GG1

Tema 1-1: Introducción a las técnicas usadas en la calidad de servicio (QoS)

José Manuel Arco Rodríguez

Índice

- Introducción a la QoS
- Necesidades del tráfico de usuario
- Técnicas utilizadas en QoS
  - Control de admisión de conexiones
  - Encaminamiento con QoS
  - Conformado
  - Función Policía
  - Control de congestión
  - Control de flujo
  - Planificadores de tráfico
- Señalización

Evolución de las redes

- Internet (IP)
  - Desde el punto de vista de los usuarios el aumento de clientes y servidores implica más capacidad de transmisión (Ancho de banda, Bandwidth (BW), Throughput o velocidad de la línea o enlace)
    - Para 2018 se prevé 22 billones de dispositivos
  - Desde el punto de vista del tráfico intercambiado entre los usuarios
    - No interactivo , aplicaciones tipo
      - Correo, ftp, www, pear to pear, radio, video (YouTube) (80% del tráfico), TVIP, computación cooperativa GRID*, etc.: Se necesita **más BW** (capacidad)
    - Interactivo, aplicaciones tipo
      - VoIP, videoconferencia (tráfico interactivo): Se necesita **Menor retardo** (<150 ms)

- En el centro de investigación de partículas, LHC se provocan 1.010 colisiones anuales que generan una información de 10.000 TeraBytes, (10 PetaBytes). Esta información se distribuye a varios centros en Europa que exige redes de transmisión de varios Giga bps (Gbps)

Transmisión de voz en redes de paquetes

Conocimientos previos

- VoIP, La voz para ser transmitida se digitaliza tomando una muestra de la misma cada 125 usec, cada muestra se codifica en un byte, (se generan 8000 muestras/sec → 64 Kbps). Para transmitirla por una red de paquetes IP, las muestras se agrupan en paquetes de unos 200 bytes.
- Lo que mas afecta a la voz es el retardo de transmisión que afecta a la calidad de la llamada, el sonido tiene que llegar rápido para que la posible contestación del otro interlocutor sea rápida y real.
- Un ejemplo de comunicación con un retardo de varios segundos es la comunicación vía satélite en un telediario con un corresponsal, se aprecia el tiempo que el corresponsal tarda en contestar.
- Hay otro efecto del retardo es el eco de la voz, producido porque parte de la voz es devuelta por el otro teléfono y es escuchada como eco (molesto), si hay un retardo superior a 300 msec.

Evolución de las redes

- Redes públicas (de los operadores)
  - Red Digital de Servicios Integrados (>68)
  - Red Digital de Servicios Integrados de Banda Ancha (>88)
  - También plantean las mismas necesidades que en Internet, lo solucionan con ATM (Asyncronous Transfer Mode)
  - Hoy en día los operadores emplean IP

Necesidad de calidad de servicio

- Usuarios utilizan aplicaciones multimedia debido a la disponibilidad de:
  - Potentes dispositivos personales
  - Acceso a redes de telecomunicación
- Estas aplicaciones requieren de la red de transmisión, el cumplimiento de ciertos parámetros, es decir, lo que se conoce como calidad de servicio, (Quality of Service, QoS)
- Estos parámetros suelen ser: ancho de banda, retardo, variación del retardo (jitter), tasa de error

Necesidad de calidad de servicio (Cont.)

- Se implementa con un acuerdo usuario y la red de transmisión
  - También se usa para especificar QoS para el tráfico que atraviesa varios operadores
- El acuerdo se refleja documento llamado SLA (Service Level Agreement), con los valores de los parámetros de transmisión
- El SLA tiene implicaciones para los dos partes
  - El que envía el tráfico debe limitar su velocidad máxima y/o media
  - El operador (en base a lo anterior) debe cumplir con los parámetros de transmisión
- El concepto de SLA también se aplica en otros entornos:
  - En las aplicaciones accesibles via Internet (e.g. Web Services) para fijar el servicio que ofrece el operador
  - En la Cloud Computing, para regular los servicios contratados

Ejemplo de QoS

- La QoS puede negociarse a la baja hasta llegar a un acuerdo o incluso rechazarse
- Un ejemplo sencillo es un operador con un router con 20 puertos conectados a sendos abonados por líneas de 10 Mbps y con un puerto al resto de la red troncal de 100 Mbps,
- Si todos los usuarios quieren el máximo de BW disponible el operador podría ofertar 5 Mbps a cada uno

Necesidad de calidad de servicio (Cont.)

- La QoS se logra por un acuerdo entre el usuario y la red
  - El usuario transmite tráfico ajustándose a un perfil declarado
  - La red planifica recursos para cumplir lo pactado
- La QoS se despliega en redes corporativas y de proveedores, por ejemplo para dar ofrecer acceso a Internet y algún tipo de llamadas gratis (VoIP)
- La QoS es difícil que se despliegue en la totalidad de Internet
  - Implica actualizar todos los routers, (10^5 muchos y de diferentes operadores)

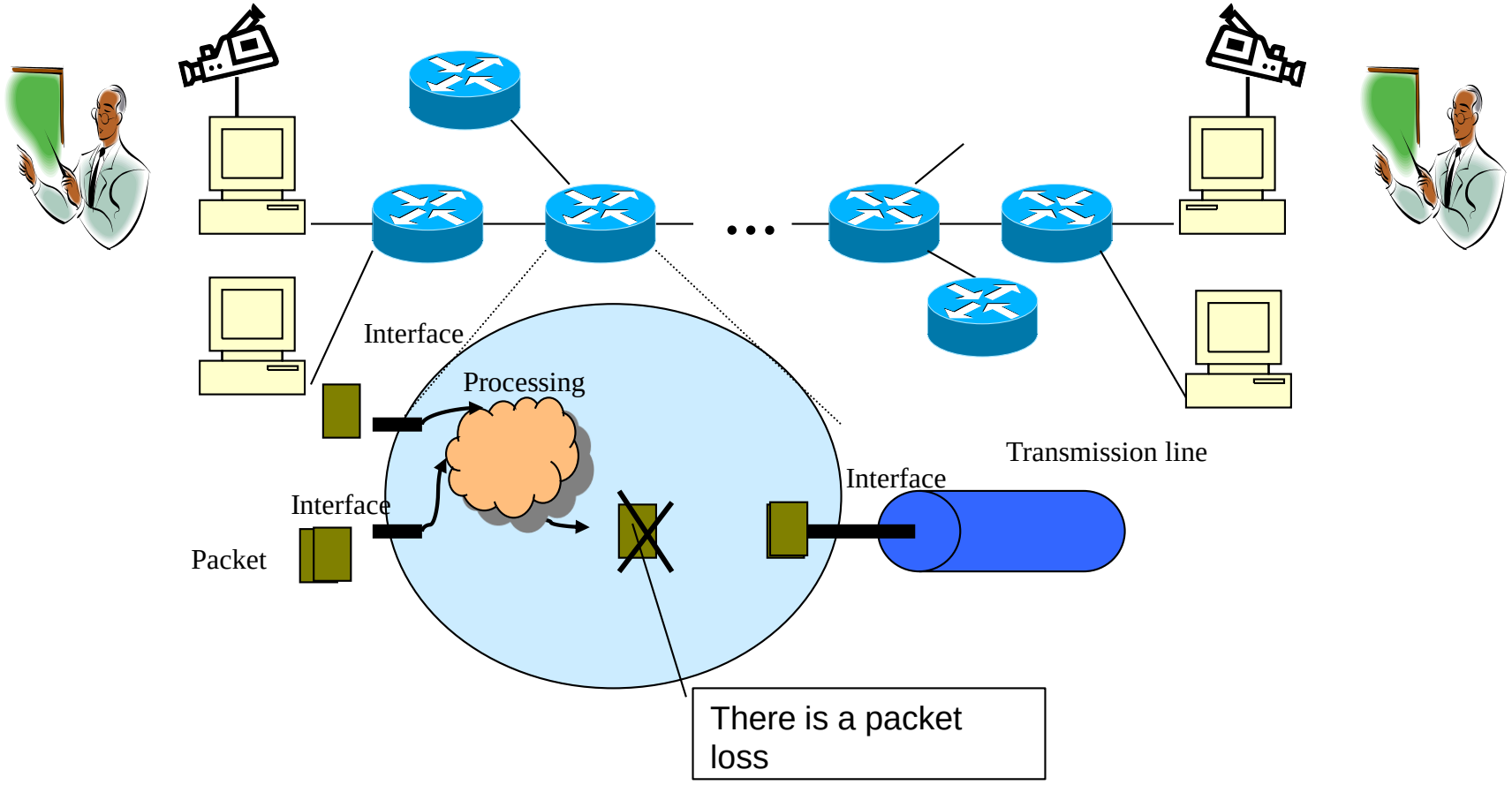
Transmisión de paquetes en la red: Funcionamiento de un router

- Cuando se recibe un paquete el router debe procesar los campos de las cabeceras 2 y 3
- Después consultar la tabla de encaminamiento para ver el interfaz de salida
- En cada etapa se puede producir congestión que se soluciona creando una cola de espera
- La congestión se manifiesta cuando en alguna etapa hay más tráfico del que se puede transmitir (o procesar), entonces se tiran paquetes

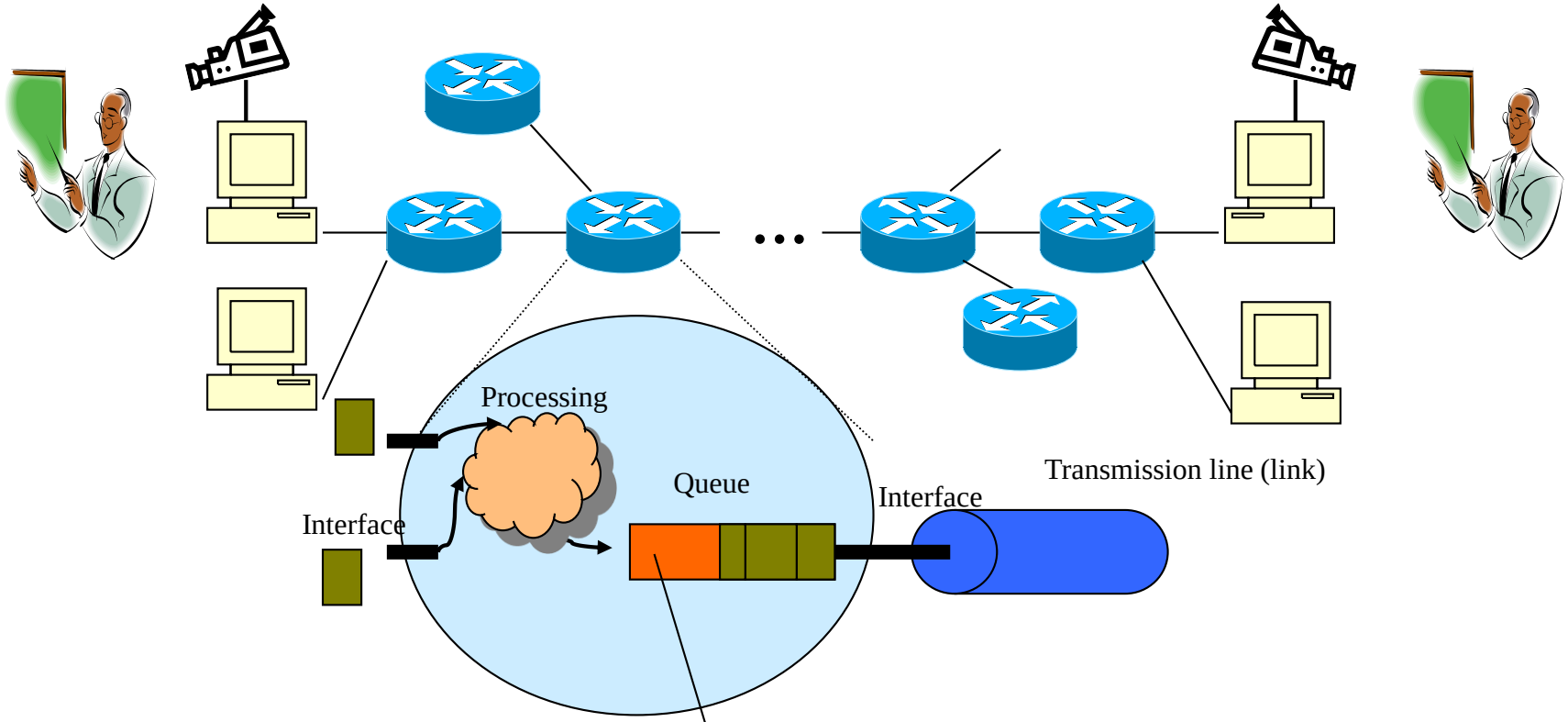
Transmisión de paquetes en la red: Línea de transmisión

- Para transmitir un paquete hay que transmitir cada uno de sus bits, cada uno en un **tiempo** de 1/Vtx (velocidad de la línea)
- Para transmitir un bit el router emisor genera una señal correspondiente a un "1" o un "0" en la línea transmisión
- Esta señal puede considerarse una perturbación electromagnética similar a la onda que se produce al lanzar una piedra en el agua
- Esta onda tarda un **tiempo** en viajar hasta el router destino, función de la distancia y del medio

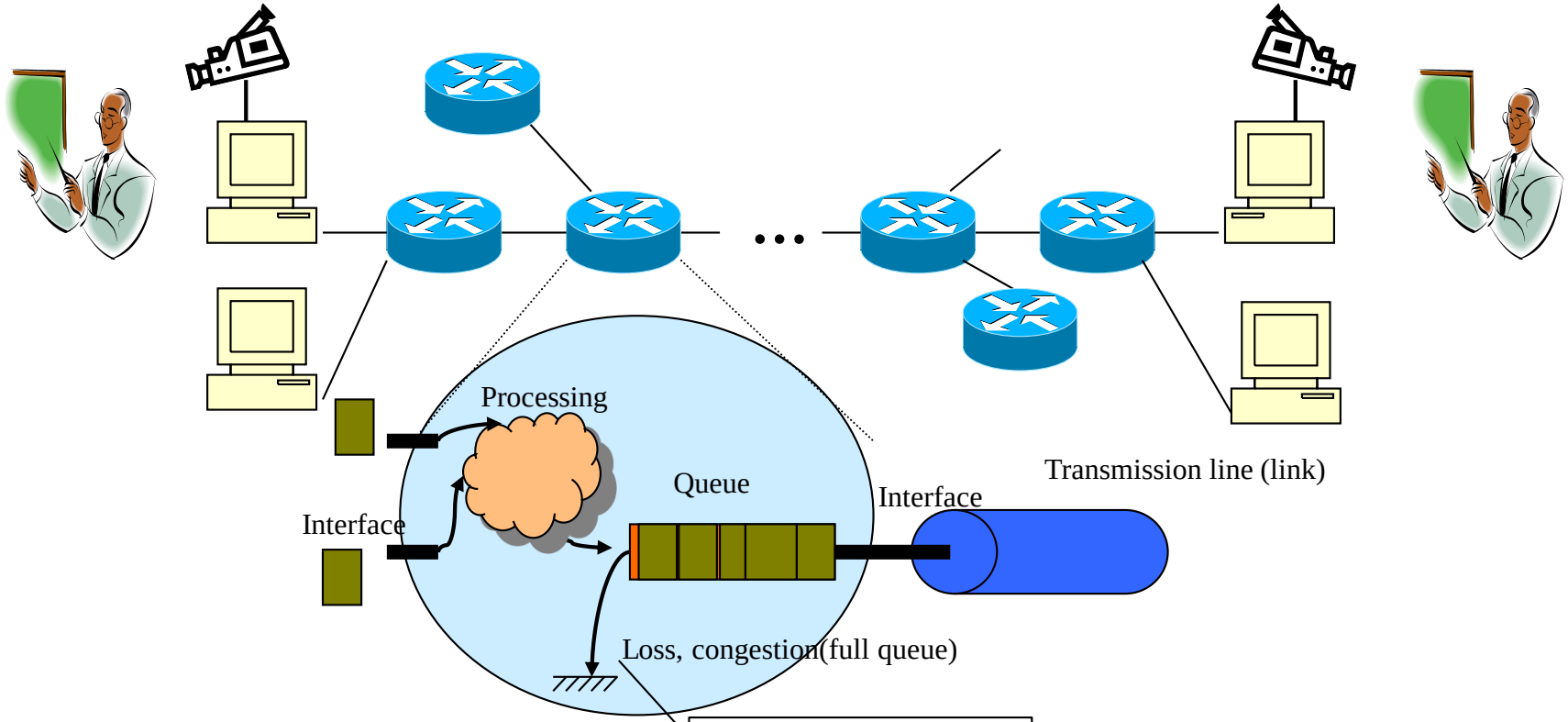
Necesidad de buffering



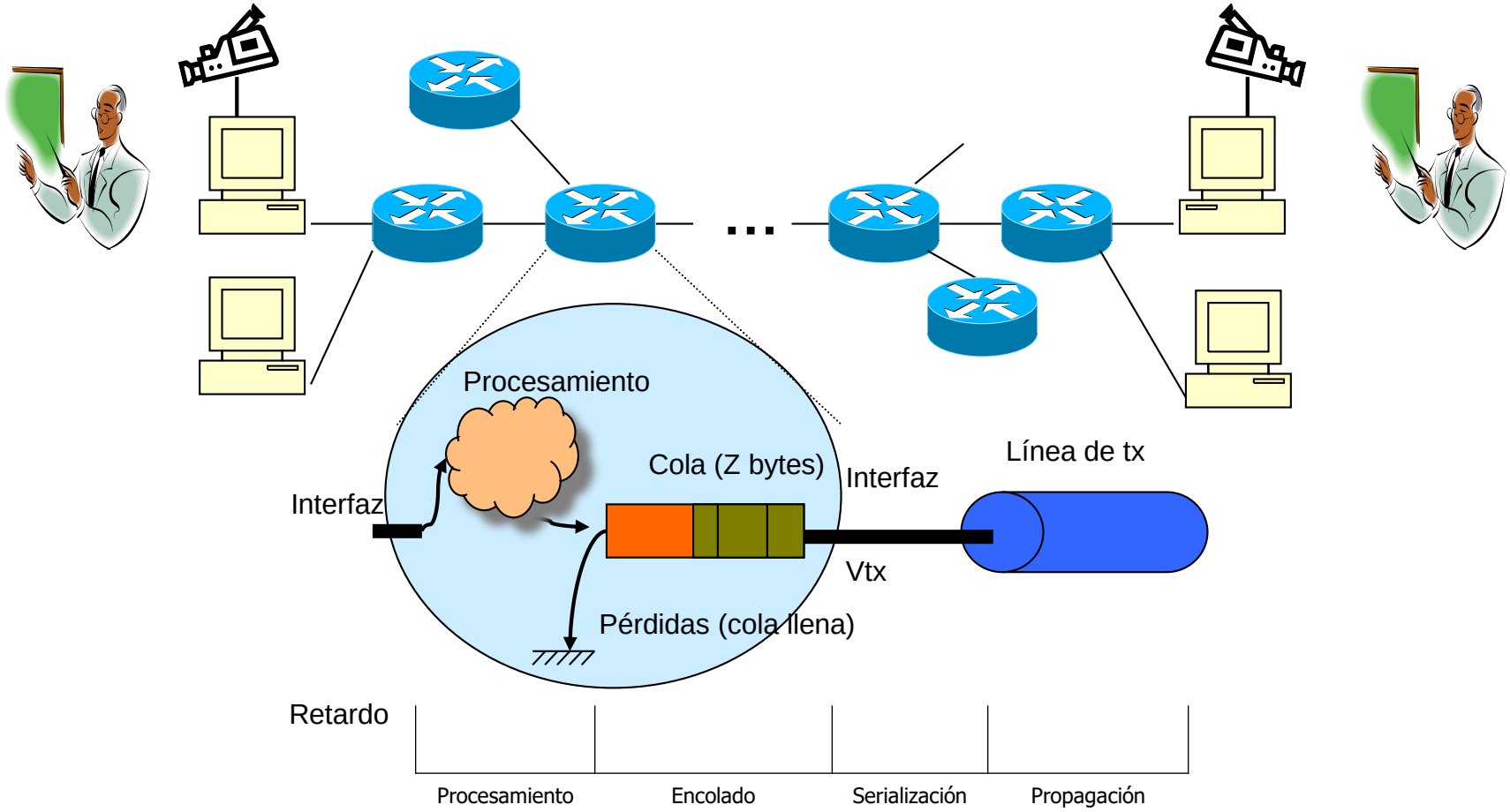
Router working



Router working



Modelo de red de transmisión sin QoS



Retardo en red sin QoS

- Retardo de transmisión de un paquete tiene las siguientes componentes:
  - Procesado en el nodo, tiempo desde que se recibe un paquete en un interfaz y se manda al planificador del interfaz de salida
    - En los nodos actuales, es despreciable (de 10 a 20 µseg). Influye en el throughput
  - Retardo en cola, depende del tipo de planificador de tráfico, tamaño de la cola, ocupación de la cola
  - Retardo de transmisión (serialización), tiempo para transmitir el paquete por la línea de transmisión
    - Depende de la velocidad de transmisión de la línea
    - A partir de 34 Mbps es despreciable (1500 bytes a 64 Kbps 187 mseg, a 34 Mbps son 0,4 mseg, a 10 Gbps son 0,0012 mseg.)
    - Se no considerar ya que el retardo en cola es varias veces este retardo
  - Retardo de propagación
    - Depende fundamentalmente de la distancia, en menor medida del medio
    - Por ejemplo en fibra óptica es de 5 mseg/1.000 km

Influencia de la cola de salida en la QoS

- Un tamaño mas grande implica
  - Tasa de errores baja
    - Beneficia al tráfico de datos, con menos errores menos retransmisiones y se completa antes la transferencia de un fichero
      - En la práctica es irelevante que acabe la tranferencia unas décimas segundo antes
  - Retardo mayor, mas tiempo de espera del tráfico interactivo, podemos superar los 150 mseg
- Tamaño mas pequeño
  - Puede disparar la tasa de errores y aumentar de forma notoria la duración de una transferencia de un fichero
  - Baja el retardo transmisión pero el aumento en la tasa de errores puede degradar la calidad del sonido/imagen

QoS en Internet (IP)

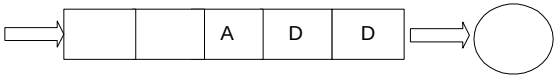
- Internet trata a todo el tráfico por igual, no tiene QoS
- También se dice que Internet tiene un único servicio llamado Best Effort
- Es un red no conectiva que puede producir los siguientes problemas al transmitir tráfico
  - Desordenar
  - Tirar
  - Duplicar
  - Fragmentar
  - Retardos grandes
  - Retardo variable

Influencia del protocolo de transporte en la QoS

- TCP, (protocolo fiable usando retransmisiones de mensajes no asentidos)
  - Usado para tráfico no interactivo (datos) para conseguir transmisión libre de errores
  - No recomendable para tráfico interactivo, en caso de error introduce un retardo adicional para su recuperación, que hace que llegue tarde el mensaje al destino y no se pueda reproducir (imagen/voz)
- UDP, (protocolo no fiable)
  - Usado para tráfico interactivo (imagen/voz) para conseguir bajos retardos
  - Errores no son recuperados, el codec del destino los enmascara

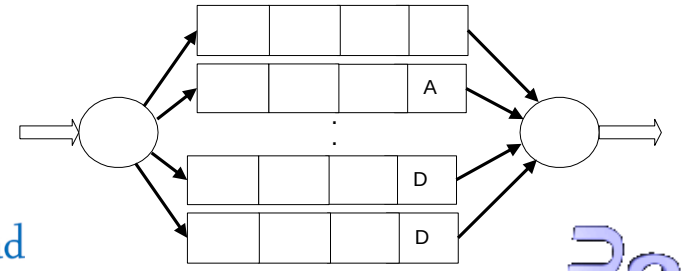
Retardo en red con QoS

- Para que haya QoS para un flujo i, se debe cumplir:
  - **Usuario**: limite el tráfico emitido, a un patrón, el más usado es (σ_i,ρ_i)
    - σ_i tamaño máximo de la ráfaga a *Vtx* para el flujo *i*
    - ρ_i velocidad media para el flujo *i*
  - **Operador:** en la cola de salida los nodos tenga un planificador de trafico, (e.g., WFQ) que garantice la velocidad media ρ_i de ese flujo
    - Equivalente en tráfico rodado: de reserva de carril



FIFO

(Sin QoS)



Retardo en red con QoS (Cont.)

- Con QoS, el retardo máximo extremo a extremo (con n nodos iguales) de un paquete del flujo *i*, será:

$$\text{Re}\,tardo = \sum_{n\_nodos} R.cola + R.procesamiento + R.tx.paquete + \sum_{n-1\,enlaces} R.propagación \Big|_{\max,\, n\uparrow\uparrow}$$

$$\text{Re}\,tardo \approx \frac{\sigma_i}{\rho_i} + \sum_{n\_nodos} \frac{L\max}{\rho_i} + \sum_{n\_nodos} \left(0 + \frac{L\max}{Vtx} + R.propagación \right) \Big|_{\max}$$

- El valor del retardo de cola está calculado en esta referencia [Zhang]:
  - H. Zhang, "Service Disciplines for Guaranteed Performance Service in Packet-Switching Networks". Proceeding of the IEEE, Octubre 1995.

Comparativa del retardo en red con y sin QoS

- El retardo de cola es:

$$R.cola = \frac{\sigma_i}{\rho_i} + \sum_n \frac{L\max}{\rho_i}\bigg|_{\sin\uparrow} \approx \sum_n \frac{L\max}{\rho_i}$$

- El retardo puede ser más o menos grande en función del peso (Φ_i) o ancho de banda del flujo i :

$$\rho_i = \phi_i * Vtx \quad 0 < \phi_i \le 1$$

$$R.cola = \sum_n \frac{L\max}{\phi_i * Vtx}$$

$$\sum_n \frac{\mathrm{Lmax}}{\mathrm{Vtx}} \le R.cola < \sum_n \frac{L\max}{\phi_{i\min} * Vtx}$$

- Comparado con el retardo de cola sin QoS:

$$R.cola < \sum_n m \frac{L\max}{Vtx}$$

- Mejora

$$\frac{\text{R. sin QoS}}{\text{R. con QoS}} \gg f_i m$$

Conclusiones del retardo

$$Retardo \approx f(\frac{1}{Vtx}, distancia, retardo\ de\ cola)$$

- Retardo configurable por el usuario mediante (σ_i,ρ_i) y el operador
  - El tamaño del buffer, debe ser un balance entre la tasa de errores y el retardo
- Velocidad alta no implica necesariamente retardo bajo
  - Por ejemplo, un enlace vía satélite (varios Mbps y 600 msec/satélite)
  - Simil, una autopista con más carriles no tiene porqué ser más rápida (semáforos, peajes, rotondas, etc)
- Menor retardo con redes con QoS

Variación del retardo en red

- No todos los paquetes sufren el mismo retardo, ya que puede encontrar las colas con diferente ocupación, de ahí el parámetro *variación de retardo* o *jitter*
- El retardo y su variación afecta negativamente al tráfico en tiempo real interactivo (bidireccional)
- El jitter se corrige en el receptor, insertando un buffer del que se sacan los paquetes, con el mismo espaciado que el de emisión
- Inconveniente: el buffer introduce un retardo adicional

Variación del retardo en red (Cont.)

- El jitter máximo extremo a extremo, sin QoS es:

$$jitter\_\max = n * m\frac{L\max}{Vtx}$$

- El jitter extremo a extremo, máximo con QoS es [Zhang]:

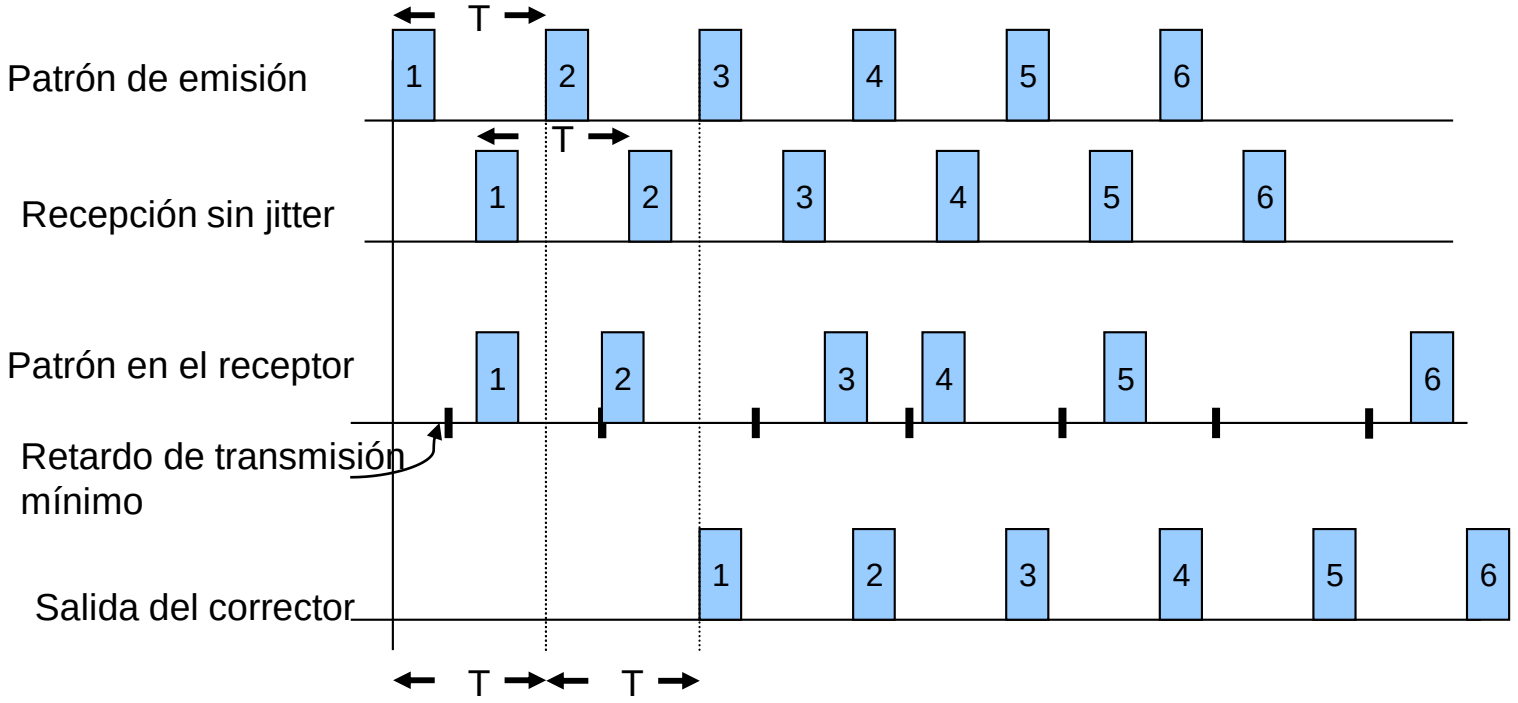
$$jitter\_\max = \frac{\sigma_i + nL\max}{\rho_i} = \frac{\sigma_i + nL\max}{\phi_i Vtx}$$

$$si\ \ \sigma_i + nL\max \approx m\,\mathrm{Lmax}$$

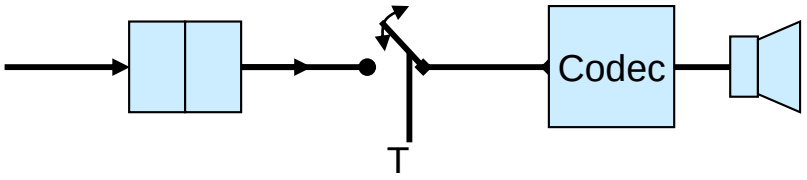
$$jitter\_\max \approx \frac{mL\max}{\phi_i Vtx}$$

reducción de jitter en un factor de » $f_i n$

Supresión del jitter



Buffer corrector del jitter



Congestión en red sin QoS

- Sin QoS no hay control de tráfico ni compromisos
- Con ocupación de la cola inferior a la saturación:
  - El retardo y la variación del retardo, pueden variar entre el mínimo y el máximo
  - La tasa de errores, se reduce a valores muy bajos (errores físicos)
  - El ancho de banda instantáneo menor o igual a *Vtx*
- Cuando se llena la cola (congestión), aumenta la tasa de errores
  - Efecto en la red
    - El retardo y la variación del retardo, pueden aproximarse al máximo
  - Efecto extremo a extremo
    - UDP no recupera errores, baja la velocidad efectiva
    - TCP baja la velocidad efectiva y sube el retardo efectivo por las retransmisiones

Congestión en red con QoS

- La ocupación (sin llenarse) de la cola influye en:
  - El retardo y la variación del retardo inferiores al máximo garantizado
  - La tasa de errores, se reduce a los errores de transmisión (≈0)
  - El ancho de banda puede ser mayor o igual al ancho de banda garantizado ρ_i (depende del planificador de tráfico)
- Congestión, en teoría no ocurre, pero si hay
  - Flujos sin QoS, los primeros en sufrir pérdidas
  - Flujos con QoS, los siguientes en sufrir pérdidas

Referencias

- J. Chao, X. Guo, “Quality of Service Control in High-Speed Networks” Editorial Wiley. “2002
- James F. Kurose & Keith W. Ross, “Redes de Computadoras”, 5a Edición, Editorial Pearson, pg 29, 312, 632
- W. *Stallings*. "*Redes e Internet de alta velocidad*: Rendimiento y calidad de servicio". Prentice Hall. 2ª Ed. Pg 16
- A.S. Tanenbaum "Redes de ordenadores, tercera edición". Capítulo 5.