UNIVERSIDAD NACIONAL DE INGENIERÍA

FACULTAD DE INGENIERÍA ELÉCTRICA Y ELECTRÓNICA

POLÍTICAS DE CALIDAD DE SERVICIO
EN UNA RED IP

INFORME DE SUFICIENCIA

PARA OPTAR EL TÍTULO PROFESIONAL DE:

INGENIERO ELECTRÓNICO

PRESENTADO POR:

ARTURO FERNANDO MENA MAEKAWA

PROMOCIÓN
2003 - I

LIMA – PERÚ
2006

POLÍTICAS DE CALIDAD DE SERVICIO EN UNA RED IP

Dedico este trabajo

A mi esposa Carmen,

que es mi amor, mi apoyo y la que me motiva a ser mejor cada día,

Al bebé que esta creciendo en su vientre,

que es mi alegría, mi fuerza y mi pasión en todo,

A mis Padres, Fernando y Sara,

que me formaron y enseñaron a lograr mis metas,

Y a mis Hermanos, Kenjy y Esteban,

que son mi esperanza de superación.

SUMARIO

El presente trabajo pretende dar a conocer cómo se puede aplicar la calidad de servicio dentro de una red IP a fin de optimizar sus recursos para realizar una óptima trasferencia de paquetes.

En el capitulo I, se ofrece una visión general sobre la convergencia de redes y las motivaciones para implementar calidad de servicio en las redes de datos.

En el capitulo II nos centramos en los conceptos y parámetros de la calidad de servicio.

En el capitulo III se describen los modelos conceptuales (mejor esfuerzo, servicios integrados y servicios diferenciados) que se utilizan para implementar calidad de servicio.

En el capitulo IV se describe un análisis en bloques del sistema necesario para implementar calidad de servicio,

En el capitulo V, por último se muestra la implementación de los conceptos expuestos en los capítulos anteriores con un ejemplo de configuración en routers Cisco.

ÍNDICE

PROLOGO

Tanto la voz como el video necesitan de garantías en su transmisión a través de una red de datos, ya que no sólo se requiere transferir los datos de forma íntegra, sino que además se requiere que se transfieran en el tiempo adecuado y a un ritmo adecuado. Para esto la implantación de calidad de servicio hace posible ofrecer garantía y seguridad.

En este documento, trataremos sobre la calidad de servicio, específicamente en redes que manejan el protocolo IP.

CAPÍTULO I
EVOLUCION DE LOS SERVICIOS DENTRO DE UNA RED

1.1. Convergencia de redes

Comenzaremos este informe explicando el por qué de la convergencia de distintas redes de comunicaciones, ya que esto fue lo que motivó la creación de la calidad de servicio.

Si recordamos un esquema tradicional de comunicaciones podemos mencionar en primer lugar a la red de telefonía convencional en la que se realiza transmisión/recepción de voz a través de centrales públicas (también llamada Red de Telefonía Pública); las redes de datos, que comúnmente trabajan con el protocolo de internet (más comúnmente conocido como protocolo IP); o la red transmisión por cable que vendría a ser un tipo de red de transmisión de video; tal como lo muestra la figura 1.1

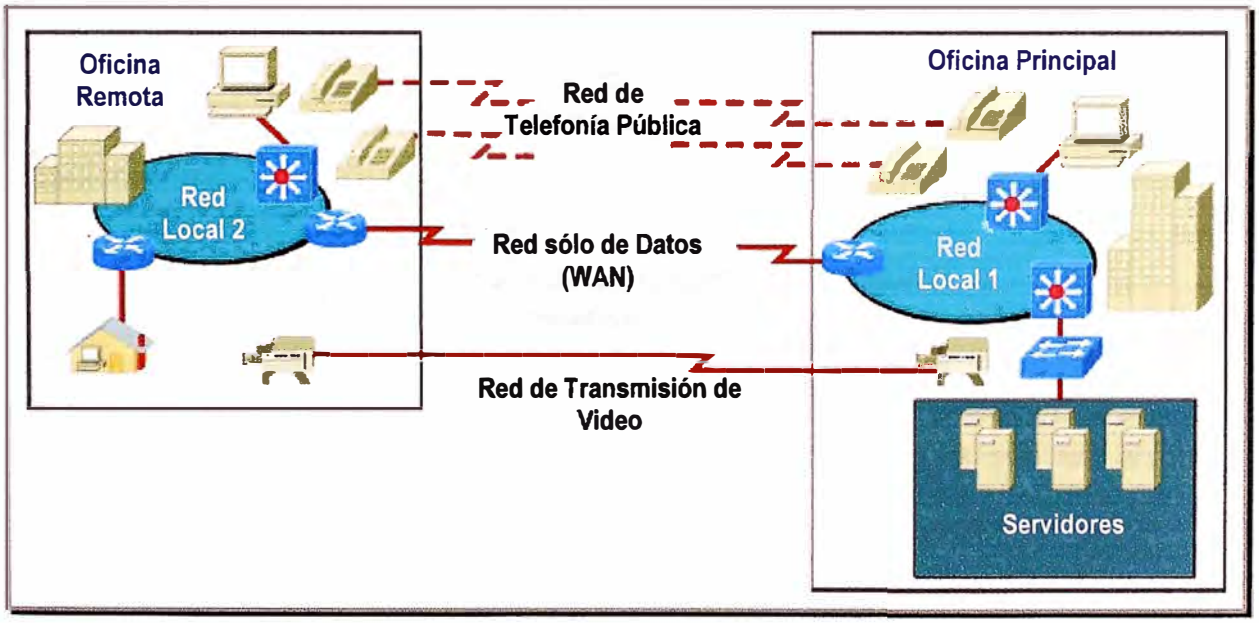

Fig. 1.1. Redes distintas, para distinto tráfico.

Lo que busca la convergencia es que la transmisión y recepción tanto de voz, datos y video, se realice a través de una misma red, la cual debe de garantizar que cada tipo de

tráfico tenga la importancia requerida para su paso de manera óptima a través de toda ella; el ejemplo unificado los mostramos en la figura 1.2.

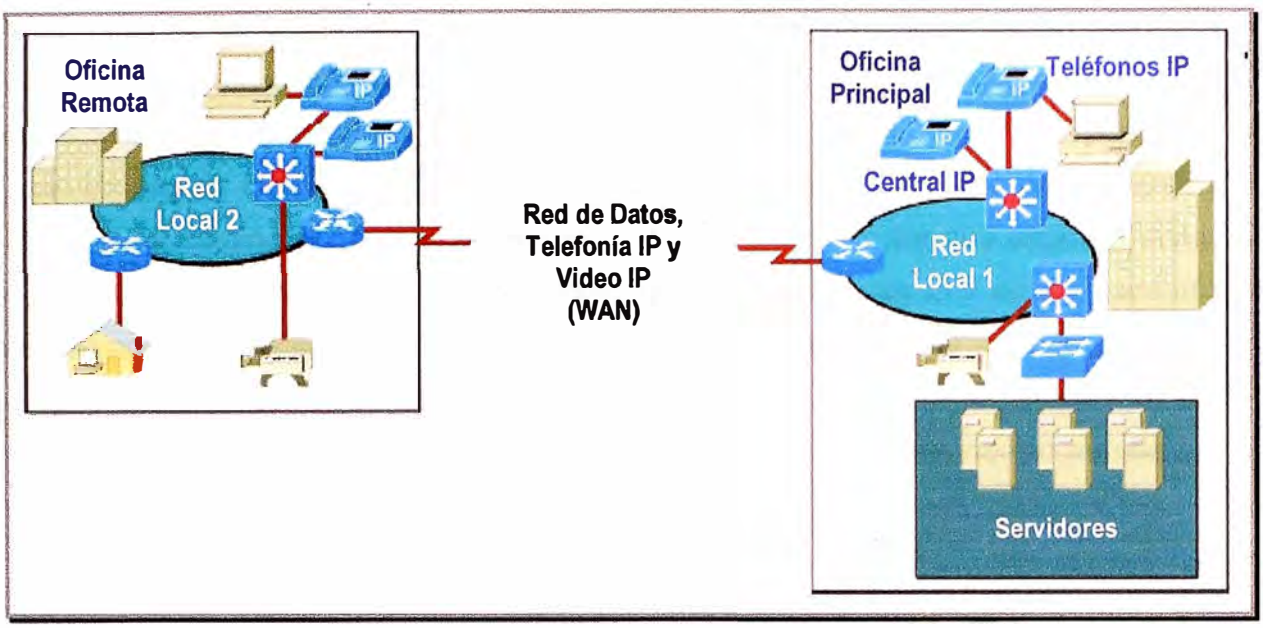

Fig. 1.2. Una misma red, para distintas aplicaciones.

Esta evolución está sustentada en una estrategia global de integración y uniformidad de servicios que unen los datos, telefonía, televisión, e Internet por medios de accesos integrados, ya que por ejemplo cada vez es mayor el número de usuarios residenciales de Internet, y las necesidades de los sistemas de información son cada vez mas especializadas. En la tabla 1.1 mostramos un ejemplo de la naturaleza de algunas aplicaciones.

Tabla Nº 1.1. Tipos de aplicaciones y su naturaleza.

Modo de Difusión	Aplicación	Naturaleza
Unidireccional	E-mail	Variable, no sensible al retardo
	Vídeo IP	Homogénea, en tiempo real.
Bidireccional	Telefonía IP	Interactiva, sensible al retardo.
Multidireccional	Audio/videoconferencia IP	Interactiva, muy sensibles a las pérdidas y al retardo.

Actualmente la red que reúne las mayores facilidades para transportar todo tipo de tráfico es la red de datos, ya que nos encontramos en una época con una "tendencia digital".

Ya que vamos a hablar de redes de datos mencionaremos que en un comienzo el protocolo IP fue diseñado bajo el esquema de Mejor Esfuerzo, en el que cada paquete IP comparte un mismo ancho de banda con otros y, por lo tanto compite con las transmisiones de los demás, es decir los paquetes llegan a su destino de la mejor forma posible, si el camino por donde tiene que ir está saturado tendrá que esperar hasta que se descongestione para poder pasar, como cuando realizamos una cola y sacamos un ticket el primero que llegue será atendido más rápido. Para mejorar este problema inicial fue que nació el concepto de Calidad de Servicio (al cual en adelante llamaremos QoS[1]), que busca la convivencia óptima de diversos tipos de aplicaciones dentro de una misma red.

Para efectos de este informe debemos aclarar que hemos tomado el protocolo IP, por ser éste el más generalizado en las redes de datos, no siendo el único valga recalcar.

1.2. Redes actuales

Actualmente no solamente existe la transmisión de Voz sobre IP, sino que también muchas empresas han tomado la opción de utilizar la Telefonía sobre IP que incorpora las mismas funciones de una central telefónica convencional, como conferencia, transferencia, llamada en espera, etc., pero manejada por un punto central que vendría a ser la central IP o Call Manager.

Viendo el lado comercial las empresas que tienen distintas oficinas en distintas zonas del país, contratan generalmente los servicios de interconexión a un operador de telecomunicaciones, y buscan cada también cada vez más calidad, nuevos servicios y nuevas tecnologías.

Telefónica del Perú, por ejemplo, brinda el servicio IP-VPN[2], que consiste en montar una VPN sobre una red con un núcleo IP-MPLS[3] que maneja calidad de servicio entre todos

---

[1] QoS: acrónimo de Quality of Service.

[2] "IP-VPN es un servicio de interconexión de redes locales sobre infraestructura perteneciente a Telefónica. Permite la creación de redes privadas virtuales ó VPN's, acrónimo de Virtual Private Network, sobre dicha

sus nodos, y de esta manera se les puede brindar a los clientes determinados caudales[4] de tráfico de acuerdo a su importancia, según la cual definimos tres tipos de tráfico:

- ✓ **Caudal Oro:** generalmente utilizado para el tráfico de voz.
- ✓ **Caudal Plata:** utilizado para el tráfico de data importante.
- ✓ **Caudal Bronce:** utilizado para el tráfico no sensible al retardo.

Por experiencia propia puedo decir que a muchos de los clientes se les instalan equipos que manejan voz sobre la red de datos, por ejemplo en el caso del fabricante Cisco, ofrece routers con tarjetas FXS, que son un tipo de tarjetas que se insertan en los routers para que de sus puertos cuelguen anexos analógicos en los puntos remotos, como los equipos Forma que la mayoría de nosotros tiene en casa, y en la oficina principal se les instala un central digital para que maneje las comunicaciones de voz.

En la figura 1.3 mostramos una topología de ejemplo.

---

infraestructura compartida manteniendo las mismas prestaciones que si fuera una red privada, reduciendo costos y aumentando el rendimiento.

El Servicio IP-VPN se implementa sobre la Red IP-MPLS de Telefónica que permite optimizar considerablemente la conmutación del tráfico IP, reduciendo los retardos de transporte. Proporciona a la red de capacidades de ingeniería de tráfico que proveen, por ejemplo, aspectos de Calidad de Servicio (QoS) y priorización de tráfico para las aplicaciones que el cliente defina como críticas." – Información obtenida de la página de Telefónica del Perú (www.telefonica.com.pe)

[3] MPLS, acrónimo de Multi Protocol Label Swiching, integra sin discontinuidades los niveles 2 (enlace de datos) y 3 (red), combinando de manera eficaz las funciones de control del routing con la simplicidad y rapidez de la conmutación de nivel 2, en sí combina lo mejor de cada nivel, en resumen la inteligencia del routing con la rapidez del switching.

MPLS usa, básicamente, un esquema de etiquetado del tráfico hacia delante, el tráfico es marcado en su entrada a la red pero no en los puntos de salida. Es independiente del protocolo utilizado (de ahí lo de multiprotocol), lo que permite que pueda ser utilizado sobre otros protocolos distintos a IP, como IPX, ATME, PPP, Ethernet, Frame Relay, sobre SONET y Token Ring.

[4] El caudal se refiere al ancho de banda contartado para cada tipo de tráfico según su sensibilidad al retardo.

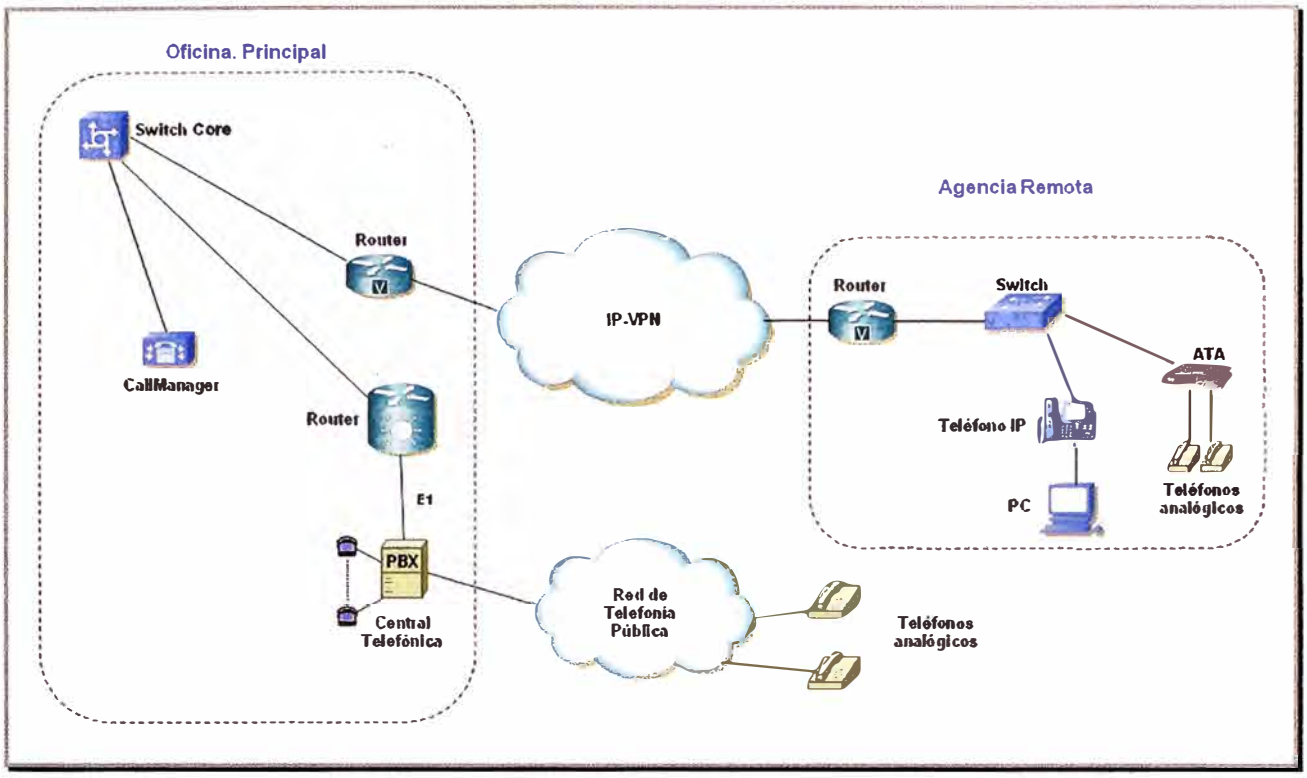

Fig. 1.3. Topología de una red de Telefonía IP

1.3. ¿Cuándo implementar QoS?

Pues bien cuestionemos algo ¿en que caso sería útil implementar QoS? En primera instancia la respuesta es simple, si tuviéramos suficiente ancho de banda en todos nuestros enlaces, no necesitaríamos utilizar QoS porque siempre habría espacio para el paso de más paquetes sea cual sea su importancia, tal como lo muestra la figura 1.4, en la que se muestra un enlace sobreprovisionado el cual obviamente sería mucho más costoso.

Fig. 1.4. Enlace que no necesitaría de QoS.

Un segundo caso sería un enlace que presenta una saturación moderada o por intervalos, en este caso utilizar QoS sería ideal, ya que en los momentos de saturación es

justo donde se aplica la optimización de recursos, y se trabajaría con un enlace con un precio acorde con nuestras necesidades, esto se muestra la figura 1.5.

Fig. 1.5. Enlace en el que QoS es útil.

Y como último caso tendríamos un enlace completamente saturado, que consume o requiere más recursos de los que posee y en el cual QoS daría alguna ayuda pero no sería la solución definitiva, ésta sería llegar al segundo escenario, el ejemplo lo tenemos en la figura 1.6, que mostramos a continuación.

Fig. 1.6. Enlace en el que incluso QoS es insuficiente.

De manera resumida tenemos dos opciones para trabajar cómodamente en los que llamamos interconexión:

- ✓ Podemos sobredimensionar adecuadamente nuestra red de transporte, lo cual implicaría aumentar costo, y recursos cuando resulte necesario; ó
- ✓ Podemos gestionar de forma inteligente los recursos disponibles, de acuerdo al tráfico que se cursa.

En una línea, podemos decir que la motivación por la que se trabajó QoS, fue para tener el manejo de la congestión[5].

---

[5] Congestión según el diccionario de la Real Academia de la Lengua Española es "la acción y efecto de obstruir o entorpecer el paso, la circulación o el movimiento de algo",

El manejo de congestión es un término general usado para nombrar los distintos tipos de estrategia de encolamiento que se utilizan para manejar situaciones donde la demanda de ancho de banda solicitada por las aplicaciones excede el ancho de banda total de la red, controlando la inyección de tráfico a la red, para que ciertos flujos tengan prioridad sobre otros. La figura 1.7 muestra de manera simple lo explicado anteriormente, la idea es dejar pasar los paquetes que mayor importancia tengan primero, en el caso del ejemplo los paquetes de voz.

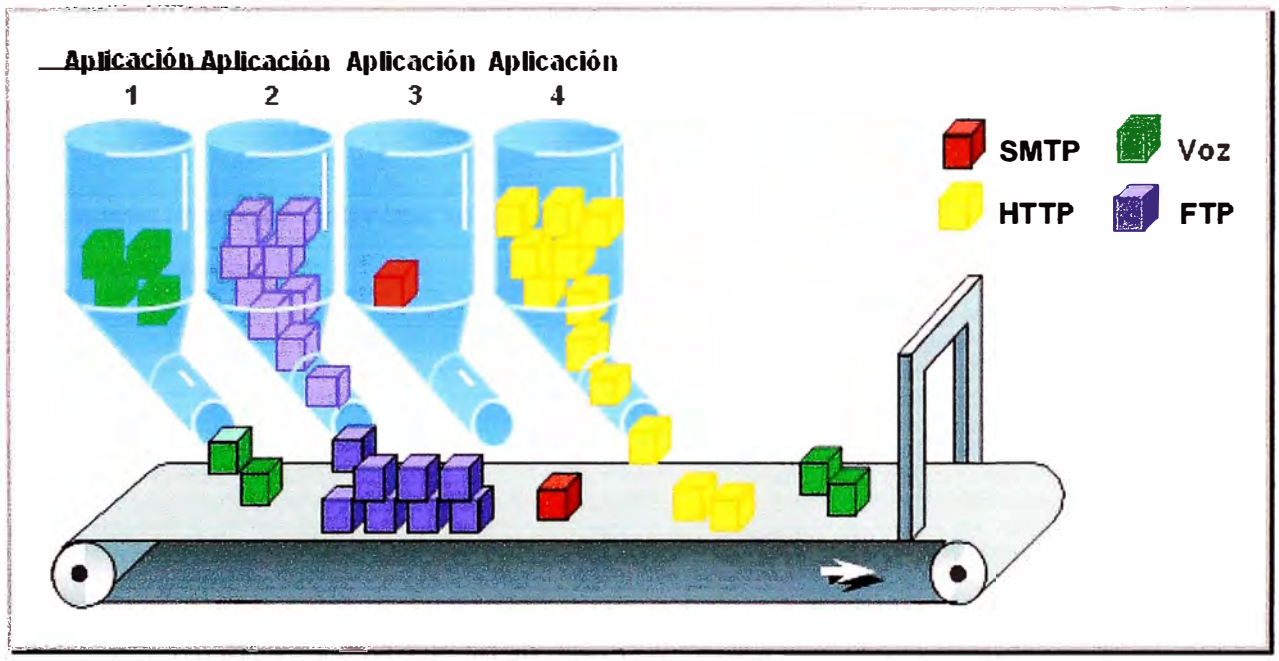

Fig. 1.7. Manejo de la congestión.

Para finalizar esta parte debo mencionar que no solo basta con tener suficiente ancho de banda, ya que también es necesario priorizar paquetes. Pongamos un ejemplo supongamos que un paquete de voz (que generalmente tiene un tamaño pequeño) llega a un router después de 5 paquetes de correo (que generalmente tienen un tamaño grande digamos el máximo de 1500 bytes), el paquete de voz aunque ha tenido espacio suficiente para su llegada al router tiene que esperar a que pasen 5 largos paquetes de correo para que recién sea atendido, lo cual provocaría un retardo considerable y por ende no se entendería el mensaje de voz.

En este caso QoS realiza el marcado de los paquetes para distinguir los tipos de servicios, los routers son configurados para manejar distintas colas, de acuerdo con las prioridades de las mismas, es decir una cola más prioritaria para la voz, otra para datos un poco menos prioritaria, etc.

Lo explicado anteriormente lo mostramos a manera de ejemplo en la figura 1.8.

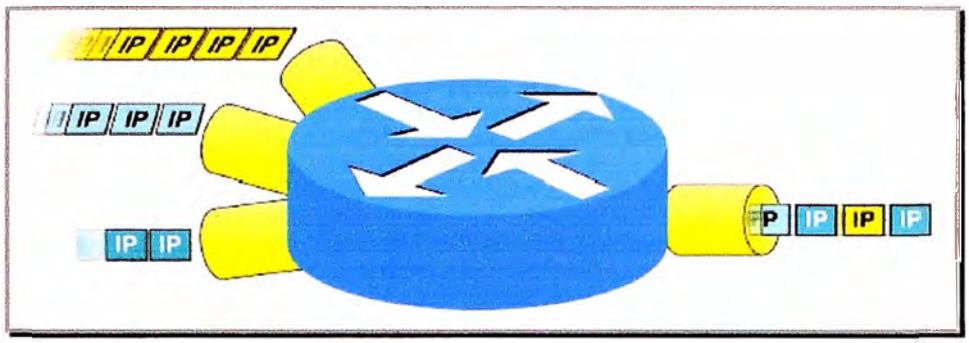

Fig. 1.8. El encolamiento y el manejo de la congestión son las claves de QoS.

En resumen hay que optimizar recursos de ancho de banda y dar la importancia debida a cada tipo de tráfico.

CAPÍTULO II
CONCEPTOS BÁSICOS

2.1. Concepto de QoS

De lo expuesto en el capítulo anterior, podríamos decir de manera concreta que QoS se entiende como: la capacidad de una red, o de un elemento de la misma, de asegurar, con un grado de fiabilidad preestablecido, que los requerimientos de tráfico, en términos de perfil y ancho de banda para un flujo de información dado, sean cumplidos.

Como dato adicional podemos anotar que según la RFC 2386, la calidad de servicio es el conjunto de requisitos del servicio que debe cumplir la red en el transporte de un flujo.

2.2. Diferencias entre QoS, CoS y ToS

Son varios los acrónimos terminados en "oS" que hacen referencia a la obtención de calidad de servicio en redes, provocando algunas equivocaciones.

2.2.1. Clase de Servicio

La Clase de Servicio, también llamada CoS[6] implica dos procedimientos: en primer lugar realizar la priorización de los distintos tipos de tráfico claramente definidos a través de la red y, en segundo lugar, la definición de un pequeño número de clases.

El priorizar es muy importante, de ello depende qué se hará con el tráfico y las clases vienen a ser cómo se agrupa este tráfico. CoS no garantiza anchos de banda.

A nivel de capa 2 del modelo OSI, también se le llama CoS a un campo de las tramas 802.1Q, lo cual detallaremos más adelante.

2.2.2. Tipo de Servicio

---

[6] CoS, es acrónimo de Class of Service.

El Tipo de Servicio, también llamado ToS[7] reserva anchos de banda con antelación y después se asigna el tráfico que necesite preferencia, de modo que este tráfico pueda utilizar el ancho de banda reservado.

A nivel de capa 3 del modelo OSI, también se le llama ToS a un campo de la cabecera de un paquete IPv4, que también detallaremos en un capítulo posterior.

2.3. Parámetros de QoS

Son varios los términos manejados en lo que refiere a QoS, pero básicamente se depende de cuatro parámetros: retardo, jitter, ancho de banda y pérdida de paquetes.

2.3.1. Retardo

Indica el retraso en la llegada de los flujos de datos a su destino, generalmente es expresada en milisegundos.

La figura 2.1 muestra el impacto de los retardo intermedios (tanto el retardo de propagación del paquete por cada enlace, y del retardo en el procesamiento en cada uno de los routers) por consiguiente, será desaconsejable todo aumento del tiempo de procesamiento en conexiones con tiempos de transmisión muy por debajo incluso de 150 ms, a menos que los beneficios obtenidos desde el punto de vista del servicio y de la aplicación sean evidentes.

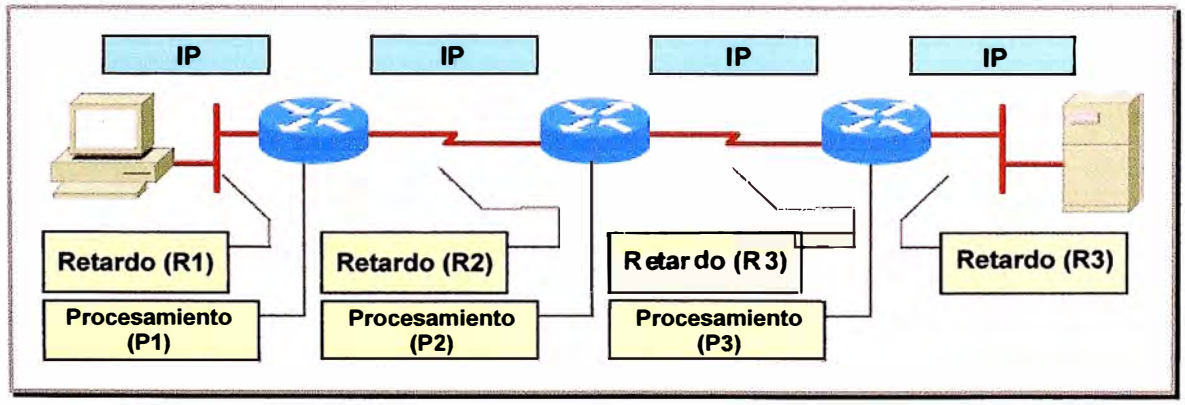

Fig. 2.1. Retardo de extremo a extremo.

[7] ToS, es acrónimo de Type of Service.

El retardo total de extremo a extremo vendría a estar dado por la suma de todos los retardos intermedios es decir, si utilizamos la variable D como el retardo total este se expresaría según la fórmula 2.1:

$$D = R1+P1+R2+P2+R3+P3+R4 \quad (2.1)$$

La tabla 1.2 que contiene las recomendaciones de la Unión Internacional de Telecomunicaciones sobre los límites de retardo admisibles

Tabla Nº 1.2. Tipos de aplicaciones y su naturaleza.

Retardo	Observaciones
De 0 a 150 ms	Aceptable para la mayoría de las aplicaciones de usuario.
De 150 a 400 ms	Aceptable siempre y cuando se conozca la influencia del tiempo de transmisión en la calidad de transmisión de las aplicaciones de usuario
Mayor a 400 ms	Inaceptable a efectos de planificación general de la red; se acepta, sin embargo, que este límite pueda ser rebasado en ciertos casos excepcionales.

2.3.2. Variación del retardo

También denominado jitter, es una distorsión de los tiempos de llegada de los paquetes recibidos, comparados con los tiempos de los paquetes transmitidos originalmente. Esta distorsión es particularmente perjudicial para el tráfico multimedia.

Digamos que el retardo de extremo a extremo del primer paquete de voz es D1 y del segundo es D2, el jitter expresado en la variable J según la fórmula 2.2 sería:

$$J = D1 - D2 = \Delta D \quad (2.2)$$

Es lo que ocurre cuando los paquetes transmitidos en una red no llegan a su destino en su debido orden o en la base de tiempo determinada, es decir varían en su retardo.

2.3.3. Ancho de banda

Es la medida de la capacidad de transmisión de datos, expresada generalmente en Kilobits por segundo (Kbps) o en Megabits por segundo (Mbps). Indica la capacidad máxima teórica de una conexión, pero esta capacidad teórica se ve disminuida por factores negativos tales como el retardo de transmisión, que pueden causar un deterioro en la calidad.

Aumentar el ancho de banda significa poder transmitir más datos (algo así como aumentar el número de carriles de una autopista), pero también implica un incremento económico y, en ocasiones, resulta imposible su ampliación sin cambiar de tecnología de red, obviamente no tienen la misma capacidad de transmisión una fibra óptica y un par de cobre.

Debemos tener en cuenta que inadecuados anchos de banda incrementan el retardo, ya que producen cuellos de botella tal como lo muestra la figura 2.2.

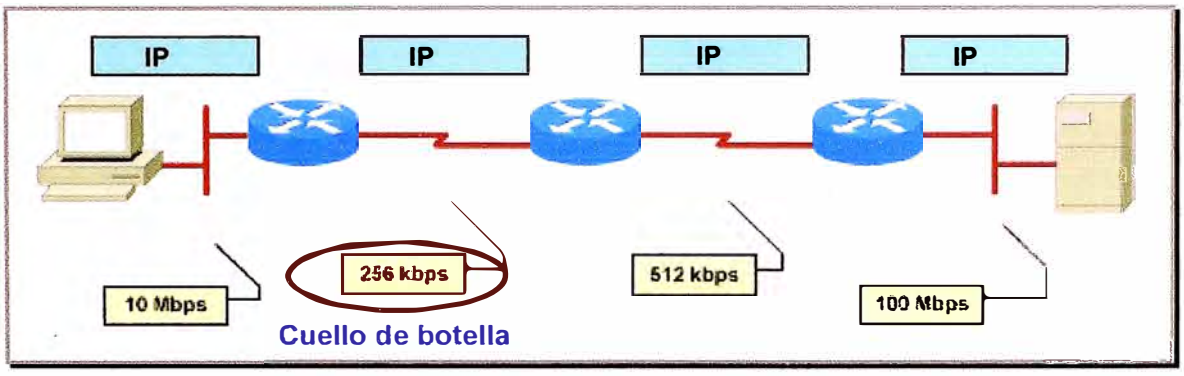

Fig. 2.2. Cuello de botella por ancho de banda insuficiente.

2.3.4. Pérdida de paquetes

Indica el número de paquetes perdidos durante la transmisión. Normalmente se mide en tanto por ciento. Por ejemplo: 1% o menos de media de pérdida de paquetes.

En redes congestionadas el ancho de banda insuficiente produce pérdida de paquetes debido al llenado de colas, ya que el los demás paquetes al no encontrar espacio se descartan, tal como lo muestra la figura 2.3.

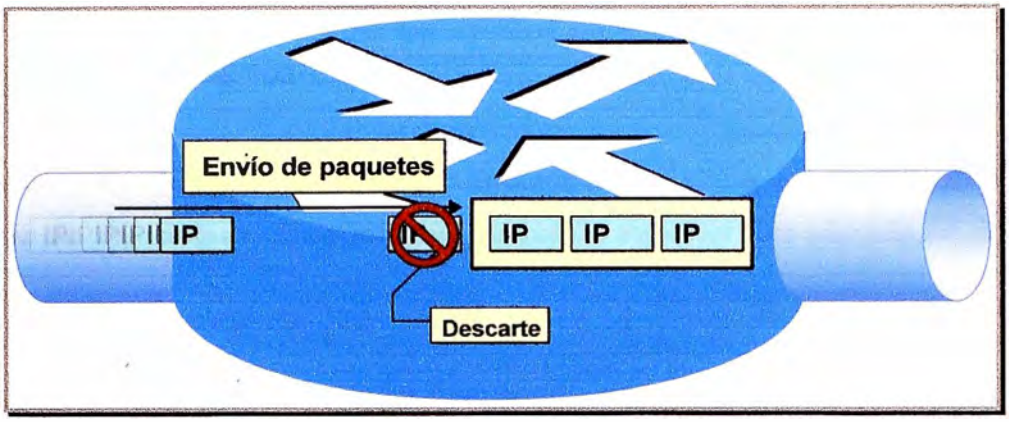

Fig. 2.3. Si no hay recursos suficientes los paquetes son desechados.

2.4. Requerimientos de QoS

Como ya hemos mencionado, las diferentes aplicaciones, protocolos y tipos de tráfico tienen diferentes requerimientos de QoS en función a la tolerancia de pérdida de paquetes, latencia o jitter que puedan tener. Reconocer estas diferencias es un elemento esencial en el diseño y configuración que ofrece QoS.

A continuación expondremos los requerimientos de QoS para el tráfico de datos, voz y video de manera separada.

2.4.1. Para el tráfico de datos

En si la data no tiene un tráfico plano en general es por ráfagas, pero estas no son bruscas sino suaves. En general no es tan sensible al retardo, ni a la pérdida de paquetes, ya que en su mayoría trabajan con TCP por lo que de descartarse algún paquete existe una retransmisión del mismo. La figura. 2.4 muestra la naturaleza del tráfico de datos.

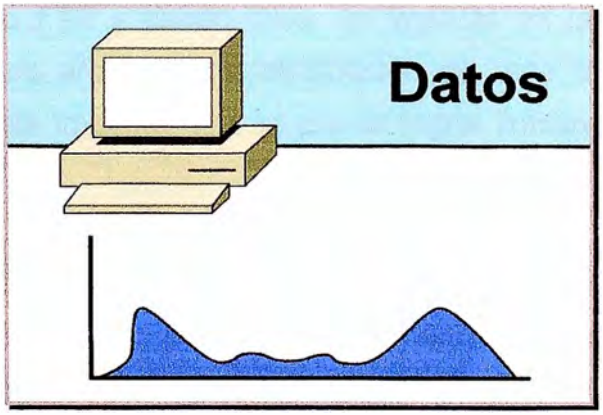

Fig. 2.4. El tráfico de datos es aleatorio, puede presentar picos.

2.4.2. Para el tráfico de voz

La pérdida de paquetes produce cortes o anulación total de la voz, el porcentaje máximo permitido es de 1%. Según los estándares se puede soportar una pérdida en una ventana de 30 ms.

Si el retardo extremo a extremo resulta demasiado largo, la conversación sonaría como la de un teléfono satelital. Según los estándares de la UIT sobre VoIP (especificación G.114) un retardo de hasta 150 ms es aceptable para tener una buena calidad en la voz. Con respecto al jitter es admisible una variación del retardo de hasta 30 ms.

La figura 2.5 muestra que el tráfico de voz es plano.

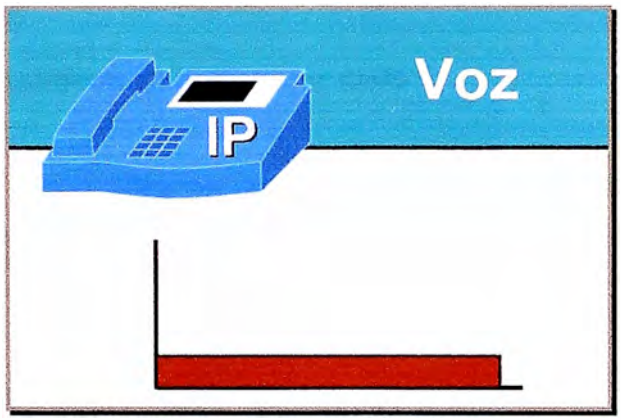

Fig. 2.5. El tráfico de voz es plano.

2.4.3. Para el tráfico de video

En el caso de la distribución de archivos de video en un solo sentido, debemos decir que es muy similar al tráfico FTP y puede tener un impacto en la performance de la red debido al tamaño de los archivos. La distribución de este tipo de tráfico debe ser manejada para evitar este impacto, ello se puede lograr limitando éste tráfico en horas pico.

En el caso de la videoconferencia, que es en ambos sentidos, los requerimientos básicos son similares a los de voz. La pérdida debe ser menor de 1%, la latencia en un sólo sentido no debe ser mayor a 150 ms y el promedio de jitter no debe ser mayor de 30 ms.

Debido a que su naturalaza es la transmisión en ráfagas, tal como lo muestra la figura 2.6, el mínimo de ancho de banda garantizado debe ser el tamaño de la sesión de videoconferencia aumentado en 20%, es decir, si tengo una sesión de videoconferencia a 384 Kbps se requiere un ancho de banda de 460 Kbps garantizado en el enlace.

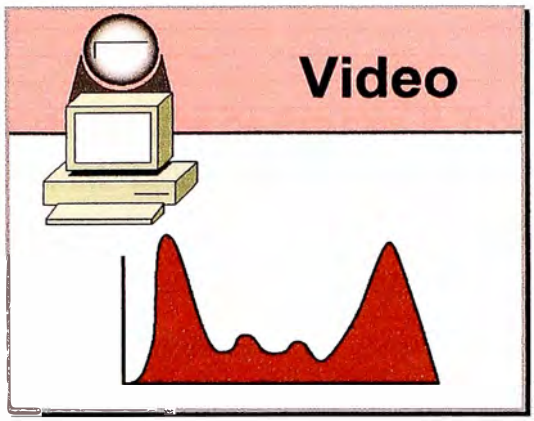

Fig. 2.6. El tráfico de video es por ráfagas.

2.5. Definición de políticas para trabajar QoS

En un sentido general, una política es una o más reglas que describen la acción que se produce cuando se da una determinada condición, pudiendo existir reglas concatenadas para ello.

La idea es definir reglas y condiciones, de acuerdo a ello tomar una acción determinada. La política define la funcionalidad requerida para la performance deseada en el acondicionamiento del tráfico, esa función depende del tipo de la tecnología escogida.

2.6. Planeamiento de políticas QoS

Existe un modelo de planeamiento de políticas denominado QPIM (QoS Policy Information Model), éste establece un estándar para especificar y representar el manejo, control y administración de los recursos de una red, éste modelo se encuentra descrito en la RFC 3644.

QPIM es usado para definir los lineamientos de los diferentes tipos de tráfico, de una manera estándar, Una política es mejor descrita usando reglas estructuradas como condición para ejecutar una acción, y todos los fabricantes de equipos de

comunicaciones siguen este modelo para implementar QoS en sus productos, he allí la importancia de conocerlo.

El modelo explica la necesidad de definir de clases para crear instancias tales que las acciones y condiciones de QoS puedan ser jerárquicamente organizadas en reglas y grupos.

En la figura 2.7 mostramos un diagrama de flujo para la implementación de una política de QoS según el modelo QPIM, la idea es definir el proceso para la generación de políticas, el proceso es dependiente de tres tipos de información, que deben ser conocidas y entendidas para su correcta especificación: la topología de la red, la metodología de QoS, y las reglas del negocio.

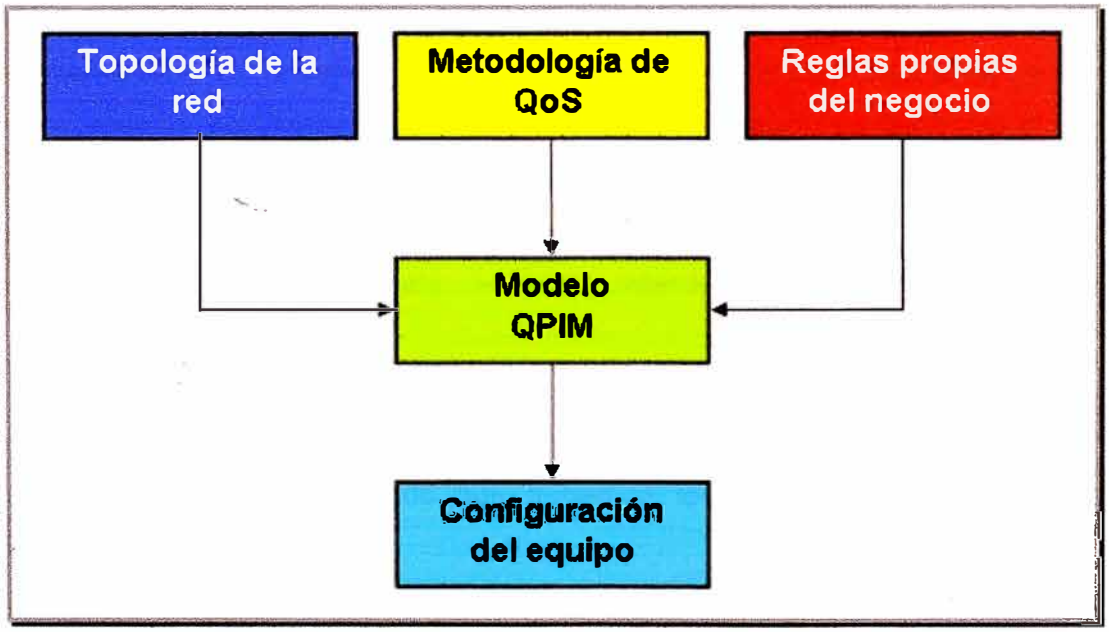

Fig. 2.7. Diagrama de flujo del modelo QPIM.

2.6.1. La topología de la red

En un escenario típico, el administrador de la red determina el papel de cada equipo dentro de la red, para identificar qué parte de la red va a brindar determinado tipo de acondicionamiento al tráfico. Es necesario realizar una auditoria a nuestra red usando herramientas de análisis, gestión u otras herramientas para saber que protocolos utilizamos y cuál es el volumen de tráfico cursado por cada enlace.

2.6.2. La metodología de QoS

Una vez identificados los recursos se planifica la metodología a utilizar para brindar QoS, con ello nos referimos al tipo de encolamiento, priorización de tráfico, separación de anchos de banda, etc.

2.6.3. Las reglas del negocio

Y todo va a depender de las reglas del negocio que no es otra cosa que la importancia de cada tipo de tráfico en la empresa que manejamos, esta decisión podría terminar siendo más difícil que una auditoria de red, porque como negocio la visión es un tanto subjetiva. Es necesario decidir el nivel de servicio que debe tener cada tipo de tráfico para lograr las metas del negocio.

Una vez conocidos los objetivos de la empresa y los tipos de tráfico éstos se deben agrupar en clases. Por ejemplo, si luego del análisis previo se tiene cinco diferentes aplicaciones de misión crítica, y requieren de las mismas características de QoS, podrían agruparse en la misma clase, de manera que los equipos de comunicación les puedan brindar los correctos niveles de ancho de banda, retardo, jitter y pérdida de paquetes.

2.7. Otros términos utilizados en QoS

Mencionaremos algunos conceptos que debemos tener claros para el desarrollo del presente informe:

Tráfico de red: Podríamos decir que es la data la atraviesa. Por ello es dependiente del tipo de aplicación que por ella circule, según esto podríamos establecer una diferenciación entre el tráfico:

a) Según tipo de aplicación: Tendremos tráfico habitual, multimedia, tiempo real, etc.

b) Según la sensibilidad al retardo: En este caso tendremos:

- ✓ **Tráfico algo sensible al retardo.** Ejemplos son los procesos de transacciones en línea, la entrada de datos remota y algunos protocolos como SNA. Este tipo de aplicaciones requieren retardos de un segundo o, incluso, menos. Retardos

mayores supondrían hacer esperar a los usuarios por la contestación a sus mensajes antes de que puedan continuar trabajando, disminuyendo así la productividad de los negocios.

- ✓ **Tráfico muy sensible al retardo.** El tráfico en tiempo real, como la videoconferencia y multimedia en tiempo real. Todos ellos requieren un retraso de tránsito muy pequeño (típicamente menos de una décima de segundo en un sentido, incluyendo el procesamiento en las estaciones finales) y un nivel de variación mínimo.
- ✓ **Tráfico muy sensible a las pérdidas.**
- ✓ **Tráfico nada sensible.**

Priorización: Consiste en la asignación de un determinado nivel al tráfico que circula por una red, asegurando así que las aplicaciones de mayor importancia sean atendidas con anterioridad a las de menor importancia, estando o no ante una situación de congestión. Es necesaria únicamente cuando la red no proporciona la suficiente capacidad para atender todo el tráfico presente en la misma.

Encolamiento: Consiste en dividir y organizar el tráfico ante un determinado dispositivo de red para su posterior retransmisión por la misma según un determinado algoritmo que define a la cola y que permite que determinados paquetes sean reexpedidos antes que otros. Es una de las herramientas más utilizadas por la QoS. La idea es ofrecer un mejor servicio al tráfico de alta prioridad al mismo tiempo que se asegura, en diferentes grados, el servicio para los paquetes de menor prioridad.

Los sistemas de colas, sin embargo, no garantizan que los datos importantes lleguen a su destino a tiempo cuando se produce congestión, lo único que aseguran es que los paquetes de alta prioridad llegarán antes que los de baja prioridad.

Las colas se suelen situar en los routers, siendo áreas de memoria o buffers dentro de los mismos.

Disponibilidad: Indica el nivel de utilización de los diferentes recursos. Suele especificarse en tanto por ciento y disminuye con cada caída de un enlace.

Rendimiento: Es definido también por algunos como la velocidad teórica de transmisión de los paquetes por la red. Esta depende directamente del ancho de banda y su variación ante las posibles situaciones de congestión de la red.

Planificación: Es el proceso de decidir qué paquetes enviar primero en un sistema de múltiples colas.

Flujo: Es el conjunto de datos pertenecientes a una misma secuencia que, debido a su gran tamaño, han de ser enviados mediante distintos paquetes. Tienen la misma dirección IP fuente y destino, el mismo puerto de destino y el mismo protocolo. El flujo, necesita, por tanto, llegar secuencialmente a su destino con una frecuencia constante.

CAPÍTULO III
MODELOS DE IMPLEMENTACIÓN

3.1. Modelos de implementación de QoS

Si bien existen diferentes modelos de implementación de QoS, todos ellos tienen en común la clasificación de flujos de tráfico.

La palabra clave en este punto es la diferenciación, debido a que antes de poder otorgar calidad de servicio a una aplicación en particular, es necesario clasificar el tráfico y determinar la forma en que será manejado a medida que circule por la red.

Durante los últimos años han surgido variados métodos para establecer QoS en equipos de red. Algoritmos avanzados de manejos de cola, modeladores de tráfico, y mecanismos de filtro como las listas de acceso, han hecho que el proceso de elegir una estrategia de QoS sea más delicado. Cada empresa puede tomar ventaja de distintos aspectos de su red en implementaciones de QoS para una obtener una mayor eficiencia.

Existen tres mecanismos para la implementación de QoS, estos son: Mejor Esfuerzo, Servicios Integrados, y Servicios Diferenciados

3.1.1. Mejor Esfuerzo

Se le llama servicio de Mejor Esfuerzo cuando se hace todo lo posible para intentar entregar un paquete a su destino, sin garantía de que esto ocurra. Una aplicación enviará datos en cualquier cantidad, cuando lo necesite, sin pedir permiso o notificar a la red. Obviamente, no es el modelo apropiado para aplicaciones sensibles al retardo, las cuales necesitan de un tratamiento especial.

3.1.2. Servicios Integrados[8]

El modelo de servicios integrados, también llamado IntServ[9] provee a las aplicaciones un nivel garantizado de servicio, negociando parámetros de red, de extremo a extremo. La aplicación solicita el nivel de servicio necesario para ella con el fin de operar apropiadamente, y se basa en la QoS para que se reserven los recursos de red necesarios antes de que la aplicación comience a operar. Estas reservaciones se mantienen en pie hasta que la aplicación termina o hasta que el ancho de banda requerido por ésta sobrepase el límite reservado para la aplicación.

El modelo de Servicios Integrados se basa en el protocolo de reservación de recursos RSVP[10] (especificado en la RFC 1633) para señalizar y reservar la QoS deseada para cada flujo en la red. Debido a que la información de estados para cada reservación necesita ser mantenida por cada router a lo largo de la ruta, es decir se asegura QoS extremo a extremo, lo cual implica tener que realizar reservas a lo largo de todo el camino, por lo tanto la escalabilidad para cientos de miles de flujos a través de una red se convierte en un problema.

RSVP es un protocolo de señalización de QoS, que posibilita:

- ✓ Dar a las aplicaciones un modo uniforme para solicitar determinado nivel de QoS.
- ✓ Encontrar una forma de garantizar cierto nivel de QoS, y
- ✓ Proveer autentificación.

RSVP ofrece dos tipos de servicios: de carga controlada y garantizado.

- ✓ **Servicio de carga controlada:** En este tipo de servicio la pérdida de paquetes debe ser muy baja o nula.

- ✓ **Servicio garantizado:** Se basa en solicitar anchos de banda y cierto retardo de tránsito máximo.

---

[8] Para esta parte se tomó como base tanto para el texto y las imágenes el artículo "Señalización para QoS en redes IP" de Manuel Moreno Martín publicado por la revista ACHIET No. 67.

[9] IntServ: acrónimo de Integrated Services

[10] RSVP: acrónimo de ReSerVation Protocol

De los dos tipos de servicios que RSVP soporta, el más adecuado para aplicaciones con requerimientos de tiempo real es el servicio garantizado, aunque es más complejo de implementar que el servicio de carga controlada.

El RSVP contempla la reserva de recursos en la red para cada flujo de información de usuario, así como el mantenimiento de un estado para cada flujo, esto es un manejo de tablas para los estados de reserva. Esto conduce a un considerable tráfico de señalización y ocupación de recursos en cada router para cada flujo, con la consiguiente complejidad en el hardware, al margen del aporte que esta señalización hace a la congestión de la red.

No es una solución escalable, no es una solución adecuada para grandes entornos como Internet, aunque si lo es para entornos más limitados y también para redes de acceso a un backbone.

RSVP define dos sentidos para la transferencia de sus mensajes de señalización, de bajada y de subida. El flujo de bajada se efectúa desde la fuente al receptor o receptores, y el flujo de subida en sentido contrario.

PATH y RESV son dos mensajes básicos del protocolo RSVP, y son en definitiva los mensajes a través de los cuales se lleva a cabo la reserva de recursos en la red previo a la comunicación.

Los mensajes PAHT´s son generados por la fuente de mensajes de usuario necesitados de garantía de QoS, e indica las características de éstos en cuanto a recursos que necesita. La ruta que deben seguir estos mensajes es la misma que siguen los datos de usuario, para lo cual se requiere previamente un diálogo entre el proceso RSVP y el proceso de ruteo, pues dicha ruta quien la determina es el protocolo de ruteo, de lo contrario para nada serviría RSVP.

En su paso por cada router RSVP los mensajes PATH´s se actualizan y se retransmiten, imprimiendo la dirección IP del router que lo actualiza y re-envía. Cada router RSVP también almacena la dirección del router anterior. Así, con los mensajes PATH´s se posibilita indicar al receptor, o receptores, no solo las características del tráfico de usuario, sino también la ruta por donde debe solicitar las correspondientes reservas de

recursos. Los routers que no soporten RSVP transfieren transparentemente los mensajes PATH´s.

Los mensajes RESV´s son producidos por el receptor (o receptores) de los flujos de información de usuario, como respuesta a los mensajes PATH´s, y solicitan a la red (a los routers RSVP) las correspondientes reservas de recursos para soportar la comunicación con cierta QoS, fluyendo hasta la fuente del stream de datos de usuario, es decir, en sentido de subida. Con la información de ruta que suministran previamente los mensajes PATH´s, los mensajes RESV´s dirigen las solicitudes de reservas a los routers RSVP apropiados, esto es, por donde fluirán los streams de datos.

Los mensajes RESV´s especifican el ancho de banda mínimo que se requiere para obtener determinada demora en un stream de datos específico. Vale decir además, que es posible efectuar reservas compartidas, esto es, una misma reserva aplicable a varios streams de datos de usuario.

Estas reservas de recursos en los routers RSVP de la red se materializan mediante estados en dichos routers, que requieren refrescarse periódicamente, por lo que durante toda la comunicación se necesita señalizar para mantener las reservas previamente efectuadas. En consecuencia, esto conlleva a cierta señalización permanente durante la fase de transferencia de información de usuario, con la consiguiente carga de tráfico que implica.

Vale decir también que la reserva de recursos extremo a extremo que posibilita RSVP será válida si, y solo si, la congestión y demora que introduzcan los routers no RSVP no es significativa.

En la figura 3.1 se muestra de forma muy simplificada el intercambio básico de mensajes RSVP, específicamente mensajes PATH´s y RESV´s entre un emisor y dos receptores (A y B), indicándose que la reserva representada por el mensaje RESV 2 prevalece sobre la reserva representada por el mensaje RESV1, de manera que esto sugiere que la reserva solicitada por el receptor A es mayor que la solicitada por el receptor B. Esto es, la reserva mayor prevalece sobre la reserva menor, así el router B sólo solicita al router A la mayor de las dos solicitudes de reservas a él llegadas desde el router C (originada por el receptor A) y desde el receptor B. Esto es una característica de RSVP.

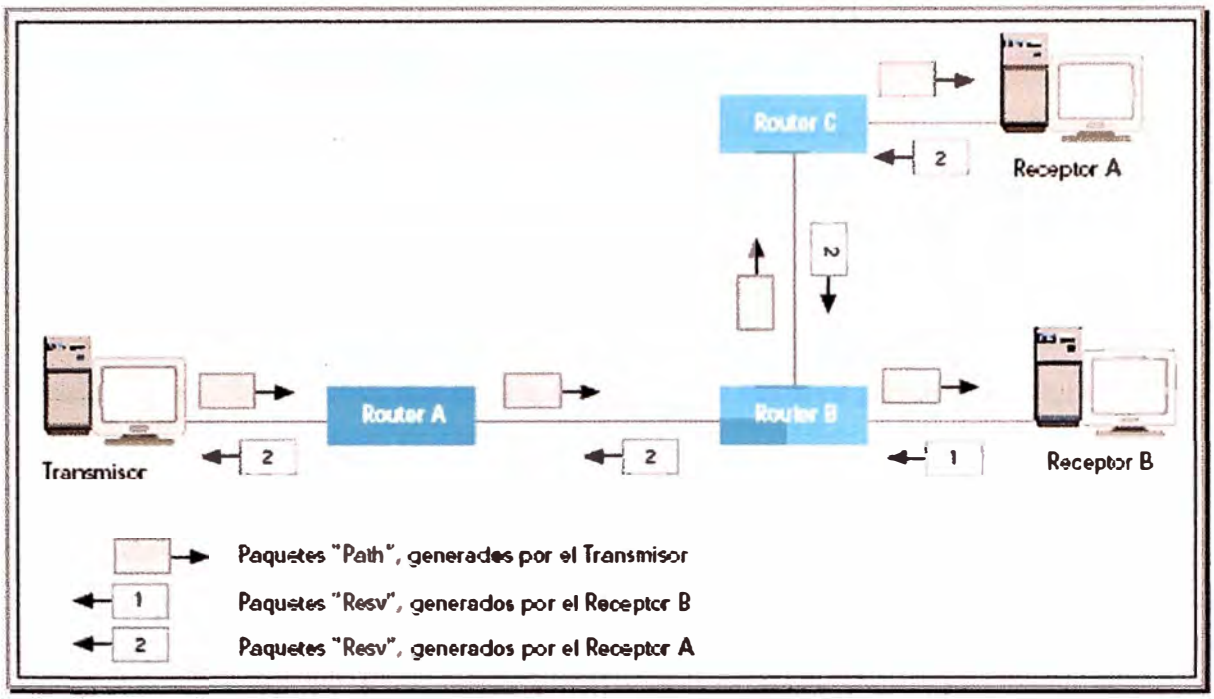

Fig. 3.1. Intercambio básico de mensajes RSVP.

Otros mensajes del protocolo RSVP son:

PATHTEAR: son mensajes generados por la fuente de datos de usuario para eliminar los estados PATH´s en todos los routers RSVP. Siguen la misma ruta que los mensajes PATH´s. También pueden ser originados por cualquier nodo cuando se agota el tiempo de espera del estado PATH.

RESVTEAR: son generados por los receptores para borrar los estados de reserva en los routers RSVP, por tanto viajan en el sentido de subida. Pueden ser también originados por nodos RSVP al agotarse el tiempo de espera del estado de reserva de los mismos.

PATHERR: viajan en sentido de subida hacia el emisor siguiendo la misma ruta que los mensajes PATH´s, y notifican errores en el procesamiento de mensajes PATH´s, pero no modifican el estado del nodo por donde ellos pasan en su viaje hacia la aplicación emisora.

RESVERR: notifican errores en el procesamiento de mensajes RESV, o notifican la interrupción de una reserva. Se transfieren en la dirección de bajada hacia el receptor o receptores apropiados.

Estas solicitudes de reserva conducen a que en cada router RSVP se establezca un estado, es decir, una reserva en cada router es un estado con un determinado tiempo de espera, que debe ser refrescada periódicamente por los receptores, de lo contrario vence el tiempo de espera y se deshace la correspondiente reserva, con la consecuente generación de un mensaje RESVTEAR.

La liberación de recursos reservados mediante RSVP se puede materializar de diferentes maneras, así la solicitud para dar baja a determinada reserva puede ser originada: por el emisor, por el receptor, o por un nodo de la red.

Por parte del emisor o de un receptor acontece cuando así lo decide la aplicación correspondiente, en cuyo caso esto se produce mediante la generación de un mensaje PATHTEAR o un mensaje RESVTEAR, respectivamente.

Por parte de un nodo se lleva a cabo cuando vence el tiempo de espera correspondiente del estado camino o del estado de reserva, lo que origina la emisión de un mensaje PATHTEAR o un mensaje RESVTEAR, respectivamente.

3.1.3. Servicios Diferenciados

El modelo de Servicios Diferenciados, también llamado DiffServ[11], es una propuesta en la cual los paquetes son marcados de acuerdo a las clases de servicios predeterminados. Este modelo incluye un conjunto de herramientas de clasificación y mecanismos de cola que proveen dar a ciertas aplicaciones prioridades sobre el resto del tráfico en la red.

Los Servicios Diferenciados se basan en marcar los paquetes IP, para que la red en base a esa marca desarrolle un tratamiento diferenciado de los paquetes. Esta diferenciación no es la misma en los diferentes nodos, sino depende de si se trata de un nodo interior o un nodo frontera. En consecuencia, y a diferencia de la solución Servicios Integrados (basada en RSVP), la red con nodos DiffServ no establece ni mantiene estados de las conexiones por flujos de paquetes, por lo que se le considera una solución escalable y puede ser fácilmente implementada en las redes IP existentes.

Se han definido dos tipos de Servicios Diferenciados:

---

[11] DiffServ: acrónimo de Differentiated Services

- ✓ **Envío explícito:** También denominado EF[12], equivale a una línea arrendada virtual, por lo que se garantiza cierto ancho de banda y reducida demora de cola. Emula un circuito.

- ✓ **Envío asegurado;** También denominado AF[13], en el que los paquetes se etiquetan con alta prioridad, aunque no se garantiza un ancho de banda. Se posibilita una QoS superior al servicio tradicional de Mejor Esfuerzo. Brinda cuatro clases de servicios, cada una con tres niveles diferentes de descarte de paquetes.

La segunda opción, es la más utilizada en la arquitectura de Servicios Diferenciados. Los cuatro grupos AF, llamados clase AF1, AF2, AF3 y AF4 son divididos en 3 sub-grupos alta, media y baja, representando la tendencia a descartar paquetes. Cada paquete será entregado a una clase de servicio mientras se apegue a un perfil de tráfico. Cualquier exceso de tráfico será aceptado por la red, pero tendrá mayor probabilidad de ser descartado según la clase de servicio y grupo. Cada nodo deberá implementar alguna forma de reservación de ancho de banda para cada clase AF, y alguna forma de otorgar prioridad para permitir políticas de esta índole.

Los tipos de routers en redes de Servicios Diferenciados se clasifican de la siguiente manera:

- ➢ **Router de primer salto:** Es el router más próximo al emisor de paquetes. Es responsable de que el tráfico esté acorde con el ancho de banda del perfil.

- ➢ **Router de ingreso:** Se sitúa en los puntos de entrada al de redes, efectuando la clasificación de los paquetes.

- ➢ **Router de egreso:** Se ubican en los puntos de salida, controlando el tráfico.

- ➢ **Router interior:** Tiene la misión de sumar flujos, realizar la clasificación y el re-envío de paquetes. Se sitúan dentro de la red.

---

[12] EF: acrónimo de Expedited Forwarding.

[13] AF: acrónimo de Assured Forwarding

Esta arquitectura permite rendir mucho mejor en ambientes de bajo ancho de banda, y provee de un mayor potencial que una arquitectura de Servicios Integrados.

CAPÍTULO IV
BLOQUES DE CONSTRUCCIÓN

4.1. Esquematización en bloques

En el siguiente capítulo trataremos de mostrar de manera esquematizada como se logra implementar la calidad de servicio, partiendo de varios bloques funcionales, que están representados en la figura 4.1, y que describiremos a continuación.

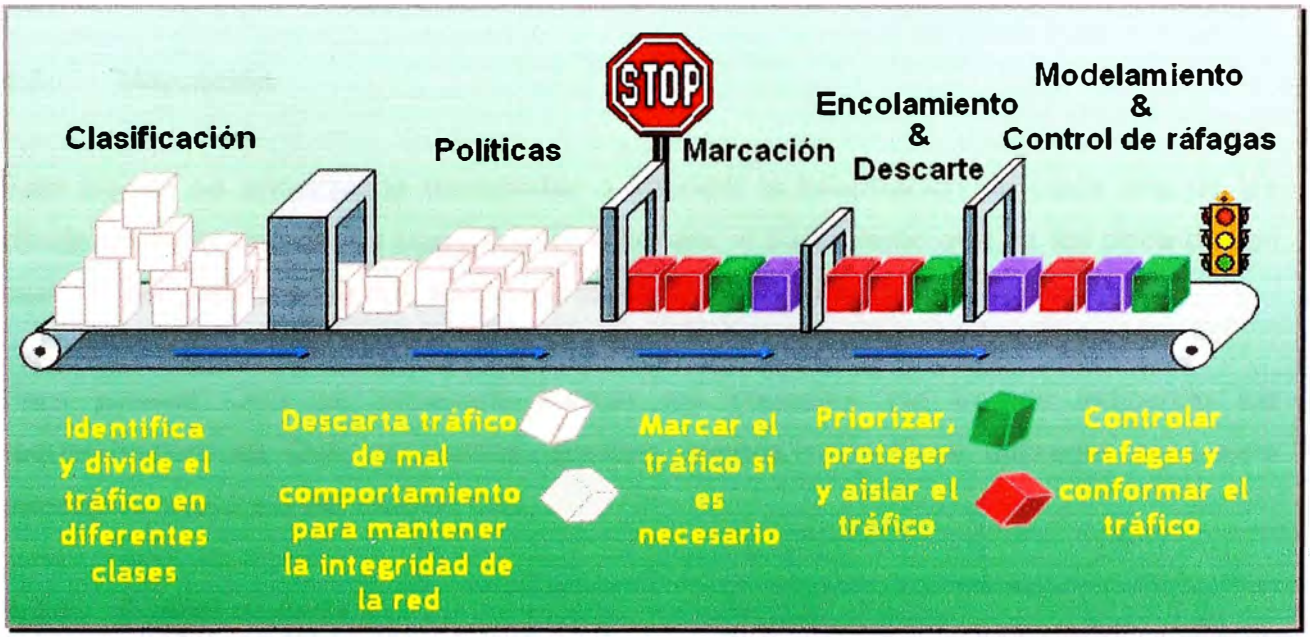

Fig. 4.1. Bloques de construcción QoS.

4.2. Clasificación[14]

Este bloque identifica cuales son los grupos de paquetes que deben recibir un servicio particular, y los separa en clases.

---

[14] Para esta parte se ha tomado como referencia básicamente dos documentos: "Introduction to QoS" de Irene Owen y el "Curso BSCI de Cisco", Semestre 3: Multilayer Switching v3.0, Cap. 8.

El criterio de clasificación puede variar en realidad podrían crearse listas de acceso en los routers de acuerdo a su dirección IP origen, IP destino o los puertos utilizados, el valor de la precedencia IP, el valor de CoS, etc. La figura 4.2 muestra un ejemplo de lo anteriormente mencionado.

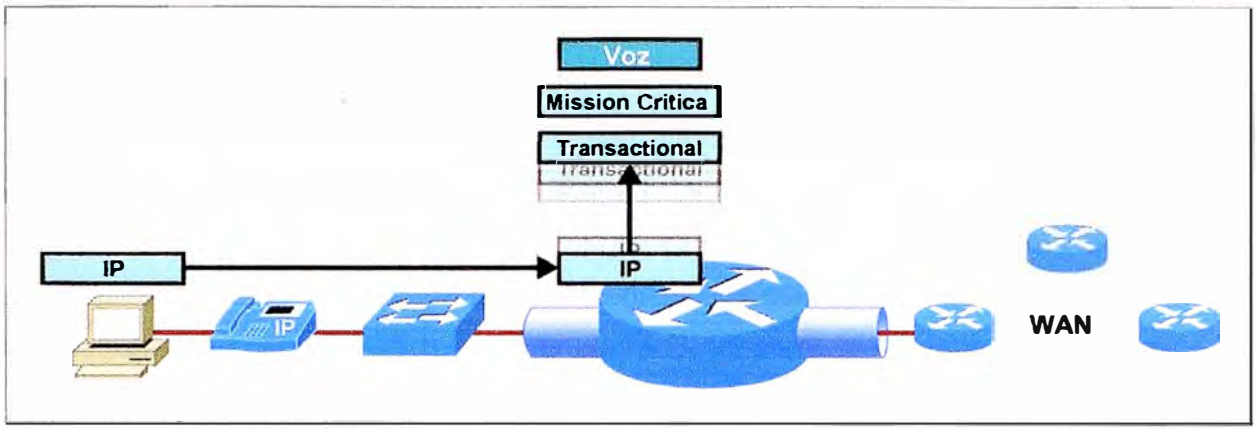

Fig. 4.2. Los routers clasifican el tráfico.

4.3. Marcación

Este bloque se encarga de transportar o imprimir la información de cada una de las clases en cada uno de los paquetes IP, ello para el tratamiento que se les deba dar en caso de congestión en la red, es decir a efectos de dar prioridad o desecharlos.

Para proveer QoS es necesario marcar los paquetes y/o tramas indicando los requerimientos de QoS necesarios. Existen campos específicos utilizados para este marcaje a nivel 2 y 3 del modelo OSI.

4.3.1. A nivel de capa 2

Las tramas Ethernet no tienen campos significativos de QoS, si es requerido a este nivel el protocolo 802.1Q puede utilizar el campo CoS para brindar esta facilidad.

A nivel de capa 2, se utilizan los 3 bits del campo CoS. Ello permite un rango de CoS entre 0 y 7, de menor a mayor prioridad. En la figura 4.3 por el campo Tag en una trama 802.1Q.

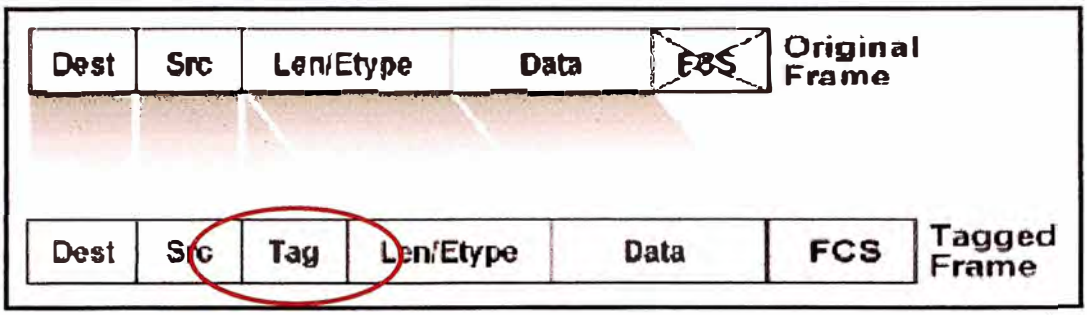

Fig. 4.3. Trama Ethernet vs. 802.1Q.

Si un equipo periférico tal como un teléfono IP o alguna aplicación de PC, es capaz de setear el valor CoS, el diseñador de la red deberá decidir si confía en esta información o no, ya que puede ser remarcada con un valor por defecto si no se especifica, trabajando valga la salvedad con un switch que maneje QoS, tal como lo muestra la figura 4.4.

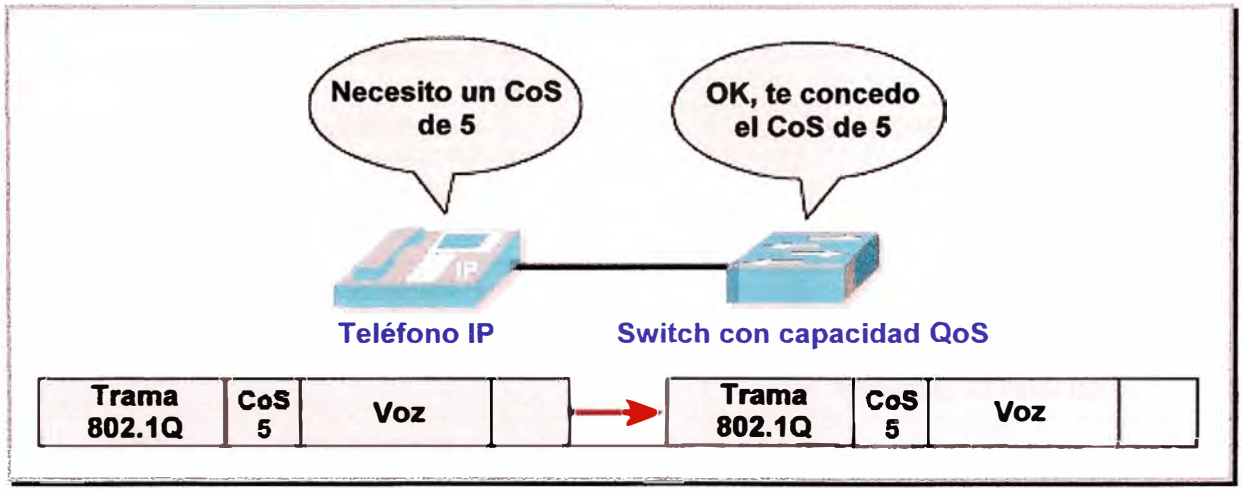

Fig. 4.4. CoS en equipos periféricos.

4.3.2. QoS en IPv4[15]

Originalmente, para el protocolo IPv4 se diseñó el campo ToS (Type of Service) para realizar el marcado de paquetes con un nivel de servicio requerido. En la figura 4.5 mostramos cómo se encuentra formada la cabecera de un paquete IPv4.

Fig. 4.5. La cabecera del paquete IPv4.

[15] Para esta parte se tomó como base el documento "Introducción a QoS" de Irene Owen, Cisco Systems.

Este campo puede ser utilizado de dos formas para la precedencia IP o en todo caso para DSCP, expondremos a continuación ambos tipos de marcado.

4.3.2.1. Precedencia IP

La precedencia IP consiste en utilizar los tres primeros bits del octeto ToS, y en base a su variación se toman las precedencias, de acuerdo a lo especificado por la RFC 1122.

Si utilizamos todas las variaciones posibles tendríamos ocho opciones, dándosele la prioridad respectiva en orden de mayor a menor las mostramos en la tabla 4.1.

Tabla Nº 4.1. Precedencias establecidas.

Valor Binario	Valor	Nombre	Tipo de tráfico
111	Precedencia 7	Network	Muy sensible al retardo
110	Precedencia 6	Internet	
101	Precedencia 5	Critical	
100	Precedencia 4	Flash Override	Sensible al retardo
011	Precedencia 3	Flash	
010	Precedencia 2	Inmediate	
001	Precedencia 1	Priority	
000	Precedencia 0	Routine	Nada sensible al retardo

En el caso del marcaje en capa 3, la precedencia IP y el valor DSCP, no pueden ser dadas simultáneamente, y de ser el caso prevalece DSCP.

4.3.2.2. DSCP

Más tarde cuando surgió el modelo de Servicios Diferenciados este mismo byte se re-utilizó como el campo DSCP (Differentiated Services Code Point), que tuvo mayor aceptación global y se asumió una interpretación estándar que permitió a las redes planificar metodologías

Se implementa de la siguiente manera, de los ocho bits se tiene los dos últimos reservados y con los otros seis bits es posible obtener 64 combinaciones, de ellas según lo especificado por la RFC 2474, 48 son utilizadas de manera global y 16 son para uso local.

En la figura 4.6 mostramos la conformación del campo DSCP en comparación con la precedencia IP.

Fig. 4.6. La precedencia IP vs. DSCP

Como información adicional presentamos la tabla 4.2 con los valores estandarizados para DSCP.

Tabla Nº 4.2. Valores estandarizados para DSCP.

6 primeros bits DSCP	Uso
111110	Reservado (Routing y control)
111100	Reservado (Routing y control)
111010	Reservado (Routing y control)
111000	Reservado (Routing y control)
110110	Reservado (Routing y control)
110100	Reservado (Routing y control)
110010	Reservado (Routing y control)
110000	Reservado (Routing y control)
101110	Premium
101100	Configurable por el usuario

101010	Configurable por el usuario
101000	Configurable por el usuario
100110	Seguro clase 4 precedencia alta
100100	Seguro clase 4 precedencia media
100010	Seguro clase 4 precedencia baja
100000	Configurable por el usuario
011110	Seguro clase 3 precedencia alta
011100	Seguro clase 3 precedencia media
011010	Seguro clase 3 precedencia baja
011000	Configurable por el usuario
010110	Seguro clase 2 precedencia alta
010100	Seguro clase 2 precedencia media
010010	Seguro clase 2 precedencia baja
010000	Configurable por el usuario
001110	Seguro clase 1 precedencia alta
001100	Seguro clase 1 precedencia media
001010	Seguro clase 1 precedencia baja
001000	Configurable por el usuario
000110	Configurable por el usuario
000100	Configurable por el usuario
000010	Configurable por el usuario
000000	Best Effort (default)

Es obvio que al tener más bits DSCP provee más opciones que la precedencia IP, pero algunos equipos no soportan DSCP por lo que en estos casos conviene la utilización de la precedencia IP.

4.3.3. QoS en IPv6

Tal fue el éxito de DSCP, que fue incluida para ofrecer las mismas ventajas en la versión 6 del protocolo IP, en el denominado campo Clase de Tráfico, que consta de 8 bits igualmente. En la figura 4.7 mostramos el formato de la cabecera de un paquete IPv6.

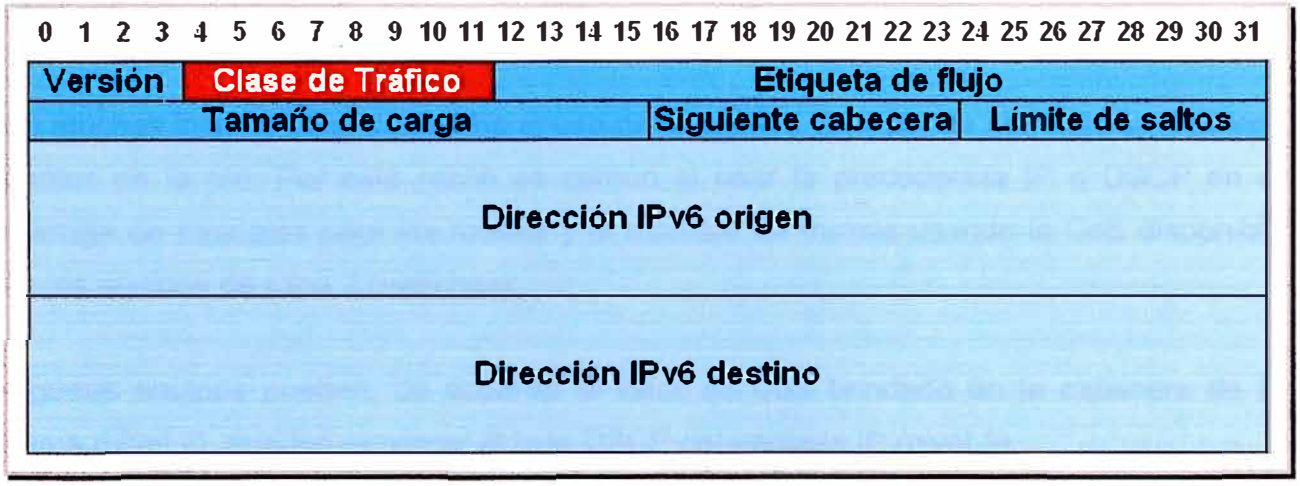

Fig. 4.7. El byte Clase de tráfico en la cabecera del paquete IPv6.

4.3.4. Relación entre nivel 2 y 3

La marcación del tráfico a nivel de capas 2 y 3 es crucial para la provisión de QoS dentro de una red, ya que de acuerdo a ello se asignan las prioridades.

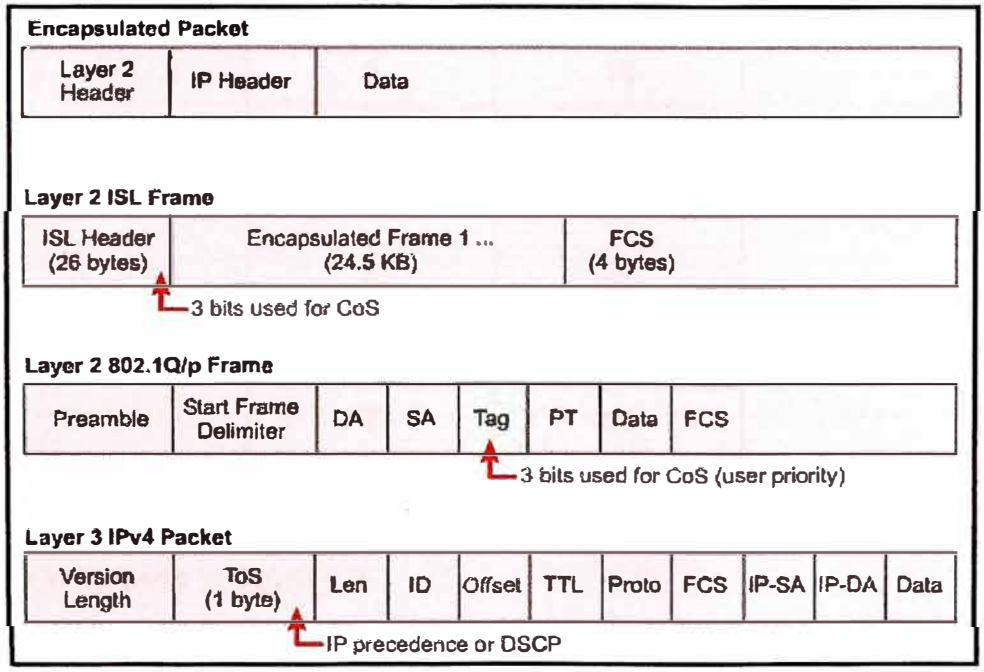

Fig. 4.8. Marcaje de QoS a nivel 2 y 3.

La decisión de si usar el marcado a nivel 2 y/o 3, debe ser tomada de acuerdo a algunas consideraciones.

- ✓ El marcaje a nivel de capa 2 (en tramas) puede ser dado para tráfico que no es IP.
- ✓ El marcaje a nivel de capa 3 (en paquetes) llevará la información de QoS de extremo a extremo. Debemos tomar en cuenta que algunos equipos antiguos no entienden DSCP.

En muchas instancias es necesario el uso de diferentes técnicas de marcado en distintos puntos de la red. Por esta razón es común el usar la precedencia IP o DSCP en el marcaje de paquetes para los routers y el marcaje de tramas usando la CoS disponible en los equipos de capa 2 (switches).

Algunos equipos pueden, de acuerdo al valor de CoS brindado en la cabecera de la trama (nivel 2), pueden remarcar el byte DSCP del paquete IP (nivel 3).

El detalle esta correspondencia lo mostramos en la tabla 4.3:

Tabla Nº 4.3. Relación CoS vs. DSCP.

CoS	DSCP
0	0
1	8
2	16
3	24
4	32
5	40
6	48
7	56

Del ejemplo expuesto en la figura 4.4, a un CoS de 5 le correspondería un valor DSCP de 40, tal como lo muestra la figura 4.9.

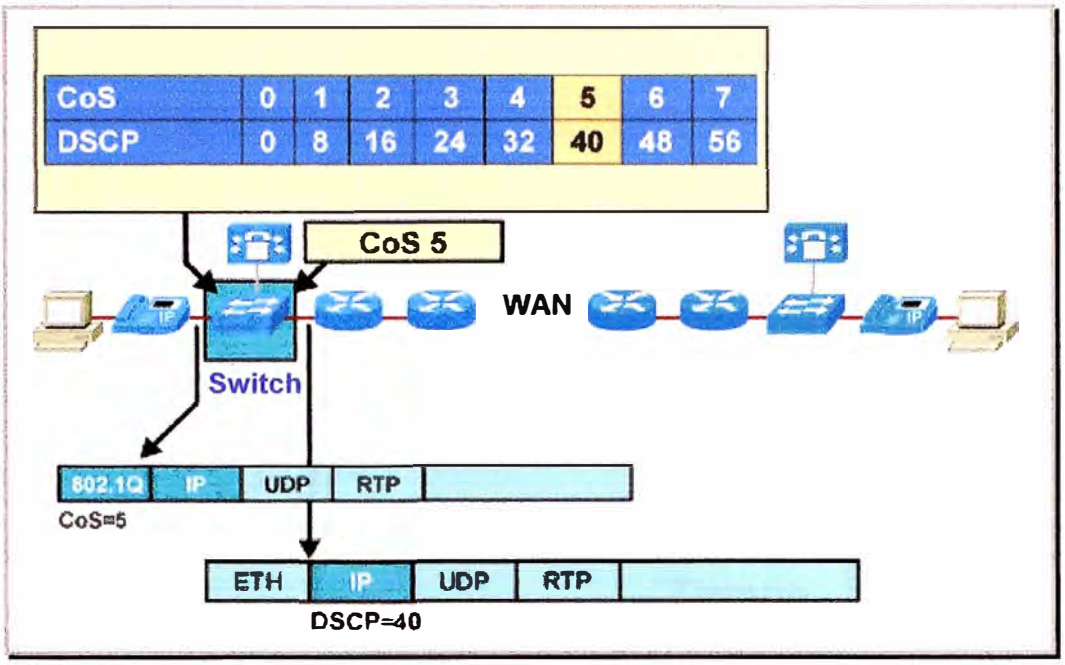

Fig. 4.9. Correspondencia entre CoS y DSCP.

4.4. Modelamiento del tráfico

Muchas veces es necesario limitar el tráfico saliente en una interfaz determinada, con el fin de administrar eficientemente los recursos de la red. Ante esta necesidad existen dos metodologías de limitación de ancho de banda: el modelamiento del tráfico y el control de ráfagas.

Las técnicas de modelamiento de tráfico son un poco más diplomáticas en el sentido en que operan. En vez de descartar el tráfico que excede cierta tasa determinada, atrasan parte del tráfico sobrante a través de colas, con el fin de modelarla a una tasa que la interfaz remota pueda manejar. El resto del tráfico excedente es inevitablemente descartado.

Es una buena herramienta en situaciones en las cuales el tráfico saliente debe respetar una cierta tasa máxima de transmisión. Este proceso es realizado independientemente de la velocidad real del circuito. Esto significa que es posible modelar tráficos de Web o FTP a velocidades inferiores a las del receptor.

La idea es manejar la congestión al forzar a que los paquetes se transmitan a una razón más predecible. Lo que se hace es regular la tasa promedio de ráfagas de transmisión de

datos, en este caso si se llega a un exceso de capacidad los paquetes son encolados, no descartados.

Como nota adicional podríamos decir que el modelamiento generalmente es aplicado a tráfico plano. El efecto de esta técnica se muestra en la figura 4.10.

Fig. 4.10. Modelamiento del tráfico.

4.5. Control de ráfagas

Mediante este bloque se especifica la limitación a un máximo de tasa de transmisión para una clase de tráfico. Si este umbral es excedido, una de las acciones inmediatas será ejecutada: transmitir, descartar, o remarcar. En otras palabras, no es posible almacenar los paquetes para posteriormente enviarlos, como es el caso del modelamiento del tráfico, ejemplo de ello es la figura 4.11.

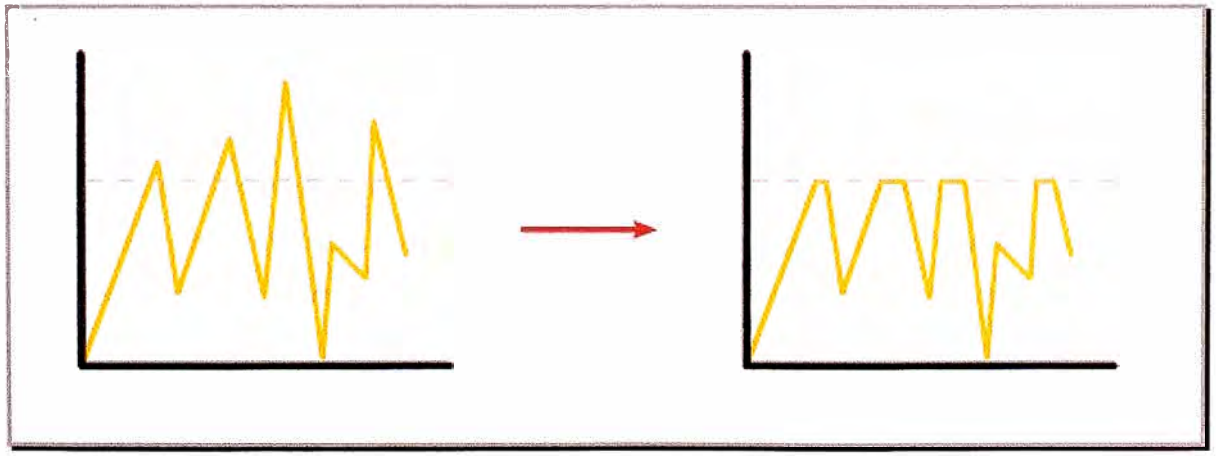

Fig. 4.11. Control de ráfagas.

Los paquetes en exceso pueden ser descartados o se les puede dar un marcado especial. El control de ráfagas descarta o remarca los paquetes en exceso si es que

sobrepasan el límite definido. El tráfico que se originado en ráfagas se propagan por la red, no son suavizados como en el modelamiento del tráfico. Controla la tasa de salida mediante descarte de paquetes, por lo que disminuye el retardo por encolamiento. Sin embargo debido a estos descartes, el tamaño de la ventana deslizante de TCP debe reducirse, afectando el rendimiento global del flujo.

En varios casos es necesario utilizar una vía con la velocidad adecuada para transmitir un paquete de alta o baja prioridad. Por ejemplo, si se tienen dos enlaces, una con mayor velocidad que el otro, sería lógico plantear la metodología de transmisión de mejor esfuerzo para los paquetes de menor prioridad sobre el enlace de menor velocidad.

La tabla 4.4 nos muestra un resumen entre las principales diferencias entre el modelamiento del tráfico y el control de ráfagas.

Tabla Nº 4.4. Diferencias entre control de ráfagas y el modelamiento del tráfico.

Control de ráfagas	Modelamiento del tráfico
En ambas direcciones, entrante y saliente.	Solamente en dirección saliente.
Los paquetes fuera del perfil son descartados.	Los paquetes fuera del perfil son encolados hasta que el buffer se llene.
Causa retransmisión de TCP	Minimiza las retransmisiones de TCP.
Soporta marcado de paquetes para cambios de prioridad.	El marcado/remarcado no es soportado.

4.6. Encolamiento

Las colas protegen y aíslan el tráfico. El encolamiento es un componente característico de QoS que determina cómo las colas de salida son atendidas. Cada cola de salida recibe una asignación de ancho de banda y buffers para su almacenamiento. Una cola puede ser reservada para tráfico de baja latencia, como se podría necesitar en el caso de VoIP.

Los algoritmos de programación reordenan las colas de transmisión para ofrecer servicio prioritario a flujos específicos de acuerdo a la política establecida.

A continuación explicaremos cuáles son las técnicas de encolamiento.

4.6.1. FIFO[16]

Es el tipo más simple de encolamiento, se basa en que: el primer paquete en entrar a una interfaz, es el primero en salir. Es adecuado para interfaces de alta velocidad, sin embargo no para bajas, ya que FIFO es capaz de manejar cantidades limitadas de ráfagas de datos.

Si llegan más paquetes de los que la cola soporta, éstos son descartados. No tiene mecanismos de diferenciación de paquetes. Es simplemente un carril de una sola vía tal como lo muestra la figura 4.12.

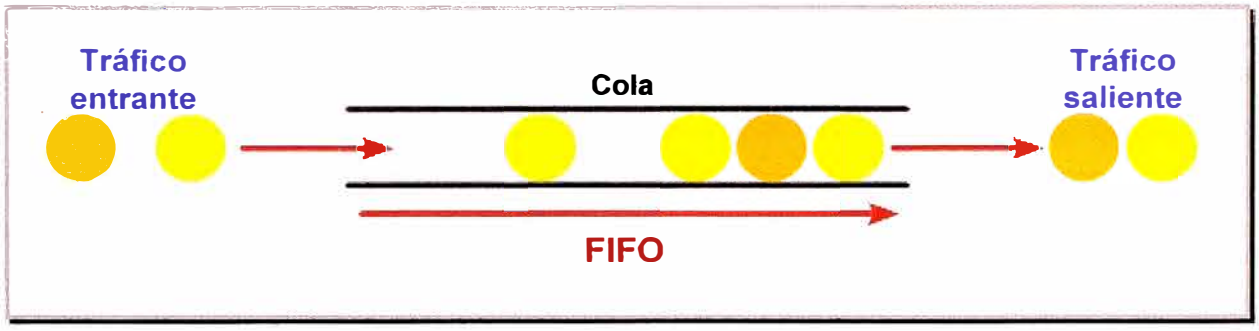

Fig. 4.12. El primero que entra el primero que sale.

4.6.2. Encolamiento Prioritario

El encolamiento prioritario, también denominado PQ[17], consiste en un conjunto de colas, clasificadas desde alta a baja prioridad. Cada paquete es asignado a una de estas colas, las cuales son servidas en estricto orden de prioridad. Las colas de mayor prioridad son siempre atendidas primero, luego la siguiente de menor prioridad, y así. Si una cola de menor prioridad está siendo atendida, y un paquete ingresa a una cola de mayor prioridad, ésta es atendida inmediatamente, tal como lo muestra la figura 4.13. Este mecanismo se ajusta a condiciones donde existe un tráfico importante, pero puede causar la total falta de atención de colas de menor prioridad

---

[16] FIFO: acrónimo de First In, First Out.

[17] PQ, acrónimo de Priority Queuing

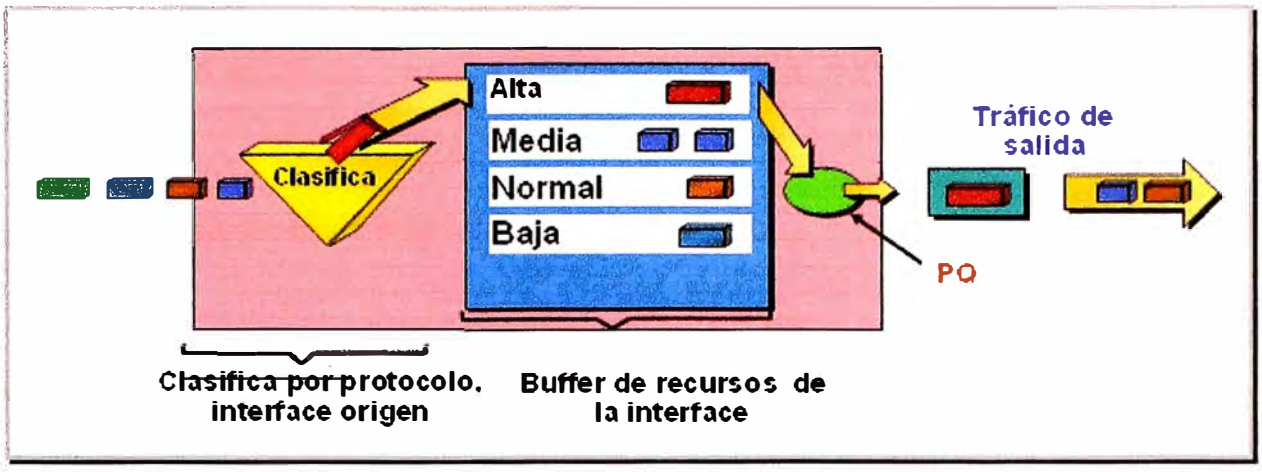

Fig. 4.13. Encolamiento prioritario.

4.6.3. Encolamiento Personalizado

El encolamiento personalizado, también denominado CQ[18], para evadir la rigidez del encolamiento prioritario, se opta por utilizar encolamiento personalizado, que permite priorizar el tráfico sin los efectos laterales de inanición de las colas de baja prioridad, especificando el número de paquetes o bytes que deben ser atendidos para cada cola.

Se pueden crear hasta 16 colas para categorizar el tráfico, tal como lo muestra la figura 4.14.

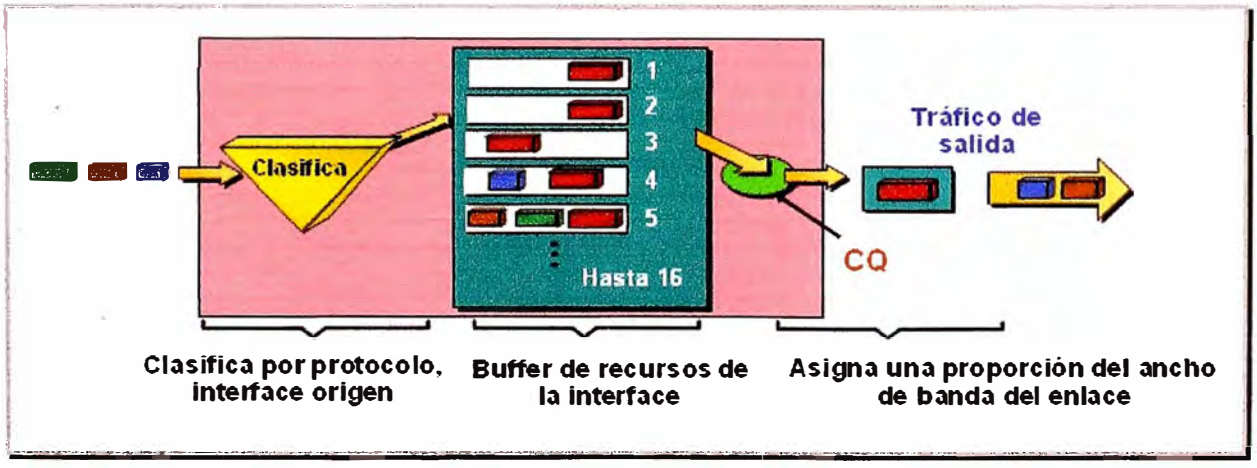

Fig. 4.14. Encolamiento personalizado.

CQ ofrece un mecanismo más refinado de encolamiento, pero no asegura una prioridad absoluta como PQ. Se utiliza CQ para proveer a tráficos particulares de un ancho de

---

[18] CQ: acrónimo de Custom Queuing.

banda garantizado en un punto de posible congestión, asegurando para este tráfico una porción fija del ancho de banda y permitiendo al resto del tráfico utilizar los recursos disponibles.

4.6.4. Encolamiento de baja latencia

También denominado LLQ[19], es una mezcla entre PQ y CBWFQ. Es actualmente el método de encolamiento recomendado para la Voz sobre IP y Telefonía IP.

LLQ consta de colas de prioridad personalizadas, basadas en clases de tráfico, en conjunto con una cola de prioridad, la cual tiene preferencia absoluta sobre las otras colas. Si existe tráfico en la cola de prioridad, ésta es atendida antes que las otras colas de prioridad personalizadas. Si la cola de prioridad no está encolando paquetes, se procede a atender las otras colas según su prioridad. Debido a este comportamiento es necesario configurar un ancho de banda límite reservado para la cola de prioridad, evitando la inanición del resto de las colas.

La cola de prioridad que posee LLQ provee de un máximo retardo garantizado para los paquetes entrantes en esta cola, el cual es calculado como el tamaño del MTU dividido por la velocidad de enlace. En la figura 4.17 ilustramos este tipo de encolamiento.

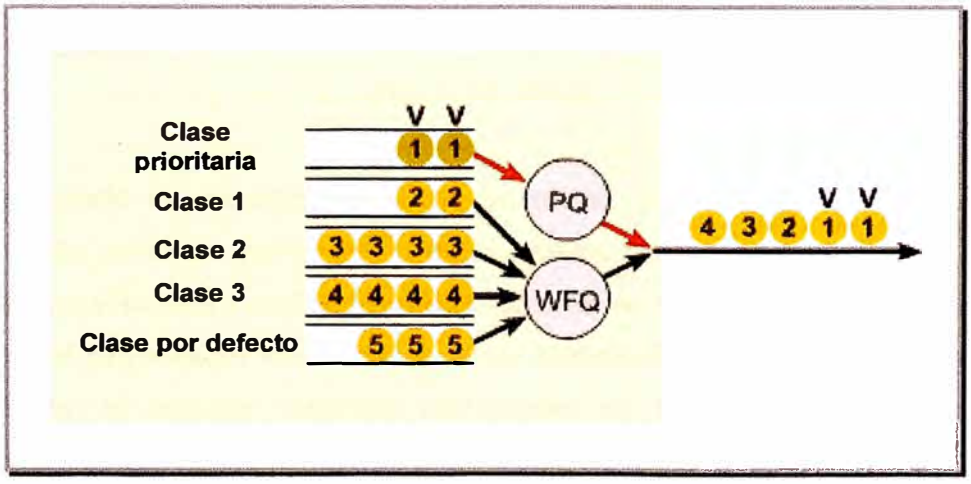

Fig. 4.17. LLQ.

[19] LLQ: acrónimo de Low-Latency Queueing

4.6.5. WFQ[20]

Esta técnica es un método automatizado que provee una justa asignación de ancho de banda para todo el tráfico de la red, utilizado habitualmente para enlaces de velocidades menores a 2 Mbps.

WFQ ordena el tráfico en flujos, utilizando una combinación de cinco parámetros: dirección origen, dirección destino, puerto origen, puerto destino, y protocolo. Una vez distinguidos estos flujos, el router determina cuáles son de uso intensivo o sensibles al retardo, priorizándolos y asegurando que los flujos de alto volumen sean empujados al final de la cola, y los volúmenes bajos, sensibles al retardo, sean empujados al principio de la cola, tal como lo muestra la figura 4.15.

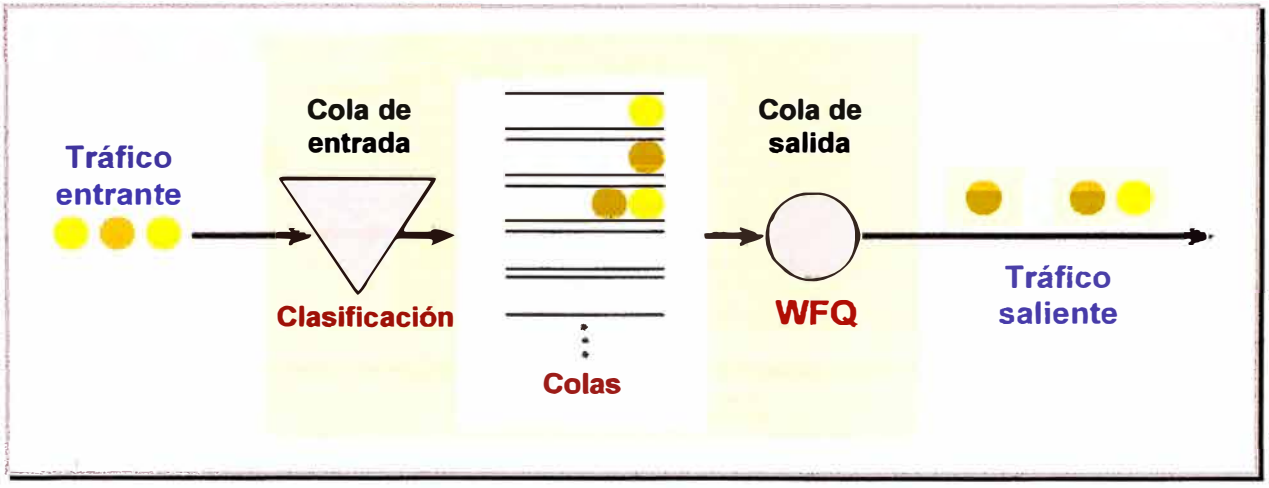

Fig. 4.15. WFQ.

WFQ es apropiado en situaciones donde se desea proveer un tiempo de respuesta consistente ante usuarios que generen altas y bajas cargas en la red, ya que WFQ se adapta a las condiciones cambiantes del tráfico en la red. Sin embargo, la carga que significa para el procesador en los equipos de enrutamiento, hace de esta metodología poco escalable, al requerir recursos adicionales en la clasificación y manipulación dinámica de las colas.

4.6.6. CBWFQ[21]

---

[20] WFQ: acrónimo de Weighted Fair Queueing.
[21] CBWFQ: acrónimo de Class-Based Weighted Fair Queuing.

WFQ tiene algunas limitaciones de escalamiento, ya que la implementación del algoritmo se ve afectada a medida que el tráfico; colapsa debido a la cantidad numerosa de flujos que analizar. CBWFQ fue desarrollada para evitar estas limitaciones, tomando el algoritmo de WFQ y expandiéndolo, permitiendo la creación de clases definidas por el usuario, que permiten un mayor control sobre las colas tráfico y asignación del ancho de banda.

En la figura 4.16 ilustramos la lógica de CBWFQ.

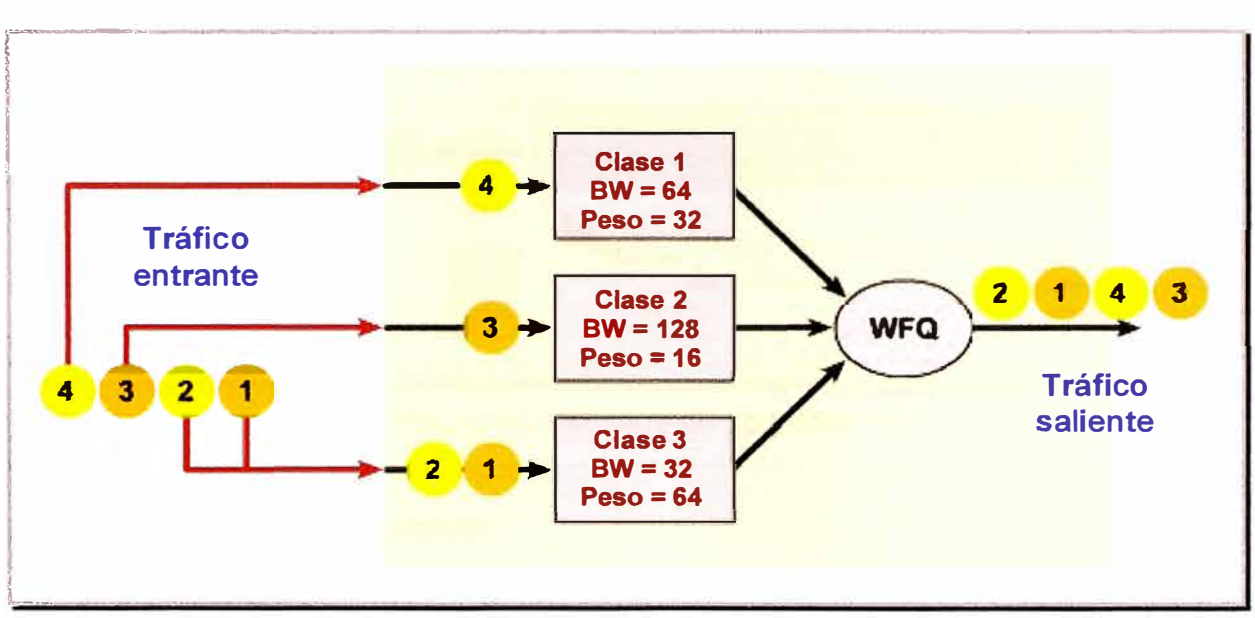

Fig. 4.16. CBWFQ.

Algunas veces es necesario garantizar una determinada tasa de transmisión para cierto tipo de tráfico, lo cual no es posible mediante WFQ, pero si con CBWFQ. Las clases que son posibles implementar con CBWFQ pueden ser determinadas según protocolo, listas de acceso, valor DSCP, o interfaz de ingreso. Cada clase posee una cola separada, y todos los paquetes que cumplen el criterio definido para una clase en particular son asignados a dicha cola. Una vez que se establecen los criterios para las clases, es posible determinar cómo los paquetes pertenecientes a dicha clase serán manejados. Si una clase no utiliza su porción de ancho de banda, otras pueden hacerlo. Se puede configurar específicamente el ancho de banda y límite de paquetes máximos (o profundidad de cola) para cada clase. El peso asignado a la cola de la clase es determinado mediante el ancho de banda asignado a dicha clase.

4.7. Descarte

En este bloque los paquetes que llegan a la cola son descartados si cola se llena, esto se toma como última opción. Los paquetes en una cola pueden tener diferentes valores de descarte, es decir es posible definir una política de descartes aleatorios.

Es lógico que al llenarse un buffer ya no hay espacio para más paquetes, por ello son descartados, ello lo ilustramos en la figura 4.18.

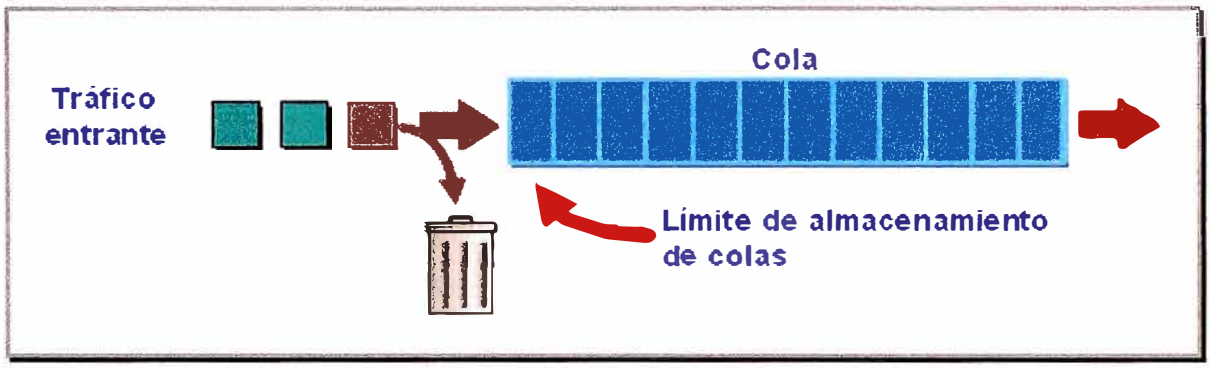

Fig. 4.18. Descarte de paquetes

4.8. Evasión de la congestión

Las metodologías de evasión de congestión se basan en la manera que los protocolos operan, con el fin de no llegar a la congestión de la red. Las técnicas de RED (Random Early Detection) y WRED (Weighted Random Early Detection) evitan el efecto conocido como sincronización global. Cuando múltiples conexiones TCP operan sobre un enlace común, todas ellas incrementaran el tamaño de su ventana deslizante a medida que el tráfico llega sin problemas. Este aumento gradual consume el ancho de banda del enlace, hasta congestionarlo. En este punto las conexiones TCP experimentan errores de transmisión, lo que hace que disminuyan su tamaño de ventana simultáneamente. Esto conlleva a una sincronización global, donde todos los flujos comienzan a incrementar su tasa de transmisión nuevamente para llegar a otro estado de congestión. Este ciclo es repetitivo, creando picos y valles en la utilización del ancho de banda del enlace. Es debido a este comportamiento que no se utiliza los máximos recursos de la red.

Los métodos de evasión de congestión tratan con este tipo de situación, descartando paquetes de forma aleatoria. RED fuerza a que el flujo reduzca el tamaño de su ventana de transmisión, disminuyendo la cantidad de información enviada. A medida que se

alcanza el estado de congestión en la red, más paquetes entrantes son descartados con el fin de no llegar al punto de congestión en el enlace.

Lo que limita a estas técnicas de evasión de congestión es que sólo sirve para tráfico basado en TCP, ya que otros protocolos no utilizan el concepto de ventana deslizante.

CAPÍTULO V
IMPLEMENTACIÓN DE QoS

5.1. Ejemplo de aplicación en router Cisco

A manera de ejemplo mostramos cómo se configuran los routers del fabricante Cisco. A continuación haremos una breve explicación y reseña de cómo se aplican estos comandos así como la metodología a utilizar según las políticas definidas. Estos son los comandos básicos:

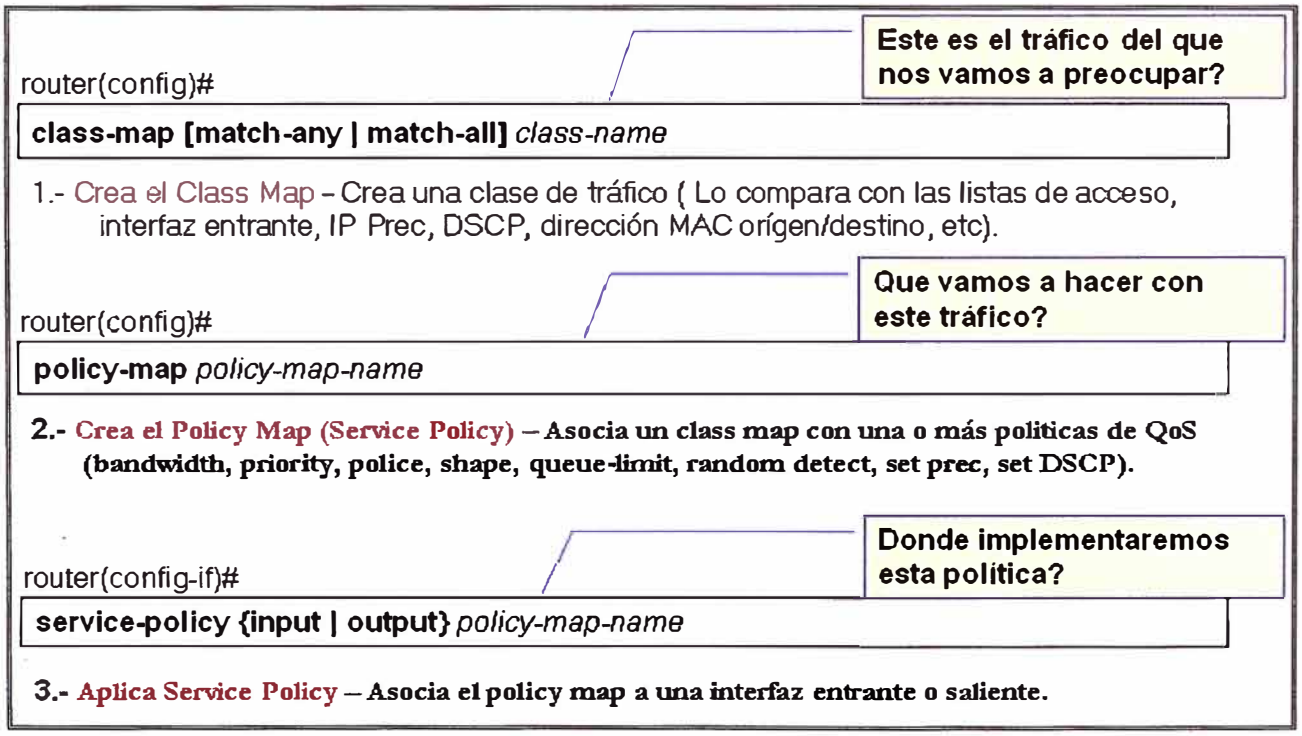

Fig. 5.1. Esquema de configuración para el caso de routers Cisco.[22]

5.2. Creación de clases

En el primer comando se está creando una clase, en el caso de Cisco pueden existir varias clases, por ejemplo podríamos utilizar una clase para lo que es datos simples (no muy sensible al retardo, por ejemplo internet, correo, etc.), otra clase para datos

[22] Imagen del documento "Introducción a QoS" de Irene Owen, Cisco Systems

importantes (por ejemplo algún software corporativo o transacciones bancarias) y por último una clase para voz (muy sensible al retardo suponiendo que también manejemos VoIP.

Un mapa de clase es un mecanismo para nombrar y aislar un flujo de tráfico específico, éste define el criterio utilizado para comparar el tráfico para más tarde clasificarlo, el cual puede incluir selecciones mediante listas de acceso. Después que el paquete es confrontado al criterio del mapa de clase, es posible clasificarlo mediante el uso de mapas de política.

Un mapa de política especifica en qué clase se actuará. Las acciones pueden ser confiar en los valores de CoS, DSCP o precedencia IP de la clase de tráfico, establecer un valor específico de éstos, o especificar las limitaciones de ancho de banda y la acción a tomar cuando el tráfico cae fuera del perfil definido en el mapa de política.

Para manipular los tráficos y otorgarles calidad de servicio, se utilizan los procedimientos básicos de clasificación y asignación de prioridad, denominados mapas de clase y mapas de política. En resumen las clases se utilizan para agrupar tipos de tráfico, dar orden y establecer una jerarquía.

La configuración de esta parte sería la mostrada a continuación:

---

router(config)# ***class-map [match-any | match-all]*** *class-name*

Supongamos que tenemos dos clases una para voz y uno para datos, la configuración de esta parte sería de la siguiente manera:

```
class-map match-all VOZ
 match access-group 100
class-map match-all DATOS
 match access-group 101
```

Teniendo en cuenta que hemos creado previamente dos listas de acceso:

- ✓ Una para las **VOZ** identificada con el access-list **100**, y
- ✓ Otra para **DATA** identificada con el access-list **101**.

En la tabla 5.1 presentamos las posibilidades de configuración en el caso de routers Cisco.

Tabla Nº 5.1. Configuración de clases.

```
Router(config)# class-map match-all video

Router(config-cmap)# match ?
  access-group          Grupo de acceso
  any                   Cualquier paquete
  class-map             Clase mapeada
  cos                   Clase de Servicio
  destination-address   Dirección destino
  input-interface       Interface entrante para hacer match
  ip                    Especifica valores IP
  mpls                  Especifica valores MPLS
  not                   Niega el resultado de este match
  protocol              Protocolo
  qos-group             Grupo de QoS
  source-address        Dirección origen
```

5.3. Creación de políticas

En segundo lugar tenemos un comando para crear políticas, éste implica la creación de una jerarquía mayor, es decir la política o policy-map como lo llama Cisco, agrupa varias clases, es aquí donde se da el tratamiento de QoS. La configuración de esta parte la mostrada a continuación:

```
router(config)# policy-map policy-map-name
```

Si suponemos que tenemos un ancho de banda del 256K, configuraremos nuestra política de la siguiente manera

```
policy-map IPVPN
    class VOZ
        priority 64
        set precedence 5
    class DATOS
        bandwidth 128
        set precedence 1
    class class-default
        fair-queue
```

Es decir se crea una política llamada **IPVPN**, y se definen los parámetros de tres clases, dos ya definidas y una por defecto:

- ✓ Una para **VOZ**, con un ancho de banda asegurado de 64 Kbps, con un encolamiento LLQ y marcando los paquetes con una precedencia IP de 5.

- ✓ Una para **DATOS**, con un ancho de banda asegurado de 128 Kbps, con un encolamiento CBWFQ, y marcando los paquetes con una precedencia IP de 1.

- ✓ Y por último indicamos una clase por defecto (**class-default**), en la cual se encolaran el resto de paquetes mediante el WFQ y por defecto se les asignará una precedencia IP de 0.

Sólo a modo de ejemplo presentamos la tabla 5.2, que nos muestra las opciones de marcación de paquetes que tiene un router Cisco.

Tabla Nº 5.2. Posibilidades de marcación de paquetes en un router Cisco.

Router(config-pmap-c)#set ?	
atm-clp	Set ATM CLP bit to 1
cos	Set 802.1Q ISLCoS service/user priority
ip	Set IP specific values
mpls	Set MPLS specific values
qos-group	Set QoS Group
Router(config-pmap-c)#set ip ?	
dscp	Set IP DSCP (DiffServ CodePoint)
precedence	Set IP precedence

5.4. Aplicación de políticas

Por último tenemos el comando que se utiliza para aplicar la política creada en determinada interfaz del router, generalmente es aplicada como una política de salida, tal como lo podemos apreciar:

router(config-int)# ***service-policy {input | output}*** *policy-map-name*

Si en nuestro caso tenemos un equipo con una interfase serial, la aplicaremos de la siguiente forma:

interface Serial0
ip address 10.10.10.1 255.255.255.252
encapsulation ppp
service-policy output IPVPN

De tal manera que la calidad de servicios quedaría configurada de la siguiente manera:

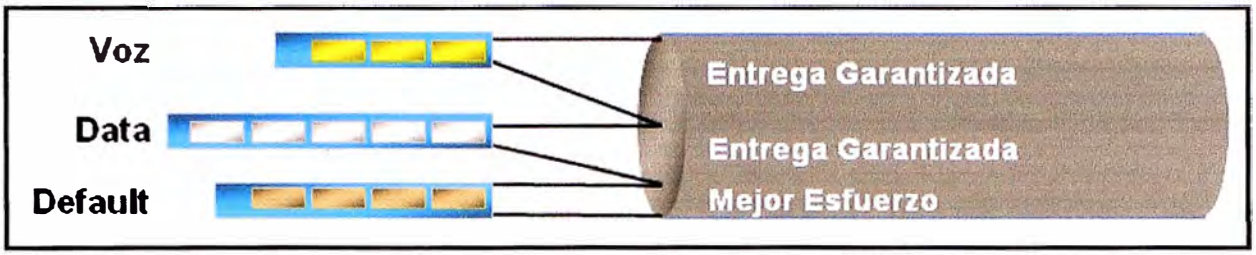

Fig. 5.2. Separación de paquetes de acuerdo a la política aplicada.

5.5. Regla del 75%

Como nota adicional podríamos mencionar la regla de 75% por la cual, por la cual el ancho de banda que deba ser garantizado deberá ser menor o igual que el 75% del ancho de banda de la interfase, tal como lo muestra la figura 5.3

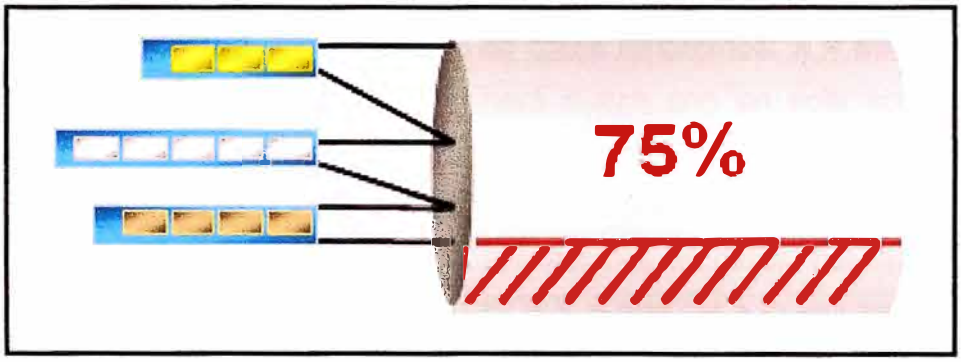

Fig. 5.3. Reserva sólo hasta el 75%.

Ya que se deja espacio libre para señalización SNMP, tráfico de enrutamiento, etc.

5.6. Ejemplo de configuración

En la tabla 5.1 se detalla la configuración básica que se debe incluir el los routers, específicamente del fabricante Cisco, para brindar QoS.

La diferencia con respecto a los ejemplos anteriores parte de que el marcado de paquetes se hace a través de la interfase entrante es decir en el caso de los datos la interfase Ethernet.

Adicionalmente se muestra la configuración completa del router.

Tabla Nº 5.1. Configuración básica de QoS en un router Cisco.

Comandos de Línea	Descripción
ip telnet tos 0	Se le asigna el tipo de servicio ToS en hexadecimal igual a 00 que corresponde a caudal bronce, es decir precedencia IP igual a 0)
class-map match-all **ORO** match access-group 100 class-map match-**X PLATA** match access-group 101 match access-group 102	El comando **class-map** hace el match con las listas de acceso predeterminadas para los tráficos de voz y datos. Si se considera un solo tipo de aplicaciones de datos tendremos: **X = all** ; y a la vez se hará match con un solo access-group de datos (access-group 101) Si se considera 02 ó más tipos de aplicaciones en datos, por ejemplo: HTTP y FTP **X = any** ; y a la vez se hará con 02 ó más access-group de datos (access-group 101 y access-group 102) respectivamente.
policy-map **IPVPN** class **ORO** priority **M** police cir **m** conform-action **transmit** exceedaction **drop** class **PLATA** bandwidth **N** shape average **n**	**IPVPN:** Nombre del Service-policy output. **ORO:** Nombre asignado para tráfico ORO. **PLATA:** Nombre asignado para tráfico PLATA. El comando **policy-map** define el ancho de banda para los tráficos de voz y datos respectivamente. Ancho de Banda = 75% de la velocidad de línea. **M + N** = 75% Velocidad de línea. Siendo **m** el caudal contratado ORO en bps.

	Siendo **n** el caudal contratado PLATA en bps.
interface FastEthernet0/1 description **<Description>** ip address **<IP WAN> <MASK>** no ip redirects no cdp enable speed 100 duplex-full service-policy output **IPVPN**	**<Description>:** Algún comentario sobre la interfase. **<IP WAN>:** Dirección IP WAN que se le asigna al router del cliente. **<MASK>:** Máscara de la subred
interface FastEthernet0/0 ip address **<IP LAN> <MASK>** speed 10 duplex-full ip route-cache policy ip policy route-map **DATOS**	**<IP LAN>:** Dirección IP LAN, asignada por el cliente. **<MASK>:** Máscara de la subred **DATOS:** Nombre de route-map para tráfico datos
access-list **100** permit tcp any eq 1720 any access-list **100** permit tcp any any eq 1720 access-list **100** permit udp any any range 16384 32767 access-list **100** permit udp any range 16384 32767 any access-list **101** permit tcp any any eq 80 access-list **101** permit tcp any eq 80 any access-list **102** permit tcp any eq 20 any access-list **102** permit tcp any any eq 20 access-list **102** permit tcp any eq 21 any access-list **102** permit tcp any any eq 21	Creación de los access-list para tráfico de voz y datos: ✓ Access-list **100**, para tráfico sólo voz. ✓ Access-list **101**, **102** para tráfico de datos IP en general. En caso de tener sólo un tipo de tráfico datos sólo crear el access-list 101. El puerto TCP=1720 permite el establecimiento de llamadas extremo a extremo. El rango de puertos UDP está entre 16384 y 32767, ello permite el intercambio de información de voz. El puerto TCP=80 corresponde a las

	sesiones http como aplicativo de datos. El puerto TCP=20 corresponde al aplicativo FTP para la transferencia de datos. El puerto TCP=21 corresponde al control para el FTP.
route-map **DATOS** permit 10 match ip address 101 set ip precedence **1** route-map **DATOS** permit 20 match ip address 102 set ip precedence **0**	Se asigna la prioridad de datos <0,1> o **<prioridad>:** Precedencia IP = 1 (priority), para datos importantes. o **<prioridad>:** Precedencia IP = 0 (routine), para datos sin importancia.
dial-peer voice 10 pots destination-pattern **<Anexo1>** forward digital all port 1/0 dial-peer voice 20 voip destination-pattern **<Anexo2>** session target ipv4: **< IP Dest.>** **ip precedence 5** codec g723r53	**Anexo1:** Anexo local **Anexo2:** Anexo remoto **<IP Dest>:** IP address WAN ó LAN del local remoto **ip precedence 5:** Setea la precedencia IP a 5 para el tráfico de voz con calidad ORO.

CONCLUSIONES

1. Las aplicaciones están consiguiendo ser cada vez más exigentes, Las denominadas críticas requieren cada vez más calidad, confiabilidad, para asegurar la puntualidad en la entrega de paquetes. Un ejemplo claro son las aplicaciones de voz o vídeo, éstas deben ser manejadas cuidadosamente dentro de una red del IP para preservar su integridad.

2. Las tecnologías de QoS cobran especial importancia hoy en día proporcionándonos utilidades para la entrega de datos críticos de un negocio en los tiempos requeridos y con unas garantías determinadas.

3. Podríamos decir que la aplicación de QoS nos servirá en caso nuestra red se encuentre congestionada.

4. Cuatro son los parámetros que degradan la calidad de servicio: el retardo, la variación del retardo, el ancho de banda y la pérdida de paquetes, para contrarrestarlas utilizamos tanto las técnicas de encolamiento, como las de separación de anchos de banda, y las predicción de saturación de tráfico.

5. La idea de las políticas es crear jerarquías y un orden para la asignación de recursos y priorización en la red.

6. Es vital tener una política de QoS definida para administrar los recursos de nuestra red de manera óptima, para ello existe el modelo QPIM, por el cual se eligen las políticas de acuerdo a tres premisas: la topología de la red, la metodología de QoS, y las reglas del negocio.

7. Podemos utilizar la combinación de diversos modelos de implementación de QoS, como: mejor esfuerzo, servicios integrados y servicios diferenciados.

8. Para implementar y entender QoS debemos saber como estructurarlo funcionalmente, para así llegar a un óptimo control de recursos. Los bloques explicados en este informe son: clasificación, marcación, modelamiento del tráfico, control de ráfagas, encolamiento y descarte.

9. En el caso de los routers Cisco, es posible la aplicación de políticas y priorización de acuerdo a clases previamente definidas, pudiendo utilizar para el marcaje de clase una lista de acceso.

ANEXO A

GLOSARIO

1. **CoS:** acrónimo de Class of Service
2. **DiffServ:** acrónimo de Differentiated Services.
3. **DSCP:** acrónimo de DiffServ Code Point
4. **FIFO:** acrónimo de First In, First Out
5. **IntServ:** acrónimo de Integrated Services
6. **IP:** acrónimo de Internet Protocol
7. **IPv4:** acrónimo de Internet Protocol version 4
8. **IPv6:** acrónimo de Internet Protocol version 6
9. **MPLS:** acrónimo de siglas de Multiprotocol Label Switching
10. **QoS:** acrónimo de Quality of Service
11. **QPIM:** acrónimo de QoS Policy Information Mode
12. **RFC:** acrónimo de Request For Comments.
13. **RSVP:** acrónimo de ReSerVation Protocol
14. **ToS:** acrónimo de Type of Service

BIBLIOGRAFIA

1. Irene Owen, "Introduction to QoS"
2. Santiago Felici, "Calidad de Servicio (CoS y QoS)"
3. Jairo Daniel Pérez, "Calidad de Servicio en Redes (QoS)", EAFIT 2003
4. Cisco Systems, "Curso BSCI de Cisco", Semestre 3, Cap. 8.
5. Cisco Systems, "Implementing Cisco Quality of Service – Volumen I y II", guía de estudiante y diapositivas del curso.
6. CBT Nuggets, "CCVP QoS", material multimedia
7. RFC 3644, "Policy QoS Information Model"
8. Cisco Systems: http://www.cisco.com
9. Revista AHCIET: http://www.ahciet.net/REVISTA/90/qos.pdf
10. IETF:http://www.ietf.org