Capítulo 7

Qualidade de Serviço

Neste capítulo será analisada a necessidade de garantias de desempenho fim-a-fim para a transmissão tempo-real de áudio e vídeo. Para tanto, inicialmente serão analisados os principais componentes das aplicações multimídia e suas influências no desempenho do sistema.

7.1 Gerenciamento de Qualidade de Serviço

Para comunicações multimídia é necessário garantias de desempenho fim-a-fim, desde a fonte da informação até o destino. Para fornecer uma única abordagem para que diferentes aplicações especifiquem as garantias de desempenho necessárias e para que sistemas forneçam as garantias requeridas, foi introduzido o conceito de Qualidade de Serviço (QoS).

7.1.1 Definição de QoS

Não existe uma definição de QoS universalmente aceita atualmente. Baseado em [Lu, 96], podemos definir QoS como:

> **QOS É UMA ESPECIFICAÇÃO QUALITATIVA E QUANTITATIVA DOS REQUISITOS DE UMA APLICAÇÃO OU DE UM CLIENTE QUE UM SISTEMA DEVERIA SATISFAZER AFIM DE OBTER A QUALIDADE DESEJADA.**

Nesta definição, o termo sistema inclui todos os elementos envolvidos na comunicação fim-a-fim entre as partes envolvidas na aplicação multimídia (p.e., Sistemas Operacionais e a infra-estrutura de rede). Além desse, o termo cliente se refere ao cliente de um serviço oferecido por um provedor de serviços de comunicação (p.e., Provedor de Serviço Internet). Baseada nesta definição, existem dois aspectos para QoS: aplicações ou clientes especificam os requisitos de QoS e o sistema fornece as garantias de QoS. A QoS é normalmente especificada por um conjunto de parâmetros, a nível de rede são normalmente utilizados a taxa de bits, taxa de erros, limites de atraso e de variação de atraso. Um ou mais valores são associados a cada parâmetro. Eles especificam o valor desejado ou um intervalo admissível.

A noção de QoS foi inicialmente usada em comunicações de dados para caracterizar o desempenho da transmissão de dados em termos de confiabilidade, atraso e vazão. Por exemplo, o modelo de referência OSI tem alguns parâmetros de QoS descrevendo a velocidade e confiabilidade da transmissão, tal como vazão, atraso de trânsito, taxa de erro e probabilidade de falha de estabelecimento de conexão. Estes parâmetros são especificados na camada de transporte e não tem seus significados diretamente observáveis ou verificáveis pela aplicação. Estes parâmetros não cobrem todos os requisitos da comunicação multimídia e são usados ao nível de transporte apenas. Além disso, nenhum mecanismo é especificado no modelo de referência OSI para garantir os requisitos de QoS especificados. Para comunicações multimídia, a QoS deve ser especificada e garantida fim-a-fim em todos os níveis. Portanto, aplicações multimídia requerem um novo modelo de QoS.

7.1.2 Estrutura Geral de QoS

[Lu, 96] propõe um modelo simplificado de operações de QoS de um sistema de comunicação multimídia, onde:

- A aplicação especifica seus requisitos de QoS e submete ao sistema.
- A partir da especificação da QoS requerida, o sistema determina se ele tem recursos suficientes para satisfazer os requisitos. Em caso afirmativo, ele aceita a aplicação e aloca os recursos necessários. Caso contrário, o sistema pode rejeitar a aplicação ou sugerir uma QoS mais baixa.

Baseado neste modelo de operação, os seguintes elementos são necessários para fornecer garantias de QoS:

- Um mecanismo de especificação da QoS para aplicações especificarem seus requisitos.
- Um processo de negociação de QoS permitindo admissão de várias aplicações.

- Controle de admissão para determinar se novas aplicações possam ser admitidas sem afetar a QoS das aplicações atuais.
- Alocação de recursos e escalonamento para satisfazer os requisitos.
- Policiamento de tráfego para tornar seguro que aplicações geram o correto conjunto de dados conforme a especificação aceita.
- Mecanismos de renegociação são necessários de modo que as aplicações possam mudar suas especificações de QoS iniciais.
- Mecanismos de monitoramento da sessão afim de que na ocorrência de problemas no provimento da QoS negociada sejam tomadas as devidas providências.
- Técnicas de degradação gradativa da qualidade e outras deveriam ser usadas juntamente com o mecanismo anterior para fornecer serviços satisfatórios para aplicações multimídia.

Todos os subsistemas devem prover os mecanismos acima e cooperar para fornecer garantias de QoS fim-a-fim. Para facilitar o desenvolvimento de uma interface QoS configurável e controle dirigido a QoS e mecanismos de gerenciamento em todas as camadas arquiteturais, vários arquiteturas de QoS foram desenvolvidas [Campbell, 94][Campbell, 96]. Estas arquiteturas fornecem uma abordagem semântica para fornecer garantias de QoS fim-a-fim.

Na seqüência, serão descritos alguns elementos importantes e problemas do gerenciamento de QoS.

7.1.3 Diferentes Níveis de Garantia

Até aqui nós falamos que a QoS deveria ser garantida. Na prática, o usuário pode especificar um grau (ou nível) de garantia. Em geral, existem três níveis de garantia:

- Garantia Determinista ou Hard: a QoS especificada pelo usuário é garantida a 100%. Esta garantia é mais custosa em termos de recursos, pois os recursos são alocados na base pior caso, e eles não podem ser usados por outras aplicações mesmo quando não estão sendo usados, resultando num baixo uso dos recursos. Por exemplo, um vídeo codificado a taxa de bits variável com uma variação na saída de 200kbits/s a 4Mbits/s, os recursos deveriam ser reservados baseado em 4Mbits/s (pior caso). Uma vantagem é que esta garantia é de fácil implementação, pois recursos são reservados estaticamente.
- Garantia Estatística ou Soft: a QoS especificada pelo usuário é garantido em uma certa percentagem. Esta garantia é mais apropriada para mídias contínuas pois elas não necessitam precisão de 100% na apresentação. Além disso, o uso de recursos é mais eficiente. Ele é baseado em multiplexação estatística: recursos não usados por uma aplicação podem ser usados por outras. Este é um modo desejado para comunicação multimídia, mas ele é de difícil implementação devido a natureza dinâmica do tráfego e uso dos recursos.
- Melhor Esforço: neste caso nenhuma garantia é fornecida e a aplicação é executada com os recursos disponíveis. Os sistemas computacionais tradicionais operam neste modo.

7.1.4 Fornecendo Garantias de QoS

QoS pode ser garantida apenas quando recursos suficientes são disponíveis e o escalonamento de processos é apropriadamente implementado. A prevenção de sobrecargas requer **controle de admissão** e a prevenção de que aplicações não utilizem mais recursos do que aquele alocado requer **mecanismos de policiamento**.

Desde que QoS fim-a-fim é necessária, cada subsistema deveria ter funções de gerenciamento de QoS, incluindo cinco elementos: especificação de QoS, controle de admissão, negociação e renegociação, alocação de recursos e policiamento de tráfego. Várias arquiteturas QoS para gerenciar estes subsistemas têm sido propostas. Estas arquiteturas fornecem uma abordagem de trabalho na qual todos os subsistemas devem cooperar para fornecer as garantias de QoS.

7.2 Um exemplo de manipulação de QoS

Um exemplo simples ilustra como os conceitos de QoS apresentados anteriormente são usados na prática. Suponha que um cliente deseja obter e apresentar uma peça de áudio de qualidade telefone de um servidor remoto. Os seguintes passos estabelecem a sessão para o usuário:

- O usuário seleciona o nome do arquivo de áudio e a qualidade de telefone através de uma interface com o usuário.
- A aplicação traduz o requisito do usuário nos seguintes parâmetros: taxa de amostragem - 8 kHz (possivelmente com uma pequena variação), bits por amostra = 8 (uma amostra deve ocorrer a cada 125 μs).
- A aplicação passa o pedido ao sistema operacional cliente, que determina se ele pode processar um byte todo 125 μs com a carga atual no cliente. Se não, a sessão é rejeitada.
- O sistema operacional passa o pedido ao sistema de transporte incluindo protocolo de transporte e todas as camadas mais baixas, que determina se ele pode suportar uma taxa de bits de cerca de 64 kbits/s. Se não, o pedido de sessão é rejeitado.
- O servidor passa o pedido ao controlador de disco para determinar se ele pode suportar uma taxa de transferência de 64 kbits/s sem afetar a QoS das sessões existentes. Se não, o pedido de sessão é rejeitado.
- A sessão é estabelecida com sucesso e o usuário ouve a peça de áudio pedida.

7.3 Qualidade de Serviço na Internet

Como visto anteriormente, a Internet de hoje fornece um serviço do tipo melhor esforço: o tráfego é tratado tão rápido quanto possível, mas não há garantias temporais ou limites de erro. O termo serviço é usado para descrever algo oferecido aos usuários fim da rede, como comunicações fim-a-fim ou aplicações cliente-servidor.

Com a rápida transformação da Internet em uma infra-estrutura comercial, o fornecimento de qualidade de serviço está sendo considerado cada vez mais um requisito essencial. Qualidade de Serviço pode ser visto aqui como a capacidade para diferenciar entre tráfego ou tipos de serviço, de forma que o sistema possa tratar uma ou mais classes de tráfego diferentemente de outros. Neste sentido, seria interessante que um Provedor de Serviços Internet (ISP – *Internet Service Provider*) pudesse fornecer a seus clientes diversas possibilidades em termos de qualidade de serviço, garantido de um certo modo a taxa de bits e atraso. É claro, quando o cliente optasse por um serviço de qualidade, ele pagaria um custo maior do que aquele pago por um tipo de serviço melhor esforço.

Para isto, o cliente de um ISP optará por uma das várias classes de serviço (garantindo diferentes qualidades). Por exemplo, uma classe de serviço forneceria serviços Internet "previsíveis" para companhias que fazem negócios na Web. Tais companhias têm interesse em pagar um certo preço para tornar seus serviços confiáveis e fornecer a seus usuários um acesso rápido a seus sites Web. Esta classe de serviço poderia conter um serviço único ou poderia conter *Serviços Ouro, Prata e Bronze*, que reduz em qualidade. Outra classe de serviço forneceria serviços de pequeno atraso e baixa variação de atraso para aplicações tal como telefonia Internet e videoconferência. Companhias despejarão pagar um preço para executar uma videoconferência de alta qualidade para reduzir custos e tempo de trabalho. Finalmente, o Serviço Melhor Esforço permanecerá para clientes que requerem apenas conectividade.

Qualidade de Serviço e o aumento da largura de banda

A necessidade de fornecer mecanismos na rede para garantir a qualidade de serviço é um ponto muito debatido. Um grupo de pesquisadores considera que com o aumento da largura de banda (por exemplo, usando fibras óticas), a qualidade de serviço será automaticamente satisfeita. Mas outro grupo afirma que mesmo com o aumento da largura de banda, novas aplicações serão inventadas para fazer uso desta largura de banda. Portanto, mecanismos serão necessários para fornecer garantias de QoS.

Trabalhos da IETF relacionados com garantias de QoS

A *Internet Engineering Task Force* (IETF) tem proposto vários modelos de serviço e mecanismos para satisfazer a necessidade de QoS na Internet, proporcionando um melhor controle sobre o tráfego na Internet, na forma de priorização de certas aplicações (com certas restrições temporais) em detrimento do restante (tráfego essencialmente melhor esforço). Entre estes trabalhos estão o modelo Serviços Integrados/RSVP [Branden, 94] e o modelo Serviços Diferenciados [Black, 98]. Eles são descritos nas duas próximas seções.

7.4 Serviços Integrados/RSVP

Os serviços Integrados (*Integrated Services ou IntServ*) [Branden, 94] foram projetados para prover um conjunto de extensões ao modelo de entrega de tráfego de melhor esforço atualmente utilizado na Internet. Em essência, eles foram projetados para dar tratamento especial para certos tipos de tráfego e prover um mecanismo para que as aplicações possam escolher entre múltiplos níveis de serviços de entrega para seu tráfego.

IntServ é baseada na reserva de recursos. Para aplicações tempo real, antes dos dados serem transmitidos, as aplicações devem primeiro configurar caminhos e reservar recursos. RSVP (visto anteriormente) é um protocolo de sinalização para configurar os caminhos e reservar recursos.

7.4.1 Protocolos de Reserva de Recurso RSVP

Para fornecer garantias de QoS, técnicas de gestão de recursos devem ser usadas. Sistemas multimídia não podem fornecer QoS confiável aos usuários sem a gerência de recursos (ciclos de processamento de CPU, largura de banda da rede, espaço em buffer nos comutadores e receptores) nos sistemas finais, rede e comutadores. Sem a reserva de recursos, atrasos ou corte de pacotes devido a não disponibilidade de recursos necessários podem acontecer na transmissão de dados multimídia.

Para possibilitar esta reserva de recursos é necessária a existência de um protocolo de reserva de recurso na camada de Rede. Este tipo de protocolo na realidade não executa a reserva do recurso em si, ele é apenas o veículo para transferir informações acerca dos requisitos de recursos e usado para negociar os valores de QoS que o usuário deseja para suas aplicações. Para a reserva de recursos, os subsistemas devem prover funções de administração de recurso que forçam e escalonam acessos a recursos durante a fase de transmissão de dados.

O RSVP (*ReSource ReserVation Protocol*) [Zhang, 94] é um protocolo projetado para aumentar o suporte para aplicações tempo-real em redes IP. Ele permite a reserva de recursos em um caminho, mas a transmissão de dados é de responsabilidade do IP. Neste sentido, ele deve ser visto como um protocolo companheiro do IP. RSVP adotou a abordagem sem conexão. Assim, RSVP permanece compatível com a filosofia IP.

No RSVP é definida uma árvore de conexões com qualidade de serviço garantida. A inovação essencial do RSVP é que a qualidade de serviço não é especificada para a rede pelo emissor da informação, mas pelo receptor. A idéia é que o receptor está mais bem colocado que o emissor para saber que qualidade de serviço é necessária. Por exemplo, não há necessidade de que o emissor envie um fluxo de vídeo a 6Mbps para o receptor se ele não tem poder de processamento para decodificar e descompactar mais que 3Mbps. A implicação desta diferença é que RSVP é mais eficiente no uso de recursos da rede, pois é reservado o estritamente utilizável, além de permitir requisitos de receptor heterogêneos.

Resumidamente, o mecanismo de reserva do RSVP trabalha da seguinte forma:

- As aplicações fonte enviam regularmente mensagens especiais chamadas Path para um endereço multicast. Estas mensagens contêm a especificação do fluxo. Esta mensagem estabelece o estado Path nos agentes RSVP intermediários que é usado na propagação dos pedidos de reserva (feita pelos destinatários) para uma fonte específica.
- Na recepção da mensagem Path, cada receptor usa informações desta mensagem e informações locais (recursos computacionais, requisitos da aplicação, restrições de custo) para determinar a QoS. Em seguida, ele responde a mensagem path por uma mensagem Reservation especificando a qualidade de serviço requerida.
- A rede reserva recurso no caminho de retorno da mensagem Reservation para a aplicação fonte. Na passagem da mensagem Reservation, os agentes intermediários reservam recursos de rede ao longo do caminho e usam o estado Path estabelecido para propagar o pedido para o grupo emissor. A propagação da mensagem Reservation termina quando o caminho emenda em uma árvore de distribuição com recursos alocados suficientes para satisfazer os requisitos pedidos.

Quando a qualidade de serviço exigida por um receptor difere do fluxo emitido pela fonte, filtros de tráfego são usados para reduzir os requisitos de QoS nos agentes RSVP apropriados. Por exemplo, se um receptor é apenas capaz de apresentar imagens preto&branco e a fonte libera dados de imagens coloridas, um filtro será usado para remover os componentes de cor. Portanto, o estilo de reserva

iniciado pelo receptor acomoda requisitos heterogêneos dos receptores. Além de propósito da filtragem é preservar a largura de banda da rede.

O protocolo RSVP permite o *tunelamento*: se um nó IP intermediário entre dois nós RSVP não implementa o RSVP, ele pode retransmitir as mensagens RSVP. Mas neste caso, nenhum recurso será reservado no nó intermediário; assim, o caminho fim-a-fim terá uma ligação melhor esforço, e as garantias fim-a-fim determinística não será mais possível.

7.4.2 Classes de Serviços do IntServ

O modelo de Serviços Integrados propõe duas classes de serviço em adição ao Serviço Melhor Esforço:

- Serviço Garantido (Guaranteed Service) [RFC 2212]: fornece limites firmes (matematicamente prováveis) em termos de atrasos de enfileiramento que os pacotes sofrerão nos roteadores. Ele garante tanto o atraso quanto a taxa de bits. Basicamente uma sessão requisitando Serviço Garantido está requerendo que os bits em seus pacotes tenham uma taxa de bits garantida. Note que este serviço não tenta minimizar a variação de atraso, ele controla o atraso máximo de enfileiramento. Para este tipo de serviço, todos os nós intermediários devem implementar os serviços garantidos. Este serviço pode ser útil para aplicações requerendo fronteiras de atraso fixa, tal como aplicações tempo-real e aplicações de áudio e vídeo tempo-real.
- Serviço de Carga Controlada (Controlled Load Service) [RFC 2211]: uma sessão requerendo tal serviço receberá uma qualidade de serviço muito próxima da qualidade que um fluxo poderia receber de uma rede não sobrecarregada. Em outras palavras, a sessão pode assumir que uma "percentagem muito alta" de seus pacotes passará com sucesso através do roteador sem serem cortados e com um atraso de enfileiramento muito próximo a zero. Note que o Serviço de Carga Controlada não fornece garantias quantitativas acerca do desempenho – ele não especifica o que constitui uma "percentagem muito alta" de pacotes nem que qualidade de serviço aproximada será fornecida por um elemento de rede não sobrecarregado. Este tipo de serviço é dirigido para aplicações tempo-real adaptativas que estão sendo desenvolvidas hoje na Internet. Estas aplicações executam razoavelmente bem quando a rede não é sobrecarregada, mas elas se degradam rapidamente quando a rede se torna congestionada.

7.4.3 Problemas do IntServ

A arquitetura Serviços Integrados/RSVP representa uma mudança fundamental na arquitetura atual da Internet, que é fundada no conceito que todas as informações de estado relacionadas aos fluxos deveriam estar nos sistemas finais. Neste sentido, existem alguns problemas com a arquitetura Serviços Integrados [Xiao, 99]:

- O montante de informações de estado aumenta proporcionalmente ao número de fluxos. Isto causa uma sobrecarga de armazenamento e processamento nos roteadores. Portanto esta arquitetura não é escalável.
- Os requisitos nos roteadores são altos: todos os roteadores devem implementar RSVP, controle de admissão, classificação MF e escalonamento de pacotes.
- Para Serviço Garantido, toda a rede deve suportar IntServ. Uma instalação gradativa de Serviço de Carga Controlada é possível pelo emprego de funcionalidades RSVP e Serviço de Carga Controlada nos nós gargalos de um domínio e tunelando as mensagens RSVP para outras partes do domínio.
- IntServ/RSVP não são muito aplicáveis a aplicações do tipo navegadores WWW, onde a duração de um fluxo típico é apenas de poucos pacotes. A sobrecarga causada pela sinalização RSVP poderia facilmente deteriorar o desempenho da rede percebida pela aplicação.

7.5 Serviços Diferenciados

Devido as dificuldades de implementar e utilizar Serviços Integrados/RSVP, os Serviços Diferenciados (DS - *Differentiated Services ou DiffServ*) [Black, 98] foram introduzidos. Eles têm sido escolhido para ser implementado na Internet2. Neste modelo, os pacotes são marcados diferentemente para criar várias classes de pacotes. Pacotes de classes diferentes recebem diferentes serviços. O campo TOS (*Type Of Service*) do cabeçalho do pacote IPv4 e o campo *Class* do cabeçalho do pacote IPv6 podem ser setados

pelas aplicações para indicar a necessidade de serviço de pequeno atraso e alta vazão ou baixa taxa de perdas. Mas as escolhas são limitadas.

7.5.1 DS é um esquema de prioridades

A meta do DiffServ é definir métodos relativamente simples (comparados a IntServ) para prover classes diferenciadas de serviço para o tráfego na Internet. A ideologia é não pregar uma nova e completa arquitetura onipresente, mas produzir um pequeno e bem definido conjunto de blocos de construção dos quais uma variedade de serviços podem ser construídos. O mecanismo é que um pequeno padrão de bits, no campo TOS do IPv4 ou *Class* do IPv6, é usado para marcar um pacote para que ele receba um tratamento de encaminhamento particular, ou PHBs (*Per-Hop Behaviors*), em cada nó da rede. PHB é o comportamento observável externamente de um pacote em um roteador suportando DS. O enfoque do DiffServ é padronizar uma estrutura comum a ser usado para o campo TOS do IPv4 ou *Class* do IPv6, agora chamado de DS (*Differentiated Services*). Modificando os formatos definidos anteriormente pela IETF, este campo é definido em [Nichols, 98]:

- Seis bits do campo DS são usados como codepoint DSCP (Differentiated Service CodePoint) para selecionar o PHB que o pacote terá em cada nó. Este campo é tratado como um índice de uma tabela que é usada para selecionar um mecanismo de manipulação de pacotes implementado em cada dispositivo. Este campo é definido como um campo não estruturado para facilitar a definição de futuros PHBs.
- Um campo de dois bits é reservado (são ignorados por nós DS-conformantes).

Marcando os campos DS dos pacotes diferentemente, e manipulando pacotes baseados nos seus campos DS, várias classes de Serviços Diferenciados podem ser criadas. Portanto, Serviços Diferenciados é essencialmente um esquema de prioridades.

Quando aos serviços oferecidos por um domínio DS-conformante, devemos notar:

- Serviços DS são todos para tráfego unidirectional apenas.
- Serviços DS são para tráfegos agregados, não fluxos individuais.

[Black, 98] define um Serviço como o tratamento global de um subconjunto do tráfego do cliente dentro de um domínio DS-conformante ou fim-a-fim. O tráfego na rede hoje geralmente atravessa uma concatenação de redes que podem incluir hosts, redes residenciais e de escritório, redes corporativas/campus e várias redes de longa distância. Redes residenciais e de escritório são normalmente clientes de redes campus ou corporativas, que são por sua vez clientes de redes longa distância. Note que existem várias fronteiras cliente/provedor em que o conceito de serviço se aplica.

Afim de que os clientes recebam Serviços Diferenciados de seus Provedores de Serviço Internet (*ISP – Internet Service Provider*), eles devem firmar um Acordo de Nível de Serviço (*SLA – Service Level Agreement*) com seu ISP. Vários aspectos dos SLAs (como termos de pagamento) são fora do escopo de padronização; é a Especificação do Nível de Serviço (*SLS – Service Level Specification*) é que especifica as classes de serviços suportados e o montante de tráfego permitido em cada classe. Um SLA pode ser estático ou dinâmico. SLA estáticos são negociados mensalmente, anualmente, etc. Clientes com SLA dinâmicos devem usar um protocolo de sinalização (Por exemplo, RSVP) para pedir por serviços sob demanda.

Os clientes podem marcar os campos DS de pacotes para indicar o serviço desejado ou estes campos são marcados pelo roteador que liga o cliente à rede ISP (leaf router) baseado na classificação MF (multicampo).

No ingresso às redes ISP, os pacotes são classificados, policiados e possivelmente atrasados para torná-los conformes a algum perfil de tráfego pré-instalado. As regras de classificação, policiamento e retardos usadas nos roteadores de ingresso são derivadas a partir dos SLAs. O montante de espaço de bufferização necessário para estas operações também é derivado dos SLAs.

Um exemplo simples de perfil de tráfego poderia ser: medir o fluxo de pacotes do endereço IP a.b.c.d e se sua taxa fica abaixo de 200 kbps, sete o byte-DS para o valor X, senão sete o byte-DS para o valor Y. Se a taxa excede 600 kbps, corte os bytes excedentes. Os perfis são configurados pelo operador de acordo com o SLAs. Como os perfis são fornecidos (configuração manual ou sinalização) é fora do escopo do diffserv. Dentro da rede (nos roteadores internos ao domínio), o byte DS é usado para

determinar como os pacotes são tratados. O tratamento, também chamado de PHB ou comportamento agregado, pode incluir diferentes prioridades envolvendo atraso de enfileiramento (escalonamento), diferentes prioridades na decisão de descarte na sobrecarga de filas (gerenciamento de fila), seleção de rota, etc. De 2^6 possíveis significados (dados pelo campo DSCP), o grupo de trabalho DiffServ especificará (padronizará) alguns PHBs globalmente aplicáveis, e deixará o resto para uso experimental. Se os experimentos indicarem que um certo PHB não padronizado é claramente útil, ele pode ser padronizado posteriormente. Isto ainda está sob debate.

É de responsabilidade das ISPs decidir que serviços fornecer. Os seguintes serviços poderiam ser fornecidos:

- Serviço Premium, para aplicações requerendo serviço de pequeno atraso e pequena variação de atraso. Neste caso, o usuário negocia com seu ISP a máxima largura de banda para enviar pacotes através da rede e as alocações são feitas em termos de taxa de pico. Uma desvantagem é o fraco suporte a tráfegos em rajada e o fato de que o usuário paga mesmo quando não usa completamente a largura de banda.
- Serviço Assegurado, para aplicações requerendo melhor confiabilidade que Serviço Melhor Esforço. Este serviço não garante a largura de banda como o Serviço Premium, mas fornece uma alta probabilidade de que o ISP transfere os pacotes marcados com alta prioridade confiavelmente. Ele não foi completamente definido, mas oferece um serviço equiparável ao Serviço de Carga Controlada do IntServ;
- Serviço Olympic, que fornece três tipos de serviços: Ouro, Prata e Bronze, que reduz em qualidade.

Serviços Diferenciados é significativamente diferente de Serviços Integrados:

- Primeiro, há apenas um número limitado de classes de serviço indicados no campo DS. Desde que o serviço é alocado na granularidade de uma classe, o conjunto de informações de estado é proporcional apenas ao número de classes e não proporcional ao número de fluxos. Serviços Diferenciados é portanto mais escalável do que Serviços Integrados.
- Segundo, as operações de classificação, marcação, policiamento e retardo são apenas necessárias nas fronteiras das redes. Roteadores ISP internos (core) necessitam apenas implementar a classificação Comportamento Agregado (BA - Behavior Aggregate), que é uma classificação baseada apenas no byte DS. Portanto, Serviços Diferenciados é mais fácil de implementar e usar.

No modelo Serviços Diferenciados, um serviço assegurado pode ser fornecido por um sistema que suporta parcialmente os Serviços Diferenciados. Roteadores que não suportam Serviços Diferenciados simplesmente ignoram os campos DS dos pacotes e fornecem a pacotes serviço assegurado o Serviço Melhor Esforço. Desde que pacotes serviço assegurado têm menos probabilidade de serem cortados em roteadores compatíveis com DS, o desempenho total do tráfego serviço assegurado será melhor que o tráfego Melhor Esforço.