

RED funciona monitoreando la carga de tráfico en puntos de la red y descartando paquetes estocásticamente si la congestión comienza a aumentar. El resultado del descarte es que la fuente detecta el tráfico descartado y ralentiza su transmisión, y está diseñada principalmente para operar en redes TCP/IP. RED comienza a descartar paquetes aleatoriamente cuando el tamaño promedio de la cola excede un valor umbral (minth). La tasa de caída de paquetes aumenta linealmente a medida que aumenta el tamaño promedio de la cola hasta que el tamaño promedio de la cola alcanza el umbral máximo (maxth). A partir de entonces, una determinada fracción (llamada denominador de probabilidad de marca) de paquetes se descarta, también de forma aleatoria.

Proportionally fair allocation implementation in real queue systems results in WFQ queue management systems.
Max-min implementation in real queue systems results in PQ-RR queue management systems.