2
Mecanismos de Controle de Admissão

As disciplinas de Controle de Admissão representam um ponto crucial para o suporte a QoS. São mecanismos de controle de congestionamento preventivo, e objetivam regular a entrada de usuários no domínio em questão. Controlando a entrada de usuários, o mecanismo de Controle de Admissão limita o volume de tráfego no domínio, de forma a satisfazer os parâmetros de QoS especificados, como taxa de descartes e retardo de pacotes.

O objetivo de um controle de admissão eficiente é a manutenção da Qualidade de Serviço previamente acordada para os usuários já admitidos, maximizando ao mesmo tempo, a utilização da rede. Em outros termos, o algoritmo utilizado deve maximizar o número de conexões admitidas sem prejuízo da Qualidade de Serviço das conexões existentes. Existem diversos algoritmos que realizam esse gerenciamento de entrada de usuários, considerando a ocupação dos recursos da rede e as necessidades dos novos usuários, podendo ser divididos em 2 grupos: Algoritmos de alocação não-estatística e de alocação estatística.

PUC-Rio - Certificação Digital Nº 0210416/CB

2.1
Algoritmos de alocação não-estatística

Os algoritmos de alocação não-estatística são aqueles que fazem a alocação de usuários pela taxa de pico do fluxo solicitante. A alocação pela taxa de pico é recomendada para fluxos nos quais a garantia de desempenho é absolutamente inviolável, ou para fluxos que operem a taxas constantes (CBR), como audio/vídeo não-comprimidos, telemetria, etc. Para fluxos com variações significativas na taxa (o que é muito comum), a reserva de recursos pela taxa de pico causará um enorme desperdício de banda e conseqüente subaproveitamento da rede. Por serem algoritmos extremamente simples, exigem pouquíssimo processamento por parte da unidade de controle de admissão.

Em geral, o super-dimensionamento para a reserva de recursos é compensado pelo baixo tempo de processamento dos algoritmos. No entanto, quando o tráfego é caracterizado por grandes variações na taxa (tráfego em rajadas), em que grandes períodos de silêncio são alternados com períodos de grande volume informações a transmitir, o super-dimensionamento poderá não ser vantajoso. Nesses casos, são utilizados algoritmos de alocação estatística.

2.2 Algoritmos de alocação estatística

São algoritmos que buscam um melhor aproveitamento dos recursos disponíveis, considerando que a grande maioria das aplicações, além de alternarem entre períodos de atividade e inatividade, têm sua taxa de transmissão de dados variando continuamente nos períodos ativos. Em aplicações como audio/vídeo comprimido (MPEG, ...), por exemplo, a taxa média de bits é significativamente menor que a taxa de pico e, nesse caso, alocar recursos pela taxa de pico seria um enorme desperdício. Uma opção para esse problema é alocar recursos considerando-se uma taxa maior que a taxa média, porém menor que a taxa de pico. Como resultado, temos um processo chamado multiplexação estatística: a soma de todas as taxas de pico das conexões aceitas pode ser maior que a capacidade do link em questão, mas como a taxa de pico não é solicitada continuamente pelas fontes (conexões), os algoritmos podem realizar essa alocação amparados por garantias estatísticas de QoS.

Obviamente, esses algoritmos são bem mais complexos que os algoritmos de alocação não-estatística, e exigem um processamento muito maior da unidade de controle de admissão. Todavia, resultam em grande economia de recursos em se tratando de tráfegos de taxa variável.

Podemos identificar dois grandes obstáculos a serem transpostos na implantação de algoritmos de alocação estatística:

1. a dificuldade de se obter uma caracterização fiel dos fluxos de chegada para servir de base aos cálculos estatísticos do algoritmo;

2. as decisões de admissão devem ser tomadas em tempo real, o mais rápido possível. Para que essa meta seja atingida, o processamento solicitado à unidade de controle de admissão deve ser reduzido ao máximo.

O maior desafio em desenvolver um algoritmo de Controle de Admissão eficiente e efetivo é a determinação da banda requerida por uma nova conexão de maneira suficientemente rápida. Para conexões com taxas constantes esse problema é simples. Porém, para conexões com taxas variáveis e com requisitos

PUC-Rio - Certificação Digital Nº 0210416/CB

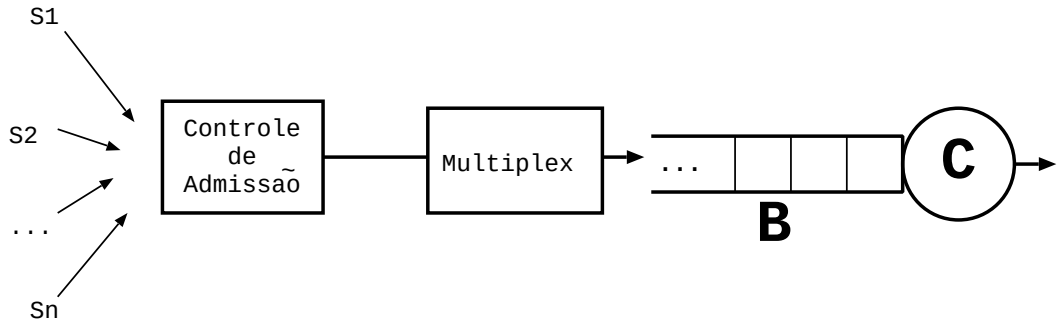

Figure 2.1: Modelo geral para análise dos algoritmos de Controle de Admissão

estritos de tempo a solução não é trivial. Segundo [1], um esquema de Controle de Admissão deve reunir várias características, dentre as quais destacam-se:

- *Simplicidade*: os algoritmos devem ser simples em termos de economia na implementação;
- *Flexibilidade*: a arquitetura do algoritmo deve ter suficiente adaptabilidade para incluir, de maneira simples, as alterações decorrentes de novos serviços que poderão aparecer no futuro;
- *Rapidez*: o algoritmo deve ser rápido o suficiente para processar as decisões de admissão em tempo real;
- *Eficiência*: o algoritmo deve alcançar alta utilização de recursos por meio da exploração da multiplexação estatística;
- *Efetividade*: deve ser capaz de garantir a QoS comprometida;
- *Controlabilidade*: o controle de tráfego deve ser alcançado sem degradar o desempenho da rede.

Há diversos algoritmos que abordam esses problemas de várias maneiras. De modo geral, podem ser divididos em duas categorias: baseados em modelos e baseados em medidas.

A seguir, são apresentados vários esquemas de Controle de Admissão. A avaliação de desempenho desses algoritmos é realizada nas referências corespondentes, sendo que em todos os casos, o esquema é proposto e avaliado com base no modelo da Figura 2.1. Este modelo corresponde à arquitetura de um nó de entrada da rede ao qual um grupo de fontes S_n podem ter acesso. Os pacotes das fontes que recebem permissão do mecanismo de Controle de Admissão são multiplexados através de agendamento com prioridade FIFO e entram em buffer de tamanho B, aguardando transmissão.

Os parâmetros de desempenho considerados são basicamente a perda no buffer e a vazão ou taxa de utilização.

PUC-Rio - Certificação Digital Nº 0210416/CB

2.2.1 Algoritmos baseados em modelos

Para que a decisão de admissão de novos usuários seja tomada, os algoritmos utilizam modelos de tráfego previamente definidos para caracterizar cada usuário. Com esses modelos, o algoritmo pode prever se a inclusão de um novo usuário vai resultar em degradação de QoS dos demais fluxos ou não.

Os algoritmos de alocação estatística baseados em modelos têm, portanto, a sua eficiência diretamente relacionada à precisão desses modelos de tráfego. Devido à dificuldade de se caracterizar o tráfego característico de determinadas aplicações, o desempenho dos algoritmos baseados em modelos pode ser bastante comprometido.

O método que pode ser considerado básico é o da Capacidade Equivalente ([14]). Na realidade, o conceito e o método da Capacidade Equivalente podem ser vistos como um dos modelos de uso mais geral, e refletem o problema básico do Controle de Admissão.

A capacidade Equivalente de um fluxo ou de um agregado é definida como a porção de banda requerida pelo fluxo (ou agregado) para obter o nível de QoS desejado. Em geral, este nível de QoS é um valor de probabilidade de perda.

De acordo com este modelo, testes de Controle de Admissão podem ser propostos para determinar se a banda disponível no enlace é maior do que a capacidade equivalente requerida pelo fluxo. O problema, portanto, está na determinação da relação entre o modelo de tráfego e a capacidade equivalente. Para isto, inúmeros métodos analíticos têm sido propostos para diversos tipos de tráfego. Podemos citar:

- Método da Aproximação de Tráfego Pesado ([41] e [42])
- Método da Reserva Rápida de Banda ([49] e [51])
- Método de Admissão por Curvas de Serviço ([39])

Estes modelos são analisados a seguir.

PUC-Rio - Certificação Digital Nº 0210416/CB

Método da Aproximação de Tráfego Pesado

Este mecanismo busca, através da caracterização dos períodos *on* e *off* e das taxas de pico das fontes, obter um resultado simples baseado no comportamento assintótico da cauda da distribuição do tamanho de fila quando a intensidade de tráfego é alta. Um pedido de conexão só é aceito caso a taxa de perda, aproximada pela cauda da distribuição do tamanho da fila, seja menor que o valor desejado. A aproximação é considerada boa para situações de alta intensidade de tráfego (acima de 80%).

Esta é uma boa aproximação quando há um grande número de fontes com taxas de pico muito pequenas comparadas com a capacidade do link. Somente neste caso, a proposta pode ser usada na região de tráfego pesado. No entanto, os cálculos envolvidos na aproximação são bastante complexos, resultando num Controle de Admissão pouco escalável.

Método da Reserva Rápida de Banda

Este esquema, voltado para redes ATM, foi desenvolvido objetivando amenizar o problema de alocação de recursos para fontes caracterizadas por alta explosividade. A sua idéia principal é que a rede tente alocar os recursos necessários para o período de duração da rajada somente quando a fonte estiver pronta para transmitir. Existem esquemas para reserva rápida tanto de banda quanto de *buffer*.

No método da reserva rápida de banda, a alocação de banda requerida pela fonte é feita através de pedidos de incremento ou decremento, podendo seu valor variar entre zero e a sua taxa de pico. Os pedidos de incremento são feitos por meio de uma célula de sinalização. O aumento requerido só será aceito por um nó da rede caso a soma do tráfego total requerido não exceda a capacidade do canal, sendo, portanto, baseado na alocação pelo pico. Se o incremento é negado por qualquer nó no caminho do circuito virtual, o incremento é bloqueado. Os pedidos de decremento são feitos através de células de gerenciamento e são sempre aceitos.

PUC-Rio - Certificação Digital Nº 0210416/CB

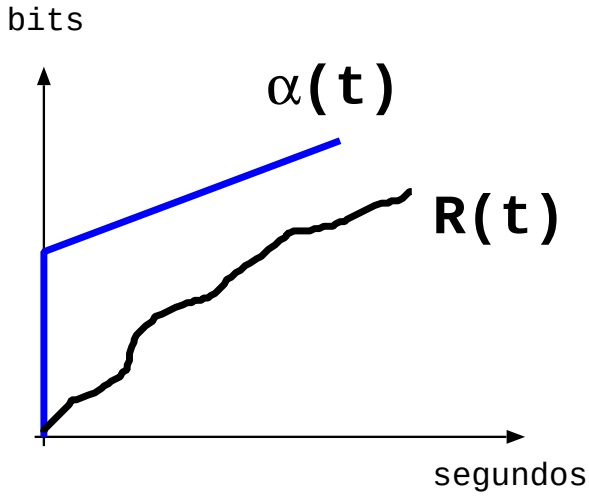

Figure 2.2: Curva de chegada limitando o fluxo de bits

Método de Admissão por Curvas de Serviço

PUC-Rio - Certificação Digital Nº 0210416/CB

Um esquema de Controle de Admissão deve operar em vários nós da rede, e assim, o volume de tráfego processado pode ser muito grande se os algoritmos não possuírem certa simplicidade. Por outro lado, um melhor aproveitamento dos recursos da rede vai implicar em algoritmos mais complexos. Portanto, um requisito-chave aos esquemas de controle de admissão é aliar simplicidade a capacidade de processamento. Dois pontos que constantemente são contraditórios. Como se sabe, o conceito de curvas de serviço [27] ; [37] permite representar situações complexas de modelagem de tráfego em problemas matematicamente simples e graficamente visualizáveis.

Curvas de serviço e curvas de chegada [27] são conceitos que podem ser bastante úteis ao dimensionamento de recursos de rede visando prover um determinado nível de serviço aos usuários. São funções utilizadas como limitantes de fluxos de dados que permitem, de maneira simples, determinar as garantias de banda e de retardo que um nó pode prover. Constituem, portanto, um processo de modelagem adequado às necessidades de um esquema de controle de admissão.

A curva de chegada de um fluxo é por definição uma curva limitante superior ao volume de tráfego associado a este fluxo. Este conceito está ilustrado na Figura 2.2, onde $R(t)$ é a função acumulativa que descreve o fluxo de bits na porta de entrada de um determinado nó da rede e $\alpha(t)$, a sua curva de chegada.

As curvas de chegada são funções não-decrescentes que podem ser uti-

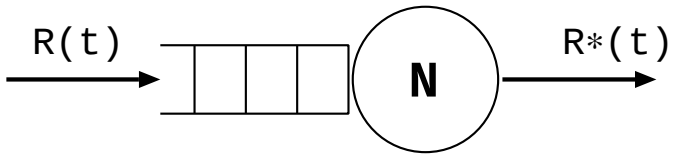

Figure 2.3: Fluxos de chegada e de saída do nó N

lizadas como modelo das fontes de tráfego pelo mecanismo de Controle de Admissão. Um exemplo de curva de chegada é o policiador *token bucket*, que, através dos parâmetros r e b (respectivamente, taxa de transmissão e tamanho do balde), limita o tráfego policiado, descartando ou modelando os pacotes não-conformes. Na Figura 2.2, a curva de chegada $\alpha(t)$ é um *token bucket* que limita o fluxo $R(t)$. Um mecanismo de Controle de Admissão pode, portanto, considerar $R(t)$ plenamente caracterizado pelos parâmetros r e b do *token bucket*.

Analogamente, temos o conceito de curva de serviço que pode ser associado a um nó N da rede que atende a um fluxo de entrada dado pela função $R(t)$ e caracterizado por um fluxo de saída dado por $R^*(t)$, como ilustrado em 2.3.

Dizemos que o nó N oferece ao fluxo uma curva de serviço $\beta(t)$ se e somente se:

$$R^*(t) \geq R(t) \otimes \beta(t)$$

Que é equivalente a $R^*(t) \geq inf_{0 \leq s \leq t}[R(s)+\beta(t-s)]$. A Figura 2.4 ilustra essa relação. Note que em qualquer instante t, a saída $R^*(t)$ estará sempre acima de $R(t) \otimes \beta(t)$, ou seja, acima do menor valor que pode ser obtido pela adição da curva de serviço ao fluxo de chegada em qualquer instante anterior. Portanto, a curva de serviço permite determinar um limitante inferior para o tráfego encaminhado em um determinado nó.

Por exemplo, se, no sistema da Figura 2.3, o nó N transmite a uma taxa C, teremos como curva de serviço desse nó $\beta(t) = C.t$.

A aplicação desses conceitos ao Controle de Admissão é direta: cada nó da rede possui uma curva de serviço $\beta(t)$ e cada fluxo que solicita serviço possui uma curva de chegada $\alpha_i(t)$. Para tomada da decisão de admissão, todos os fluxos (inclusive os solicitantes) são representados por uma curva de chegada única, $\alpha(t)$, e assim o algoritmo pode determinar quais as garantias que o nó pode oferecer e compará-las com os requisitos. No caso da Figura 2.5, por exemplo, d representa o atraso máximo a que os fluxos estão sujeitos, e $\alpha(t)$ é a soma das curvas de chegada dos fluxos já admitidos e do fluxo solicitante. Se o requisito de atraso for menor que d, o fluxo solicitante não é aceito.

PUC-Rio - Certificação Digital Nº 0210416/CB

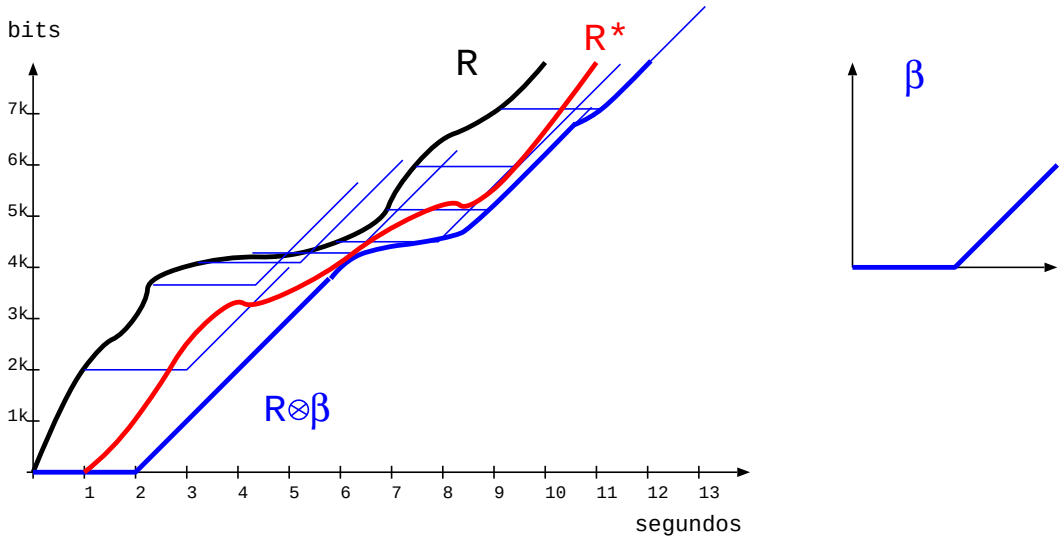

Figure 2.4: Curva de serviço do nó N

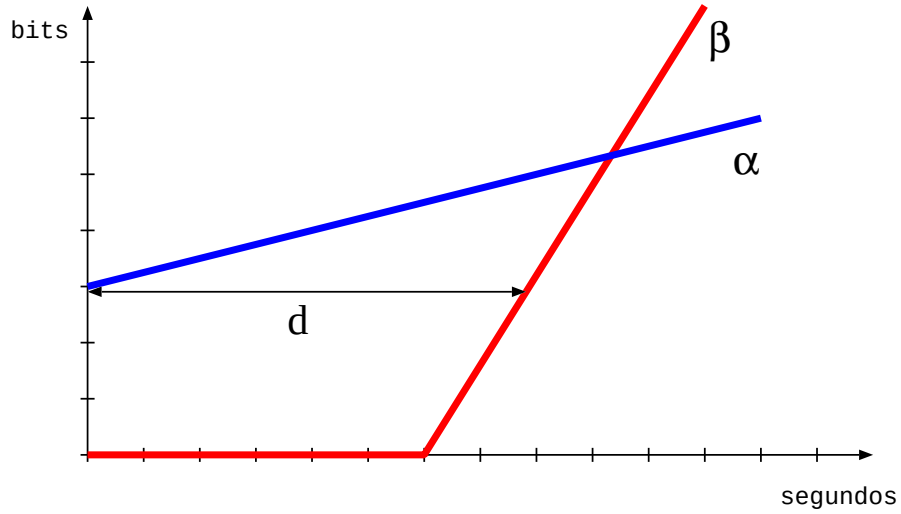

Figure 2.5: Decisão de admissão por curvas de serviço

PUC-Rio - Certificação Digital Nº 0210416/CB

2.2.2 Algoritmos baseados em medidas

Considerando-se a dificuldade em se caracterizar o tráfego de determinadas aplicações, os algoritmos baseados em medidas são utilizados como uma opção computacionalmente mais simples e, muitas vezes, de resultados superiores aos algoritmos baseados em modelos. Nesses, o módulo de controle de admissão realiza as suas decisões baseado em cálculos frequentemente bastante complexos, o que implica em consumo de tempo e capacidade de processamento, e que nem sempre apresentam a precisão desejada. Os algoritmos baseados em medidas contornam essa dificuldade utilizando medições das

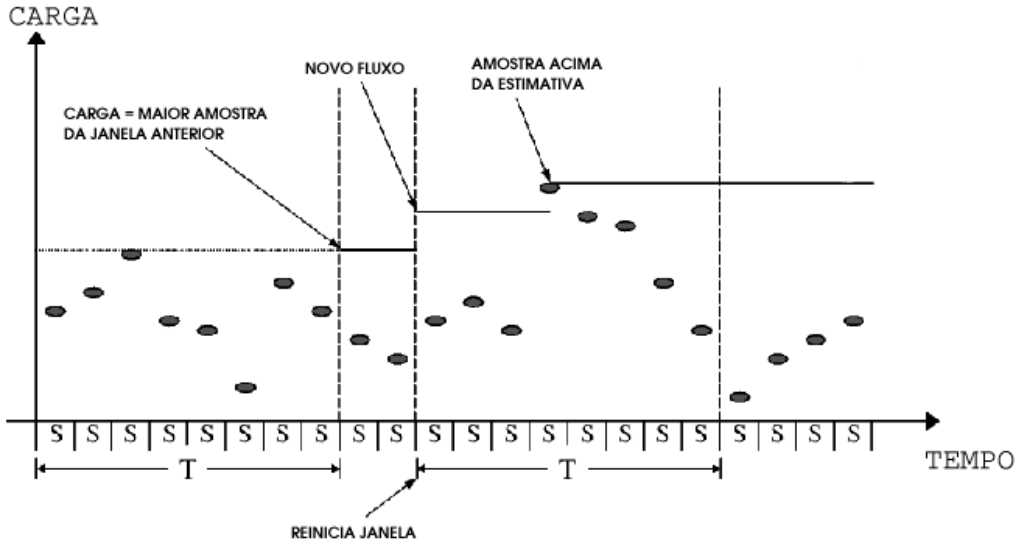

Figure 2.6: Medição de carga na rede baseada em Janelas de Tempo

PUC-Rio - Certificação Digital Nº 0210416/CB

condições de tráfego da rede para a tomada de decisões. Dessa maneira, as decisões consomem menos recursos do sistema e são mais rápidas.

A principal diferença entre os algoritmos dessa categoria é o método de medição. Há várias técnicas que realizam essa tarefa, que são utilizadas para determinação do estado da rede. A seguir são comentados três dos principais mecanismos de medição: as técnicas de janelas de tempo (*time window*), amostras de pontos (*point samples*) e média exponencial (*exponential averaging*).

Janelas de Tempo: Esse método mede a carga de tráfego da rede por meio de uma amostragem tomada em intervalos regulares (S). A cada período T, correspondente a uma certa quantidade de amostras S, a carga da rede é estimada como a maior amostra em T. Quando um novo fluxo é admitido no sistema, a estimativa é incrementada e a janela é reiniciada. O valor da estimativa é também atualizado imediatamente se uma amostra medida é maior que a estimativa atual. A Figura 2.6 a seguir apresenta graficamente o esquema descrito.

Amostras de Pontos: Este método simplesmente toma uma amostra da carga instantânea a cada intervalo S e trata esta medida como a carga média.

Média Exponencial: Similar aos mecanismos anteriores, a técnica da média exponencial toma uma amostra da carga do tráfego a cada intervalo S. Entretanto, a carga média $v^{'}$ é atualizada como uma função da medição passada v e a medida de carga instantânea v_i, ou seja:

$$v' = (1 - \omega) * v + (\omega) * v_i$$

onde ω é um peso de avaliação que determina quão rápido a média estimada adapta-se às novas medidas ($0 < \omega < 1$). Um coeficiente ω grande (próximo a 1) resulta em uma reação mais rápida à dinâmica da rede.

Uma vez determinado o estado da rede em função do parâmetro de desempenho avaliado (seja banda disponível, taxa de perda de pacotes, atraso médio no domínio, ...), o algoritmo avalia se a inclusão do fluxo solicitante é possível. Essa avaliação pode ser feita de várias maneiras, entre as quais:

- Através de um modelo de tráfego do fluxo solicitante. Dessa maneira, a imprecisão da modelagem de tráfego não é um ponto tão crítico, pois esta estará presente apenas na caracterização de um fluxo, enquanto todos os outros serão representados pelos dados obtidos por medições.
- Através da emissão dos chamados pacotes de sondagem. São pacotes de teste da rede, enviados e monitorados pelo módulo de controle de admissão, que então compara a Qualidade de Serviço que o tráfego de sondagem recebeu (perda, atraso, etc) com a QoS solicitada pelo usuário.

Os seguintes métodos de Controle de Admissão baseados em medidas serão brevemente abordados neste Capítulo:

- Método da soma medida ([45])
- Método do limitante da taxa de perda de células ([3],[13],[40])
- Método da banda equivalente utilizando limitante de Hoeffding ([15])

PUC-Rio - Certificação Digital Nº 0210416/CB

Método da soma medida

Consiste no mesmo algoritmo que a alocação não-estatística, com a diferença que a banda ocupada pelas conexões existentes é determinada por medições da carga da rede. Devido à natureza explosiva da maioria dos tipos de tráfego, na maior parte do tempo cada usuário utilizará apenas a taxa média, que é muito inferior à taxa de pico, e o algoritmo permitirá a admissão de um número muito maior de usuários.

Em comparação com a alocação não-estatística, as decisões de admissão são muito menos conservadoras, amenizando assim os efeitos de *over-provisioning*. Como conseqüência, a soma das taxas de pico de todos os usuários

alocados poderá ser maior que a capacidade C. Para minimizar esse problema, o algoritmo realiza as decisões baseado no limiar $\alpha.C$, com $0 < \alpha < 1$, reservando um excedente de banda para os instantes de maior atividade da rede.

Método do limitante da taxa de perda de células

Há vários algoritmos de Controle de Admissão desenvolvidos para as redes ATM, que têm a taxa de perda de células (CLR) como parâmetro de decisão. A diferença entre os algoritmos está na abordagem matemática. O algoritmo proposto em [3] utiliza os números médio (*Average Number of Arrivals - ANA*) e máximo (*Maximum Number of Arrivals - MNA*) de chegada de células (provenientes das conexões existentes) em um determinado intervalo de tempo ΔT para cálculo do limite superior da CLR em cada nó da rota especificada pelo solicitante.

Os parâmetros ANA e MNA são obtidos por medição, e utilizados no cálculo da probabilidade P_i de chegada de células originárias de cada uma das i conexões estabelecidas.

Pode-se, então, provar que o limitante superior do CLR é função de P_i e do intervalo ΔT. Através dessa relação, cada nó da rota toma a decisão de admissão individualmente.

Método da banda equivalente utilizando limitante de Hoeffding

Esse algoritmo busca determinar a capacidade equivalente das fontes de tráfego através de limitantes de Hoeffding. Em [15], o autor propõe um método de estimação da banda equivalente de uma fonte ou de um agregado de fontes homogêneas utilizando limitantes de Hoeffding, tendo como base dos cálculos os parâmetros de Token Bucket e as medidas de tráfego.

A partir dos parâmetros de Token Bucket, estabelece-se um limitante superior [15] para a taxa de pico de cada classe de tráfego. O teorema de Hoeffding ([18],[19]) define um limitante superior para a banda equivalente, dado por:

$$\hat{C}_H(\mu_S, \{p_i\}_{1\leq i\leq n}, \epsilon) = \mu_S + \sqrt{\frac{ln(1/\epsilon)\sum_{i=1}^{n}(p_i)^2}{2}}$$

PUC-Rio - Certificação Digital Nº 0210416/CB

Sendo μ_S a taxa média medida, e ϵ a probabilidade da taxa exceder a capacidade do link.

PUC-Rio - Certificação Digital Nº 0210416/CB