Calidad de servicio en redes IP*

Evaluation of Quality of Services in IP networks

Federico Gacharná

Ingeniero de Sistemas de la Universidad Autónoma de Colombia. Especialista en Seguridad Informática de la Universidad Sergio Arboleda. Diplomado en Docencia Universitaria de Uniminuto y candidato a Magister en Seguridad Informática de la Universidad Oberta de Cataluña. Docente del programa de Tecnología en Redes de Computadores y Seguridad Informática de Uniminuto.
fgacharna@uniminuto.edu

Resumen

La calidad de servicio (QoS), ofrecido por una red de datos es importante, ya que tiene relación directa con la cantidad de información que se puede transmitir, el tiempo que tarda en llegar a su destino y las variaciones en estos retardos. Por tanto, resulta útil una recopilación sucinta de las principales estrategias y tecnologías para enfrentar posibles problemas de rendimiento, basadas en QoS.

Palabras clave: QoS, Redes IP, VoIP, Tráfico de Red, Latencia, Retardo.

Abstract

The Quality of Services (QoS), provided by data network is relevant due to a direct relation with the amount of data that can be transmitted, the time it takes to reach its final destination, and variations caused by delays. Thus, it`s usefull to know the main strategies and technologies to face possible issues related to performance based on QoS technology.

Keywords: QoS, IP networks, VoIP, Network Traffic, Latency, Delays.

* Internet Protocol (Protocolo de Internet)

Introducción

Se entiende por Calidad de Servicio la posibilidad de asegurar una tasa de datos en la red (ancho de banda), un retardo (delay) y una variación de retardo (jitter), teniendo en cuenta los valores contratados. En las redes Frame Relay o ATM la calidad de servicio se garantiza mediante un contrato de CIR (Commitated Information Rate) con el usuario. En redes IP se han diseñado herramientas como protocolos de tiempo-real RTP [1] y de reservación RSVP [2]. Cuando se soporta un servicio de voz sobre IP (VoIP), los paquetes son cortos y el encabezado es largo, comparativamente; en este caso se requiere un encabezado reducido y un proceso de fragmentación e intercalado LFI, por tanto, mediante QoS (Quality os Service) se tiende a preservar los datos con estas características.

Por otra parte, cabe resaltar que los servicios tradicionales de Internet (SMTP o FTP, por ejemplo), disponen de una calidad denominada "best effort"; es decir, que la red ofrece el mejor esfuerzo posible para satisfacer los retardos mínimos, lo cual no es mucho, pero es suficiente para servicios que no requieren tiempo-real como el web, sin embargo, para servicios de tipo "real-time" (voz y vídeo) se requiere una latencia[1] mínima.

La discusión aquí presentada girará entorno al planteamiento de diferentes estrategias y tecnologías, así como elementos necesarios para ofrecer medios de control de intrusos, con el fin de evitar problemas de rendimiento y escalabilidad de la red.

Variantes de servicio

Los servicios de datos y multimedia tienen distintos requerimientos de calidad en cuanto a latencia y jitter[2]. Para satisfacerlos, se acude al manejo de colas de paquetes, reservación de ancho de banda y gestión de tráfico. Para obtener estos resultados se han definido variantes de servicios.

En la Tabla 1 se encuentran las variantes de servicios, las cuales se complementan con las características de calidad de servicio.

1) **CALIDAD DE SERVICIO: VARIANTES Y CARACTERÍSTICAS**	
CoS (Class of Service)	CoS usa 3 bits en un campo adicional de 4 Bytes (etiqueta Tag) dentro del protocolo MAC. Estos 3 bits permiten definir prioridades desde 0 (máxima) a 7 (mínima) y ajustar un umbral en el buffer de entrada y salida del switch LAN para la descarga de paquetes.
II. IEEE 802.1P	Determina el uso de un Tag en el encabezado de MAC con 3 bits de precedencia. Se define el protocolo para registración **GARP** (Generic Attribute Registration Protocol). Las aplicaciones específicas GARP son el registro de direcciones multicast **GMRP** (Multicast GARP) y de usuarios VLAN con protocolo **GVRP** (VLAN GARP).
III. IEEE 802.1Q	Servicio VLAN para realizar enlaces troncales punto-a-punto en una red de switch.
IV. IEEE 802.3X	Examina el control de flujo en enlaces Ethernet del tipo full-dúplex. Se aplica en enlaces punto-a-punto (Fast y Gigabit Ethernet). Si hay congestión se emite, hacia atrás, un paquete llamado "pause frame" que detiene la emisión por un período de tiempo determinado. La trama "time-to-wait zero" permite reiniciar la emisión de paquetes.
V. IEEE 802.1D	Define el protocolo **STP** (Spanning-Tree Protocol). Permite que en una red de bridge y switch de muchos componentes se formen enlaces cerrados para protección de caminos. Se intercambia información de topología de la red que permite construir el árbol. Así se crean puertas redundantes en el cableado, el protocolo STP inhabilita automáticamente una de ellas y la habilita en caso de falla de la otra. Cada puerto tiene una ponderación en costo (el administrador de la red puede modificar el costo para dar preferencia a cierta puerta).
ToS *Type of Service QoS*	Es sinónimo de CoS en la capa 3. Sobre el protocolo IP se define el ToS con 3 bits para asignar prioridades. Se denomina señal de precedencia. En redes IP se define la tasa de acceso contratada CAR (Committed Access Rate). La calidad QoS se garantiza mediante protocolos de reservación RSVP y de tiempo real RTP

Tabla 1. Variantes de servicios para redes IP, Fuente el autor

[1]Se denomina latencia a la suma de los retardos en la red. Los retardos están constituidos por el retardo de propagación, el retardo de transmisión (dependiente del tamaño del paquete), el retardo de procesamiento "store-and-forward" (los switch o enrutador emiten el paquete luego de haber sido recibido completamente en una memoria buffer) y el retardo de procesamiento (necesario para reconocimiento de encabezado, errores, direcciones, etc).

[2]Un tiempo de latencia variable se define como jitter (fluctuación de retardo) sobre los datos de recepción. La solución al jitter es guardar los datos en memorias buffer.

Caching

Existen tres técnicas que utilizan los enrutadores para mejorar la eficiencia de la red, reduciendo el tráfico que circulante:

- Manejo de nombres y direcciones mediante DNS,
- Servicios proxies (Elemento de la red que actúa en representación de otro) y,
- El cache local.

Un servidor proxy es confundido con un servidor cache, sin embargo hay diferencias. Un proxy es un intermediario entre el usuario e Internet. El proxy no necesariamente incluye una memoria cache. El Cache está localizado junto al enrutador y se utiliza para reducir la carga de tráfico hacia Internet.

Un cache es un bloque de memoria para mantener a mano los datos requeridos frecuentemente por varios procesos. Cuando un proceso requiere información, primero consulta el cache, si la información se encuentra allí se produce una mejora en el desempeño, reduciendo el retardo de procesamiento. Si no se encuentra en el cache se buscará en otras ubicaciones de memoria y luego estará disponible para una próxima consulta.

Una ventaja adicional de algunos cache, es la posibilidad de reducir el dialogo para transferencia de información. Puede reducirse la cantidad de paquetes transferidos mediante una sesión en paralelo de objetos.

Algunos tipos de memoria cache son:

- Cache del procesador: Es parte del procesador y de más fácil acceso que la memoria RAM y a una velocidad mayor
- Disco cache: pertenece a la memoria RAM y contiene información del disco. En algunos casos se mueve en forma anticipada la información desde el disco al cache en la RAM.
- Cache cliente-servidor: se trata de un banco de memoria ubicado en el cliente para agilizar el flujo de datos.
- Cache remoto: permite reducir los retardos cuando se accede la información de un sistema remoto en una WAN. Se resuelve mediante un caching del terminal remoto ubicado en el sistema local.
- Cache de servidor intermedio: entrega información a un grupo de clientes en un sistema cliente-servidor.

WEB-CACHING. El uso de cache en puntos de presencia, puede reducir el tráfico en la red (aumentando la velocidad de respuesta al usuario y el costo de la conexión WAN). El cache se conecta directamente al enrutador, el cual deriva todos los paquetes de requerimientos al cache, de esta forma puede verificar si la información está disponible. Su ventaja se incrementa en la medida que el número de usuarios es mayor.

Los componentes son los siguientes:

- La memoria cache o Cache Engine. El cache posee suficiente memoria (por ejemplo, 100 Gbytes), y capacidad de transacciones (algunos miles de sesiones TCP simultáneas).
- El enrutador o Home Enrutador. El cache se conecta directamente al enrutador de borde de la red, en la conexión hacia la Internet.
- Un enrutador puede poseer varios cache que se denominan “cache farm”. En este caso se forma una jerarquía entre cache para sucesivas investigaciones sobre el requerimiento del usuario.
- Un enrutador que administra el cache dialoga con la memoria mediante un protocolo WCCP (Web Cache Control Protocol). El cache puede trabajar también en modo Proxy sin el protocolo WCCP y dialogando con un navegador configurado en forma manual

CONTROL DE CONGESTIÓN EN EL BUFFER DE DATOS.	
FIFO First In, First Out	Este es el mecanismo de QoS por defecto en las redes IP. Es válido solo en redes con mínima congestión. No provee protección, no analiza el ancho de banda ni la posición en la cola de espera.
PQ Priority Queuing	Se basa en la prioridad de tráfico de varios niveles que puede aportar el encabezado del datagrama IP (ToS Type of Service). Se trata de 3 bits disponibles en el Byte 2 del encabezado de IPv4 (bits de precedencia).
CQ Custom Queuing	Se basa en garantizar el ancho de banda mediante una cola de espera programada. El operador reserva un espacio de buffer y una asignación temporal a cada tipo de servicio. Es una reservación estática.
WFQ Weighted Fair Queuing	Asigna una ponderación a cada flujo de forma que determina el orden de tránsito en la cola de paquetes. La ponderación se realiza mediante discriminadores disponibles en TCP/IP (dirección de origen y destino y tipo de protocolo en IP, número de Socket –port de TCP/UDP-) y por el ToS en el protocolo IP. La menor ponderación es servida primero. Con igual ponderación es transferido con prioridad el servicio de menor ancho de banda. El protocolo de reservación RSVP utiliza a WFQ para localizar espacios de buffer y garantizar el ancho de banda.

Herramientas para *Quality of Service* (QoS)

A. Manejo de congestión y tráfico

En la Tabla 2 se relacionan distintas herramientas para asegurar QoS en una red IP. Son mecanismos que previenen o manejan congestión, distribuyen tráfico o incrementan la eficiencia de la red.

CONTROL DE TRÁFICO.	
WRED Weighted Random Early Detection	Monitorea la carga de tráfico en algunas partes de la red y descarta paquetes en forma aleatoria si la congestión aumenta. Diseñado para aplicaciones TCP debido a la posibilidad de retransmisión. Esta pérdida en la red obliga a TCP a un control de flujo reduciendo la ventana e incrementándola luego en forma paulatina.
GTS Generic Traffic Shaping	Se usa para control del flujo de tramas en una interfaz en particular. Reduce el tráfico saliente limitando el ancho de banda de cada tráfico específico y enviándolo a una cola de espera. Así permite un mejor desempeño en topologías con tasa de bit diferentes.
INCREMENTO DE LA EFICIENCIA. SEÑALIZACIÓN.	
LFI Link Fragmentation and Interleaving	El tráfico interactivo como Telnet y VoIP es susceptible de sufrir latencia y jitter con grandes paquetes en la red o largas colas en enlaces de baja velocidad. Se basa en la fragmentación de datagramas y el intercalado de los paquetes de tráfico.
RSVP Resource Reservation Protocol	Implementa el concepto de Señalización. Dispone de dos tipos de señalización: en-banda (por ejemplo los bits de precedencia para ToS) y fuera-de-banda (mediante un protocolo de comunicación como RSVP). Este protocolo permite que un host o un enrutador asigne la reservación de ancho de banda a lo largo de la red IP.
RTP-HC Real-Time Protocol-Header Compression	La compresión del encabezado permite mejorar la eficiencia del enlace en paquetes de corta carga útil. Se trata de reducir los 40 bytes de RTP/UDP/IP a una fracción de 2 a 5 bytes, eliminando aquellos que se repiten en todos los datagramas.

Tabla 2. Herramientas para asegurar calidad de servicio en una red IP.
Fuente: Autor.

B. Priorización de tráfico.

ToS/IEEE 802.1Q. Define el VLAN Tagging Switch que permite identificar la VLAN posibilita la priorización del servicio. La trama del paquete incluye 4 Bytes adicionales que se colocan luego de las direcciones MAC y antes del Type/Length. Los 4 Bytes son indicados a continuación (Obsérvese que hay 3 bits para prioridad de tráfico y 12 bits para identificación de VLAN). Ver tabla 3.

TPID Tag Protocol Identifier	2 Bytes. Usados para identificación del protocolo.
TCI Tag Control Information	2 Bytes usados para las siguientes funciones: ➢ .UP 3 bits para prioridad del usuario (User Priority). Se trata de CoS de 0 a 7. ➢ .CFI (Canonical Format Indicator). 1 bit para ser usado por Token Ring. ➢ .VLANI (VLAN Identifier) 12 bits, permite identificar la VLAN (válido desde 1 a 1005). Permite la interoperación entre diferentes productores.

Tabla 3. Campos para manejo de Tipo de Servicio en IEEE 802.3
Fuente: Autor

QoS/RFC. Se trata de 3 bits que pueden ser usados para asignar prioridad. Se aplica un control de acceso extendido EACL para definir la política de la red en términos de congestión.

Con los bits de precedencia se pueden realizar 3 acciones: routing basado en políticas PBR (Policy-Based Routing) (por ejemplo direcciones IP, port de TCP, protocolo, tamaño de paquetes). Propagar la política de QoS mediante el protocolo de routing BGP-4 y desarrollar una política de tasa de acceso contratada CAR.

La CAR (Committed Access Rate) se ofrece especificando políticas de tráfico y ancho de banda. El umbral de CAR se aplica a la puerta de acceso para cada puerta IP o por flujo de aplicación individual. Algunas opciones de política de CAR son:

- *Política de prioridad:*

CAR máximo (el exceso de ancho de banda es descartado).
CAR premium (el exceso es señalado con un nivel de preferencia más bajo).

CAR best effort (por encima de un umbral se cambia la preferencia y sobre otro los paquetes son eliminados).

- *Política de asignación:*

CAR por aplicación (diferentes políticas son usadas en distintas aplicaciones).
CAR por puerto (los paquetes que ingresan por un puerto son clasificados con alto nivel de prioridad).
CAR por dirección (puede diferenciarse entre la dirección IP de origen y destino y asignar la prioridad en cada caso).

Requisitos de Telefonía IP

Ancho de Banda: En el entorno LAN, donde se utiliza tecnología Switch a 10 o 100 Mbs, se elige la compresión G711 con un ancho de banda de 84,7 Kb/s ya que se obtiene mayor calidad y se dispone suficiente ancho de banda. En el entorno WAN, donde es más escaso y costoso, se elige compresión G723 con ancho de banda de 27,2 Kb/s. Ver Tabla 4.

Codec de Audio	Ancho de banda comprimido	Ancho de banda paquetizada	Ancho de banda en Ethernet
G723	6.3 Kbps	17 Kbps	27.2 Kbps
G729	8.0 Kbps	24 Kbps	28.3 Kbps
G711	64 Kbps	74.6 Kbps	84.7 Kbps
FAX	4.8 Kbps	12.8 Kbps	28.4 Kbps

Tabla 4. Especificaciones de Ancho de Banda
Fuente: Autor

En el futuro y en cambio el ancho de banda en la WAN tiende a aumentar a la vez que los precios se reducen. Esto permitirá que cada vez sea más barato aumentar el número de comunicaciones de voz en la WAN.

El ancho de banda puede reducirse entre 30-40% cuando se utiliza la detección de silencios (VAD). Los Enrutadores pueden utilizar la compresión de cabeceras IP (cRTP) para reducir las cabeceras de 40 a 2-4 bytes, pudiendo reducir hasta 16,41 Kb/s en el caso de G723.

Porcentaje de Pérdida de Paquetes: Los equipos de la red como Enrutadores y Cortafuegos debido a la prioridad de flujo y a los picos de tráfico pueden perder paquetes y producir retardos en la transmisión (son retransmitidos), no obstante mientras en aplicaciones de datos no tienen impacto, en VoIP es crìtico, la pérdida de paquetes debe ser inferior al 5%.

Retardo: Tiempo de tránsito de los paquetes desde el origen al destino y de vuelta. Las personas son capaces de mantener una conversación cómodamente aunque exista cierto retardo, sin embargo llegado a un umbral puede empezar a ser incómodo para mantener una conversación, el retardo debe ser de 400ms (calidad media) y de 150ms (calidad alta).

Jitter: Variación del tiempo de tránsito de los paquetes. No todos los paquetes sufren un retardo constante, este retardo variable o Jitter, disminuye la calidad de la voz al pasar de cierto umbral, el jitter debe ser inferior a 50ms.

La calidad de la voz resultante depende de la combinación de estos tres últimos parámetros (Pérdida de Paquetes, Retardo y Jitter). Siendo los niveles de calidad los siguientes. Ver Tabla 5:

	Calidad Alta	Calidad Media	Calidad Baja
Pedida de Paquetes	1%	3%	5%
Retardo	150 ms	400 ms	600 ms
Jitter	20 ms	50 ms	75 ms

Tabla 5. Niveles de calidad
Fuente: Autor

Conclusión

Cuando se habla de Calidad de Servicio se hace referencia al efecto colectivo del rendimiento de la red, lo que determina el grado de satisfacción de los usuarios del servicio y está caracterizada por la combinación de aspectos tales como: soporte, operabilidad, seguridad y otros factores de cada servicio. Para lograr calidad, las técnicas

de QoS deben ser instaladas en todos los dispositivos de la red.

Por otra parte, para garantizar un correcto funcionamiento de una red IP, es importante una transmisión eficiente, donde la reserva de recursos sea mínima. Para esto, es posible usar arquitecturas Cliente-Servidor, que controlen el tamaño de la información almacenada en el cliente para minimizar este tiempo.

Reducir la cantidad de recursos necesarios para transmitir información sobre Internet, no es tarea sencilla y requiere conocer el extremo cliente, el servidor y la red, para transmitir los datos.

Resulta complejo analizar el estado de la red para determinar los parámetros óptimos, aquí es donde QoS permite potenciar el desempeño de los activos de red en la transmisión de los datos.

Referencias bibliográficas

[1] H. Schulzrinne, S. Casner, R. Frederick, V. Jacobson, "RTP: A Transport Protocol for Real-Time Applications", Request for Comments RFC 1889, January 1996.

[2] R. Braden, L. Zhang, S. Berson, S. Herzog, S. Jamin, "Resource ReSerVation Protocol (RSVP) – version 1 functional specification", Request for Comments RFC 2205, September 1997.