Capítulo II

Garantia de Qualidade de Serviço

II.1 Introdução

O acréscimo em desempenho propiciado pelos recentes avanços na tecnologia de transmissão de dados estimulou a demanda por novas aplicações com os mais diferentes requisitos de qualidade de serviço. As aplicações a seguir ilustram a variedade de requisitos a serem garantidos de forma a se obter uma qualidade de serviço aceitável.

O primeiro exemplo é a visualização dos dados de um satélite (tempestades, queimadas, camada de ozônio, etc.). O objetivo é ter um efeito similar ao obtido pela TV, compondo uma seqüência de quadros de imagens de alta definição (por exemplo 4 Moctetos). Esta aplicação permite mostrar, por exemplo, através de uma seqüência temporal, a evolução de uma tempestade; ou de uma seqüência espacial, a localização de pontos de queimada, de maior destruição da camada de ozônio, etc.. Para tanto no exemplo citado, é necessário uma banda passante que não somente forneça os 640 Mbps requeridos para uma taxa de 20 quadros por segundo, mas também garanta que a entrega destes dados não sofrerá nenhuma pausa ou atraso longo que poderia degradar a visualização em tempo real.

O segundo exemplo é uma classe de aplicações como telesseminários, teleconferência e outras formas de interação eletrônica entre um grande número de pesquisadores e organizações industriais (trabalho cooperativo). Para estas aplicações, espera-se que os requisitos de banda passante sejam bem menores (50 Mbps para cada canal de vídeo), mas com requisitos mais fortes em relação à confiabilidade, atraso e variação no atraso (*jitter*). De fato,

neste tipo de aplicação é bastante importante atingir não somente baixos atrasos na comunicação, mas também uma boa sincronização entre as mídias recebidas por todos os participantes de, por exemplo, um telesseminário.

Estas novas classes de aplicações tem a propriedade de requererem um sistema distribuído capaz de prover garantias sobre a qualidade de comunicação entre as aplicações. Aplicações multimídia, com seus fortes requisitos de alta velocidade e serviços de alta qualidade, têm estimulado o desenvolvimento de novas arquiteturas e protocolos de redes de comunicação.

Dentre os principais desafios das futuras redes de alto desempenho, está a garantia da qualidade de serviço (QOS) das aplicações em tempo real. A questão principal reside em prover uma enorme variedade de serviços com seus diferentes requisitos de desempenho, maximizando a possibilidade de admissão de novos usuários sem alterar as características das comunicações já existentes, otimizando a utilização dos recursos da rede e obviamente garantindo a qualidade de serviço negociada. Isto não requer somente altas bandas passantes, baixos atrasos e baixa taxa de erro (ou perda), mas também limites superiores e inferiores garantidos para estes índices de desempenho.

A capacidade de garantir um desempenho específico torna-se particularmente importante em redes suportando aplicações em tempo real. Entende-se por tempo real, um serviço que possibilita a um usuário especificar os seus requisitos de desempenho e obter garantias quanto ao cumprimento destes requisitos [Fer90]. A interação usuário-rede deve portanto ser bastante ampla e a sua interface bastante flexível de forma a possibilitar a especificação dos requisitos de qualquer tipo de fluxo. Com isso, novos modelos para redes são requeridos e propostos, onde tráfegos com diferentes requisitos de comunicação serão diferentemente gerenciados, de forma que aplicações com necessidades maiores sejam privilegiadas e protegidas do tráfego de dados usual que flui nas redes de computadores atuais. O conceito de usuário adotado neste trabalho se refere tanto ao usuário final de uma aplicação

quanto a qualquer entidade usuária de um determinado serviço oferecido por uma entidade provedora do mesmo.

No entanto, persiste a dúvida de quais os tipos de requisitos de comunicação uma aplicação em tempo real deve ter. Uma resposta precisa a esta questão é muito difícil devido ao alto número destas aplicações que foram, têm sido e serão propostas na literatura. Enquanto que algumas aplicações, por exemplo, distribuição de vídeo, ainda que bastante exigentes em termos de banda passante, podem admitir alguma degradação na confiabilidade da entrega e nos atrasos na comunicação, outras, como vídeo-conferência e trabalho cooperativo (com caráter interativo) impõem fortes requisitos na qualidade de serviço geral. Existe, porém, um consenso de que a especificação dos requisitos de comunicação deve ser flexível o suficiente de forma a abranger ao máximo as necessidades das mais diversas, atuais e futuras, aplicações em tempo real.

Este capítulo identifica os parâmetros considerados a priori necessários e suficientes para descrever qualquer requisito de fluxo, enfocando tanto a interface quanto as formas de interação usuário-rede e descreve brevemente a política adotada por diversos protocolos propostos na literatura. A seção II.2 descreve as principais características a serem especificadas e garantidas: vazão, atraso, confiabilidade e nível de garantia de serviço. A seção II.3 apresenta e analisa as formas de negociação usuário-rede.

II.2 Caracterização da Qualidade de Serviço

Para caracterizar uma comunicação de tempo real normalmente faz-se uso de uma entidade abstrata denominada fluxo (*flow*). Um fluxo é uma cadeia de unidades de informação (pacotes, células, octetos, ...) com os mesmos requisitos de desempenho que segue uma rota simplex fixa entre a fonte e o destino.

Dentre as propostas de requisitos para a caracterização da qualidade de serviço citadas na literatura, a mais geral é a *flowspec* (*flow specification*) proposta em [Par92], onde são apresentados os parâmetros considerados necessários e suficientes para especificar qualquer requisito de fluxo. Nesta seção, à medida que os índices de desempenho forem apresentados, serão descritos os respectivos parâmetros da *flowspec*.

A qualidade do serviço de uma aplicação em tempo real é determinada pelos valores de alguns parâmetros de desempenho fim-a-fim. Tais parâmetros podem ser especificados em termos de três índices de qualidade e do nível de garantia da qualidade de serviço:

- **vazão do tráfego**: este índice está relacionado com a necessidade de caracterizar como o tráfego do fluxo será injetado na rede. Ele expressa a quantidade de dados que será enviada através da rede por unidade de tempo. Geralmente especifica as necessidades do tráfego da comunicação em termos da banda passante requerida;

- **atraso**: este índice de qualidade está relacionado com a sensibilidade das aplicações em relação ao atraso introduzido pela rede. Algumas aplicações são mais sensíveis que outras tanto em relação ao atraso quanto à variação no atraso;

- **confiabilidade**: este índice está relacionado com a sensibilidade das aplicações em relação à perda de dados. Este índice de qualidade é fortemente dependente do mecanismo de armazenamento e reserva de recursos envolvido na transmissão de dados ao longo da rede;

- finalmente, é esperado que a rede seja capaz de prover uma faixa de garantias de serviços: garantias fortes (perfeitas), suaves (imperfeitas) ou preditivas (probabilísticas). O **nível de garantia de qualidade de serviço** negociado e selecionado é fundamental para um bom desempenho da rede, pois ele influi nas probabilidades de sucesso na admissão de um determinado número de aplicações e na otimização da utilização dos recursos da rede.

II.2.1 Vazão do Tráfego

A necessidade de caracterizar o tráfego da comunicação (doravante chamado de fluxo) que será injetado na rede é primordial para se obter garantias, principalmente em relação à banda passante. Estas características do tráfego podem ser especificadas, por exemplo, pela taxa de transmissão de um pacote ou pelo intervalo entre pacotes, podendo ser referentes ao pico ou à média em um intervalo de tempo específico.

Além de representar um requisito do usuário, a especificação da vazão pode ser usada pela rede para determinar a quantidade de tráfego que será produzida por um determinado usuário. Neste sentido, este índice pode ser interpretado pela rede como um compromisso do usuário de obedecer a certas restrições de tráfego.

A *flowspec* caracteriza o tráfego injetado através de quatro parâmetros: o comprimento máximo do pacote, a taxa de transmissão média (taxa de recebimento de créditos), a taxa de pico (taxa de transmissão máxima) e o tamanho de rajada (número de créditos máximo), conforme descritos a seguir.

- A **Unidade de Transmissão Máxima** (MTU - *Maximum Transmission Unit*) é o comprimento máximo em número de octetos do maior pacote passível de ser transmitido neste fluxo. Este campo serve para dois propósitos: limita a quantidade de tempo que um fluxo pode transmitir, ininterruptamente, num meio compartilhado e é uma unidade conveniente para expressar as propriedades referentes às perdas. Usar a MTU original da rede, por exemplo 64 koctetos do IP, seria inapropriado, pois as aplicações podem ser sensíveis a perdas de quantidades de dados bem menores.

- A **Taxa de Recebimento de Créditos** (*Token Bucket Rate*) é a taxa com que os créditos (*tokens*) de um dado fluxo são inseridos no seu "depósito imaginário de créditos" e

corresponde à taxa de transmissão média. Para enviar um pacote, o transmissor deve retirar do depósito um número de créditos igual ao tamanho do pacote. Se não houver créditos suficientes o transmissor deve esperar até que uma quantidade suficiente de créditos esteja armazenada no depósito.

- O **Número de Créditos Máximo** (*Token Bucket Size*) controla a máxima quantidade de dados que um fluxo pode enviar à taxa máxima, correspondendo ao tamanho da rajada máxima. O depósito de créditos é preenchido à taxa de entrada de créditos e o número máximo de créditos limita quantos créditos um fluxo pode armazenar. Quando o depósito está cheio, os novos créditos depositados são descartados.

- A **Taxa de Transmissão Máxima** limita o quão rápido os pacotes são enviados através do mecanismo descrito a seguir. Considere que o depósito está cheio. Se o número máximo de créditos for grande, o tempo de transmissão dos pacotes poderá ser longo o suficiente para inibir a multiplexação com os demais fluxos. Pode-se pensar o controle da máxima taxa de transmissão como um "vazamento no depósito de transmissão". Quando um pacote é enviado, um número de créditos igual ao tamanho do pacote é colocado num depósito vazio que drena créditos à taxa de transmissão máxima. Nenhum pacote poderá ser enviado até que este depósito esteja novamente vazio. Portanto, a taxa de transmissão máxima é a taxa na qual este depósito é esvaziado. O seu valor deve ser maior ou igual a taxa de entrada de créditos. A MTU especifica o tempo máximo que um pacote individual pode consumir e a taxa de transmissão máxima limita a freqüência na qual os pacotes podem ser injetados na rede. Assim, o tamanho da MTU usado em conjunção com a taxa de transmissão máxima limita o tempo que pacotes individuais bloqueiam outras transmissões.

II.2.2 Sensibilidade ao Atraso

O atraso pode ser expresso tanto em termos de um valor absoluto quanto de um valor probabilístico. Além do limite para o atraso, um limite na variação do atraso, ou *jitter*, pode também ser especificado. Dependendo dos protocolos de tempo real usados, estes limites podem ser tanto requisitos de qualidade especificados pelo usuário, quanto uma medida feita pela rede do desempenho do serviço de tempo real existente no momento. No primeiro caso, estes índices são utilizados pela rede para avaliar a possibilidade de uma conexão e a quantidade de recursos requeridos para o seu estabelecimento. No segundo caso, são o resultado da avaliação das capacidades disponíveis na rede a serem submetidos ao usuário, que pode aceitar ou recusar a QOS proposta.

Conforme a sensibilidade das aplicações ao atraso e à variação no atraso, a rede deve, entre outras coisas, escolher entre rotas com diferentes características de atraso, por exemplo, tanto rotas usando satélites (que possuem longos atrasos) quanto rotas usando linhas terrestres (que possuem atrasos menores) podem possibilitar a comunicação entre a fonte e o destino.

De acordo com Partridge [Par92], existem dois problemas em permitir às aplicações especificarem um atraso máximo aceitável. Primeiro, observe que uma aplicação pode estar bastante satisfeita com um atraso de 100 ms entre o Rio de Janeiro e Tóquio, porém, bastante insatisfeita com um atraso de 100 ms numa comunicação local, dentro de uma mesma cidade. Esta observação sugere que o atraso máximo é de fato variável e função do atraso considerado atingível. No entanto, o atraso atingível é fortemente determinado pela distância geográfica entre os pares comunicantes, e este tipo de informação geográfica, em geral, não está disponível para a rede. Além disso, com o advento dos sistemas de comunicação móveis, este tipo de informação se torna bastante difícil de se obter. Portanto, existem razões para acreditar que as aplicações devam ter sérias dificuldades em escolher um atraso máximo considerado razoável, tanto para o usuário quanto para a distância geográfica considerada.

O segundo problema com atrasos máximos é que eles tentam quantificar o desempenho que é aceitável para o usuário, e uma aplicação em geral não sabe qual desempenho será aceitável para o seu usuário. Por exemplo, uma justificativa comum para especificar um máximo atraso aceitável é que considera-se humanamente difícil conversar com atrasos maiores que 100 ms. Certamente tal atraso pode tornar a comunicação menos agradável, mas é ainda possível se conversar mesmo com atrasos de vários segundos de duração, e dada a escolha entre a não conexão e um atraso longo, muitos usuários escolheriam o atraso longo.

Como parte da especificação da sensibilidade ao atraso do fluxo, deve-se também caracterizar o quanto o fluxo é sensível à variação no atraso (*jitter*). Pacotes injetados na rede de acordo com um determinado padrão de tráfego, normalmente não saem conforme o mesmo padrão. O padrão original é distorcido pelos efeitos das filas e da multiplexação nos nós da rede. Portanto, espera-se que as aplicações sensíveis a este tipo de distorção requeiram que seus receptores reservem uma certa quantidade de *buffers* para restaurar o fluxo à sua forma original.

A quantidade de *buffer* requerida para remover a distorção no sistema receptor é fortemente determinada pela variação do atraso na transmissão fim-a-fim. Se o atraso na transmissão é um atraso médio (D), mais ou menos uma variação (V) e a taxa máxima de transmissão é R, o sistema receptor necessitará de um *buffer* de tamanho $2 \times V \times R$. Para entender esta equação considere dois pacotes, A e B, enviados com T unidades de tempo de intervalo. No pior caso, A chegará depois de um atraso de $D - V$ unidades de tempo (atraso mínimo) e B chegará depois de um atraso $D + V$ unidades de tempo (atraso máximo). O receptor não pode entregar B até que ele chegue, o que no máximo ocorrerá em $T + 2 \times V$ unidades de tempo depois de A. Para assegurar que A seja entregue T unidades de tempo antes de B, A deve permanecer no *buffer* por $2 \times V$ unidades de tempo.

A *flowspec* caracteriza a sensibilidade ao atraso através de dois parâmetros: o menor atraso perceptível e a variação no atraso máxima, conforme descritos a seguir.

- O **Menor Atraso Perceptível** (*Minimum Delay Noticed*) serve apenas como um aviso à rede de quando parar de tentar reduzir o atraso, indicando que a aplicação é insensível a melhoramentos no atraso fim-a-fim abaixo deste valor. A rede é encorajada a manter o atraso próximo a este valor, mas não necessita tentar diminuí-lo. De acordo com o valor deste campo a rede pode optar por evitar ou não enlaces de satélites.

- A **Variação no Atraso Máxima** (*Maximum Delay Variation*) indica se aplicação receptora requer que os dados a serem entregues estejam com o mesmo padrão de tráfego que foram transmitidos. Este campo permite ao receptor indicar a quantidade de *buffer*, necessário para restaurar o padrão de transmissão, que ele deve reservar para este fluxo. Um exemplo simples disto é o caso em que uma aplicação deseja enviar dados com amostras de voz, geradas a intervalos regulares e que assim devem ser reproduzidas. Estes intervalos regulares podem ser distorcidos pelo efeito das filas e da multiplexação nos nós da rede e o receptor deve ser capaz de restaurar o espaçamento entre pacotes original. A quantidade de *buffer* limitada pelo receptor para o fluxo determina a máxima variação no atraso permitida. Os implementadores devem notar que a taxa de transmissão máxima pode causar uma distorção no padrão do tráfego antes mesmo dos pacotes serem injetados na rede, e esta distorção deve ser levada em consideração para determinar o tamanho do *buffer* do receptor.

II.2.3 Sensibilidade a Erros

A sensibilidade de cada aplicação à perda de dados é diferenciada, sendo algumas pouco sensíveis (no entanto sensíveis a perdas consecutivas, por exemplo aplicações de voz), enquanto outras requerem um nível de confiabilidade de 100 % de que os dados sejam transferidos e entregues na ordem correta e livres de erros (por exemplo transmissão de dados bancários, transferência de arquivos, etc.). Algumas aplicações em tempo real são por vezes sensíveis a perdas e também impossibilitadas de esperar pela retransmissão dos dados. Para

estas aplicações particularmente sensíveis, os protocolos devem implementar o *forward error correction* (FEC) para tentar minimizar a perda.

Em redes de comutação de pacotes, os pacotes recebidos pelos nós intermediários são armazenados em *buffers* até serem transmitidos para o próximo nó. Como o tamanho dos *buffers* é limitado pode ocorrer, devido à congestionamento, a perda de pacotes. Um limite probabilístico em tais tipos de perdas pode ser usado como medida da confiabilidade do serviço. O seu valor irá influenciar na quantidade de recursos requeridos para o estabelecimento da conexão. De maneira similar ao atraso, a confiabilidade pode ser um parâmetro especificado pelo usuário ou um parâmetro medido pela rede.

A *flowspec* caracteriza a sensibilidade a perda de dados através de três parâmetros: a sensibilidade a perdas, a sensibilidade a perdas consecutivas e o intervalo de perdas, conforme descritos a seguir.

- A **Sensibilidade a Perdas** expressa o número máximo de perdas de pacotes de tamanho MTU aceitável em um dado intervalo de tempo. Este campo pode também ser usado para, por exemplo, sugerir que a rede selecione a rota de transmissão com menor taxa de erro (ou perda).

- A **Sensibilidade a Perdas Consecutivas** indica o número máximo de pacotes de tamanho MTU consecutivos que podem ser perdidos.

- O **Intervalo de Perdas** determina o período no qual o máximo número de perdas é medido. Em outras palavras, dado um intervalo arbitrário, o número de perdas neste intervalo não pode exceder o número indicado no campo de sensibilidade a perdas.

II.2.4 Nível de Garantia de Qualidade de Serviço

Finalmente, é desejável que a rede seja capaz de prover uma faixa de garantias de qualidades de serviços. Os parâmetros que especificam os índices de desempenho mencionados nas subseções anteriores podem ser usados em separado ou em diversas combinações na especificação da qualidade de serviço. Como visto, em alguns casos são especificados pelo usuário, em outros são avaliados pela rede e propostos ao usuário para aprovação. Em ambos os casos, eles devem ser considerados pela rede como um compromisso para prover o serviço de acordo com os parâmetros de desempenho especificados. Entretanto, um problema crucial continua sem solução: como fazer para que este compromisso seja forte, e quais as garantias que o usuário deve receber do serviço provedor (da rede).

A maioria das soluções propostas durante os últimos anos tentaram resolver este problema usando uma dentre duas diferentes abordagens:

- **garantia forte**: a rede se compromete a oferecer um serviço, cuja qualidade é precisamente especificada através de um número de parâmetros de desempenho e de tráfego. Neste caso, um contrato é requerido entre o usuário e a rede, no qual o usuário especifica as características do seu tráfego e se compromete a respeitá-las enquanto durar o contrato; e a rede promete prover o serviço de acordo com os requisitos do usuário. Para proteger o garantido nível de qualidade dos efeitos dos usuários "mal comportados" e de flutuações na carga da rede, mecanismos de reserva de recursos, controle de admissão do tráfego e controle da taxa tem que ser reforçados. O serviço não é provido, e logicamente o fluxo não é estabelecido, se os recursos disponíveis e as características do tráfego do usuário não permitirem que o serviço esperado seja garantido;

- **garantia suave**: neste caso, quando um usuário requer um serviço de tempo real, a rede avalia a carga do tráfego corrente e as características do tráfego do usuário. Tais

características não têm que ser referentes a situação de pior caso como na abordagem de garantia forte, mas podem ser descritoras da média do comportamento do usuário. Com estes dados a rede avalia o nível de qualidade atingível e submete ao usuário para aprovação. Como este nível de qualidade depende da hipótese de que a carga da rede vai corresponder à situação corrente, e que os usuários atuais e futuros não vão modificá-la, nenhum compromisso forte pode ser feito pela rede. De fato, sempre que as condições de carga da rede mudarem, a qualidade do serviço provido também mudará; isto requer usuários adaptativos, usuários capazes de tolerar e compensar flutuações, e mesmo rompimentos temporários, no serviço da rede.

A *flowspec* caracteriza o nível da garantia da qualidade de serviço através dos parâmetros de qualidade da garantia descrito a seguir.

- A **Qualidade da Garantia** especifica o tipo de garantia a ser oferecido pela rede. Existem duas questões relacionadas a garantias. Primeiro, é importante considerar os casos em que não é possível para a rede oferecer garantias de QOS fortes. O exemplo clássico é o da rede Ethernet que freqüentemente oferece um desempenho aceitável, mas com certeza a rede não pode garantir que a Ethernet não vai saturar durante o período da conexão. A segunda questão é referente a alguns trabalhos que sugerem que algumas aplicações serão capazes de se adaptar a modestas variações no desempenho da rede e, com isso, os projetistas poderão explorar esta flexibilidade para obter uma melhor utilização dos recursos da rede. Neste modelo, a rede poderia suavemente se desviar dos parâmetros prometidos durante os períodos de sobrecarga. Esta classe de serviço é conhecida como serviço preditivo (distinto do serviço garantido).

A diferença entre serviço preditivo e serviço que não pode ser perfeitamente garantido (como no exemplo da Ethernet) é que garantias imperfeitas não fazem nenhuma promessa estatística sobre o seu "mal comportamento".

Estão previstas atualmente 6 formas de especificar a qualidade da garantia desejada, conforme descrito a seguir:

- nenhuma garantia é requerida. Neste caso, a aplicação está simplesmente expressando os parâmetros de desempenho desejado para o seu fluxo;

- uma garantia imperfeita é requerida;

- um serviço preditivo é requerido e se não for possível, então o fluxo não deve ser estabelecido;

- um serviço preditivo é requerido, mas uma garantia imperfeita é aceitável;

- um serviço garantido é requerido e se uma garantia forte não for possível, então o fluxo não deve ser estabelecido;

- um serviço garantido é requerido, mas uma garantia imperfeita é aceitável.

Espera-se que optando-se por um serviço preditivo, ou aceitando-se uma garantia imperfeita, aumentam-se substancialmente as possibilidades de o fluxo ser estabelecido.

II.3 Controle de Admissão de Conexões

Vários protocolos têm sido apresentados na literatura que permitem a uma aplicação obter serviços com garantias de desempenho. A maioria das soluções propostas são baseadas numa abstração da comunicação consistindo de uma conexão (ou fluxo) com características de desempenho especificáveis. O estabelecimento de uma conexão e os recursos associados a ela são controlados pela rede através de uma política de admissão. Em alguns esquemas, a admissão de uma nova conexão depende da disponibilidade de uma suficiente quantidade de

recursos nos nós da rede, para prevenir esta nova conexão de interferir no desempenho das conexões já estabelecidas.

Uma das formas de um sistema distribuído prover garantias é através da negociação do usuário com a rede pelo direito de utilizar certa parte dos recursos da rede. Uma outra alternativa é a rede estimar as necessidades do serviço de comunicação através das características do seu tráfego de dados. Atualmente ainda não está claro como fazer para este último esquema funcionar, exceto para um limitado conjunto de classes de tráfego. Existem várias formas de se efetuar a negociação usuário-rede. Por exemplo, uma negociação pode ser feita *in-band* ou *out-of-band*. Pode também ser feita antecedendo o envio de dados (possivelmente com dias de antecedência), como a primeira parte de uma fase de conexão, ou concorrentemente com o envio. Neste caso a conexão é dita implícita e o envio dos dados é feito simultaneamente com a negociação para assegurar a continuidade desta comunicação.

A maioria dos protocolos para tempo real propostos na literatura incluem um esquema bastante simples de interação entre o usuário e a rede. Um usuário de uma comunicação tempo real possui em geral um conjunto de requisitos de desempenho que refletem a qualidade de serviço que ele espera obter da rede. No caso dos protocolos que fornecem uma garantia forte para a qualidade de serviço, a rede pode ser considerada como um servidor de comunicação que, se existirem recursos suficientes disponíveis, é capaz de satisfazer completamente os requisitos do usuário por toda a duração da conexão.

Em um modelo simples de interação usuário-rede, primeiro o usuário envia um pedido ao servidor de comunicação para estabelecer a conexão que melhor se adapta ao seu tráfego e aos seus requisitos de desempenho. Em seguida a rede avalia o pedido e responde através de uma mensagem de conexão estabelecida ou conexão rejeitada. No primeiro caso, o servidor foi capaz de reservar os recursos necessários para o estabelecimento da conexão e assegurar a requerida qualidade de serviço, habilitando o usuário a enviar pacotes através do canal estabelecido. No caso da conexão ser recusada, uma breve descrição da causa da rejeição é

provida, especificando o recurso cuja deficiência resultou na rejeição. Este tipo de informação não ajuda muito o usuário sobre o que fazer, já que este não possui dados (quantitativos) suficientes para formular um novo pedido aceitável. Em alguns casos a rede pode estar muito perto de estabelecer a conexão, em outros, as condições de carga correntes praticamente impossibilitam a aceitação de quase todos os pedidos.

Neste caso, o usuário possui três possibilidades. A primeira é abandonar o pedido e tentar novamente mais tarde, esperando que nesta hora seus pedidos sejam aceitos. A segunda abordagem é repetir o mesmo pedido até que seja aceito. Além de aumentar o fluxo do tráfego relacionado à estabelecimentos na rede, esta opção obriga o usuário a esperar, desperdiçando parte dos seus recursos computacionais. Como rejeições no estabelecimento do canal ocorrem em situações de alta carga da rede, esta abordagem não parece ser muito apropriada. A terceira possibilidade é o usuário modificar o seu pedido, se isto for aceito pela sua aplicação, alterando alguns valores dos seus requisitos e submetendo-o novamente a rede, especialmente se informações adicionais explicando (e quantizando) a razão para a rejeição forem incluídas na resposta da rede, fornecendo ao usuário informações detalhadas sobre o estado corrente da rede. Uma vez a conexão estabelecida, não existe nenhuma outra forma de interação entre o usuário e a rede, exceto quando o usuário eventualmente requisitar a liberação da conexão.

A simplicidade deste modelo pode levar a algumas ineficiências na utilização dos recursos da rede e a um grande número de pedidos rejeitados. Em alguns casos, a rede poderia acomodar o pedido com pequenas alterações no desempenho ou nos parâmetros do tráfego, porém o usuário, em geral, não possui informação suficiente sobre o estado da rede para modificar os seus requisitos de acordo. Em outros casos, uma situação de saturação poderia facilmente ser resolvida modificando as características das conexões existentes.

Uma das formas de controle de admissão e o conseqüente estabelecimento de uma nova conexão é descrita a seguir em quatro passos:

1. o usuário determina os parâmetros da rede que caracterizam uma conexão a ser estabelecida com um destino específico. Assume-se que o usuário não possui, em geral, nenhuma informação sobre o estado da rede;
2. a rede recebe o pedido e o avalia. Esta avaliação não é centralizada em um nó particular, mas é executada através de pedidos a cada nó no caminho selecionado entre a fonte e o destino. Se um nó tem condições de aceitar esta nova conexão sem comprometer o desempenho das conexões já existentes, uma adequada quantidade de recursos é condicionalmente reservada para este novo canal;
3. o pedido segue em direção ao destino até que um dos seguintes eventos ocorra: (i) um dos nós no caminho não pode acomodar este novo canal, (ii) o destino é atingido com sucesso e pode acomodar este novo canal. No caso (i), o nó envia uma resposta negativa à origem. No segundo caso, o destino envia uma resposta positiva de volta à origem;
4. começando do último nó atingido, a mensagem de resposta segue em direção à origem, seguindo o mesmo caminho da mensagem de pedido. Se a resposta for uma rejeição, os recursos reservados são liberados, caso contrário são confirmados e o canal é estabelecido.

Apesar dos diversos esforços em pesquisa feitos nos últimos anos para prover garantias de qualidade de serviço em redes em tempo real, ainda não há um consenso em quais devem ser as características e mecanismos de tal rede. A seguir é feita uma breve descrição da política adotada por alguns protocolos propostos na literatura.

No *ReSource Reservation Protocol* (SRP) [AHS90], os usuários especificam os seus requisitos através de três parâmetros de tráfego e um índice de desempenho. Os parâmetros são respectivamente o tamanho máximo da mensagem, a taxa máxima da mensagem e o tamanho máximo da rajada. O índice de desempenho é a especificação do máximo valor para o atraso de transmissão fim-a-fim. Duas classes de serviços são oferecidas: a garantida e a *best-effort*.

O *Stream Protocol Version II* (ST-II) [Top90] permite que o usuário especifique os seus requisitos de tráfego através de um conjunto de parâmetros. A partir destes dados e da carga do tráfego momentânea, a rede avalia a vazão e o atraso realmente alcançáveis e os propõe ao usuário. A vazão é expressa pelos tamanho máximo do pacote e taxa de transmissão permitidos, enquanto o atraso é representado pela média e pela sua variação. Este protocolo visa garantir a média da taxa de transmissão (através da reserva de banda passante) de cada comunicação, e reserva recursos adicionais da rede para compartilhar com outras conexões de modo a acomodar eventuais rajadas. Portanto, seus mecanismos oferecem uma garantia de QOS suave.

O *Resource ReSerVation Protocol* (RSVP) [ZDE93] é bastante similar ao ST-II, sendo a sua principal característica o fato de, diferentemente dos demais, ser orientado ao receptor. O objetivo é prover a capacidade de receptores heterogêneos reservarem recursos especificamente ajustados às suas necessidades. É portanto um protocolo de reserva de recursos para aplicações multiponto, bastante flexível de forma a se adaptar facilmente a mudanças nas listas de participantes (por exemplo numa conferência multimídia), possibilitando filtragem de pacotes, e ainda assim oferecendo garantias fortes no que se refere a banda passante da comunicação.

Uma outra abordagem para obtenção de garantias fortes é proposta pelo Tenet Group [FeV90]. O usuário especifica os seus requisitos de vazão em termos do mínimo e da média de tempo entre pacotes, sendo esta média especificada pelo usuário. Os requisitos de atraso podem ser especificados em termos de um limite superior absoluto no atraso de transmissão (serviço determinístico) ou de um limite probabilístico (serviço estatístico), ou de um limite superior no *jitter* (serviço com *jitter* limitado) [VZF91]. Um requisito de confiabilidade na comunicação também pode ser especificado através de um limite probabilístico na perda de pacotes.

Uma solução diferente é proposta no *Flow Protocol* [Zha91], onde um algoritmo de controle do tráfego é introduzido para controlar a taxa de transmissão do fluxo de dados. O usuário pode especificar o seu tráfego por meio de uma taxa de transmissão média e um intervalo médios, de forma que a rede possa reservar os recursos requeridos para satisfazer as necessidades do usuário. O algoritmo proposto tem também um efeito benéfico no atraso que os pacotes irão sofrer ao longo da rede. Entretanto, nenhuma garantia é dada em termos de um limite no atraso da comunicação. Por estas razões o *Flow Protocol* pode ser considerado como oferecendo garantias fortes apenas no que concerne a vazão do tráfego.

Em [CSZ92] uma classe particular de aplicações multimídia, chamadas aplicações *playback*, é apresentada. Os usuários destas aplicações são capazes de se adaptar a variações na qualidade de serviço provida pela rede e mesmo suportar rompimentos temporários no serviço. Para este tipo de aplicação, um novo tipo de serviço tempo real é introduzido, chamado de preditivo. Neste serviço, a rede tenta entregar um serviço que satisfaça os pedidos do usuário, reservando uma quantidade de recursos relacionados a recente medida da carga do tráfego. Entretanto, quando as condições do tráfego mudam, a qualidade de serviço provida geralmente também muda.

A abordagem de garantia suave provavelmente apresenta uma importante vantagem quando comparada a soluções que apresentam um serviço fortemente garantido: como as computações do protocolo e reservas são feitas com base na situação corrente do tráfego, a quantidade de recursos a ser reservado a cada usuário nos nós da rede podem ser menores. Isto permite o estabelecimento de um número maior de comunicações e um melhoramento na utilização dos recursos da rede. Isto ocorre devido ao fato de os usuários de um determinado tipo de serviço serem capazes de suportar reduções e interrupções no serviço, como conseqüência da admissão de um novo tráfego, mais exigente, ou de mudanças não previstas no comportamento de algum outro usuário.

A validade da abordagem de garantia suave conta principalmente com a existência da classe particular de usuários e aplicações multimídia que suportam serviços preditivos. De fato, é verdade que algumas aplicações tem requisitos menos restritivos que outras, e possuem um padrão de tráfego bem regular. Entretanto, isto não significa necessariamente que tais aplicações e seus usuários aceitarão um serviço que não é garantido em todas as condições.

O nível de flexibilidade na interação usuário-rede disponível nos protocolos de tempo real atuais é insatisfatório, o que leva à ineficiência na utilização dos recursos da rede. Os futuros protocolos devem melhorar a sua capacidade de negociar a qualidade de serviço com a rede, permitindo a negociação do serviço de comunicação que melhor se ajuste às suas necessidades, e o aumento da probabilidade de ter tal serviço estabelecido.

Uma alternativa é explorar as características dos usuários flexíveis. Um usuário é dito flexível quando aceita uma faixa de qualidades de serviço. O usuário indica a sua flexibilidade especificando faixas de valores aceitáveis para um ou mais componentes do conjunto de parâmetros de desempenho e tráfego. Além do mais, a rede pode tirar proveito de uma maior quantidade de informação que um usuário provê no momento do estabelecimento de uma conexão. Para uma determinada aplicação, um usuário pode especificar como um índice de desempenho uma faixa de valores aceitáveis em vez de um único valor. A rede poderia portanto, escolher o valor que melhor se ajuste à situação atual.

As novas aplicações exigem um gerenciamento dinâmico das comunicações. Um usuário possui um comportamento dito estático se, uma vez estabelecido o seu fluxo, os seus requisitos de QOS se mantém constantes. Entretanto, um gerenciamento dinâmico pode ser feito no caso dos usuários flexíveis, onde a rede teria a liberdade de reconfigurar os parâmetros do fluxo, conforme a faixa de valores aceitável pelo usuário, em determinadas condições do tráfego (por exemplo saturação dos recursos), ou onde o usuário poderia solicitar que a rede adaptasse o seu canal às novas condições de desempenho solicitadas pela aplicação.

Tanto flexibilidade quanto configurações e comportamentos dinâmicos são requisitos fundamentais nas aplicações multimídia e são sempre abordados nos estudos visando garantias de qualidade de serviço. No entanto, não se pode deixar de considerar que quanto mais flexível o serviço, mais complexo se torna o protocolo, aumentando o processamento dos sistemas de comunicação e limitando portanto o desempenho atingível para uma determinada plataforma.