Novas Políticas de Escalonamento para Controle de Admissão em Redes ATM

Daniel Antunes Maciel Villela   Otto Carlos Muniz B. Duarte*

daniel@gta.ufrj.br   otto@gta.ufrj.br

Grupo de Teleinformática e Automação – GTA
COPPE/EE - Programa de Eng. Elétrica/Departamento de Eletrônica
Universidade Federal do Rio de Janeiro
C. P. 68504 - CEP 21945-970 - Rio de Janeiro - RJ - Brasil
http://www.gta.ufrj.br

Resumo

A integração de serviços em redes ATM oferece um compartilhamento dos canais de comunicação e garantia de qualidade de serviço (QoS - *Quality of Service*) para as conexões. Deve-se visar multiplexação estatística entre as várias conexões em um mesmo canal sem contudo afetar o nível de QoS que lhes é proporcionado e, ao mesmo tempo, realizar um controle de congestionamento preventivo. Portanto, a resolução do problema de alocação de recursos e a aceitação da chamada adquire importância para que se atinja estes objetivos. Este trabalho apresenta alguns modelos de caracterização de tráfego, essencial para o tratamento adequado do problema, alguns mecanismos de Controle de Admissão de Chamada (CAC - *Call Admission Control*) e seu tratamento sob novas políticas de escalonamento.

Abstract

ATM service integration provides communication channel sharing and quality of service (QoS) guarantees. The connections on the same channel must share resources with statistical allocation without lowering QoS levels. That is also a preventive congestion control. Thus, the resource sharing and call acceptance are really important to reach these objectives. This paper shows traffic characterization models, suitable for the correct problem approach, as well as Call Admission Control (CAC) mechanisms and his treatment under new sheduling policies.

*Este trabalho foi realizado com recursos da UFRJ, FUJB, PROTEM-CC, CNPq, CAPES e COFECUB.

1 Introdução

As Redes Digitais de Serviços Integrados permitem a integração de variados serviços com um amplo espectro de características. Possibilita desde serviços típicos de uma rede local como transferência de arquivos até aplicações multimídias, teleducação, HDTV (*High Definition Television*) etc. Para sua implementação, adotou-se o Modo de Transferência Assíncrono (ATM - *Asynchronous Transfer Mode*). As redes ATM oferecem um suporte aos mais diversos serviços, incluindo-se aplicações multimídias, com garantia de QoS.

Por meio de propriedades como multiplexação estatística, uma alta eficiência pode ser alcançada nas conexões existentes em ATM. Para isto, é importante a especificação de parâmetros de QoS pretendidos na chamada. Faz-se uma avaliação em relação aos recursos necessários para atender a QoS requerida pela nova conexão bem como para manter a QoS das conexões já existentes no canal compartilhado. Caso seja possível atender os requisitos da nova chamada, sem comprometer as conexões já existentes, a conexão será estabelecida.

Esta será, em resumo, a ação desempenhada pelo Controle de Admissão de Chamada (CAC)[1]. Como realiza uma tarefa que irá limitar o tráfego na rede, pode-se afirmar que esta ação faz parte de um controle de congestionamento preventivo [2]. Logo, para este objetivo, será necessário em primeiro lugar a especificação da QoS desejada. Isto será possível através da caracterização do tráfego gerado pelas fontes que compõem as aplicações. Em segundo lugar, será importante a alocação de recursos na rede. Relacionado a isto, uma estimativa acurada e confiável é essencial para atingir-se uma alta eficiência na multiplexação estatística entre as conexões. E, por último, caso esta alocação de recursos seja possível, será realizada a aceitação da chamada.

Portanto, os problemas relacionados a CAC consistem no desenvolvimento de mecanismos que o implementem de maneira confiável, rápida e eficiente [3, 4, 5]. Paralelamente, deve-se buscar uma modelagem precisa do tráfego para traduzir os parâmetros necessários ao CAC de forma mais realista [6, 7]. Ao mesmo tempo, a disciplina de serviço a ser considerada é também importante para uma estimativa acurada do número de conexões a serem comportadas em um canal de comunicação. É interessante a utilização de políticas que permitam maior distinção entre as diferentes classes de serviço e seus requisitos de QoS e flexibilidade em sua implementação. Um exemplo seria oferecer maior prioridade a uma conexão com tráfego com maiores restrições ao atraso máximo tolerado. Por isso, algumas políticas de escalonamento têm sido objeto de estudo para se obtenha este grau de funcionalidade [8, 9, 10].

Este trabalho realiza uma abordagem ampla sobre todos estes aspectos citados. Inicialmente, discute-se os conceitos e as características da multiplexação estatística na seção 2. Em seguida, na seção 3, aborda-se o problema de caracterização de tráfego, descrevendo-se alguns modelos propostos. A seção 4 apresenta alguns mecanismos de CAC, destacando-se suas principais vantagens e limitações. Algumas políticas de escalonamento são apresentadas na seção 5. Por fim, são apresentadas as considerações finais na seção 6.

2 Multiplexação Estatística

Um aspecto interessante a ser visado em uma rede de serviços integrados é a multiplexação estatística. Considerando-se que, em um canal de comunicação, haverá vários fluxos correspondentes às conexões, deve-se ter uma divisão dos recursos para as conexões que os geram.

Nas redes em que o acesso ao meio é compartilhado, esta alocação de recursos é dinâmica, no sentido de que todos possuem a taxa máxima permitida no canal, se houver disponibilidade total. Caso contrário, irão compartilhar com outras conexões o que lhes é disponível.

Nas redes ATM, é prevista uma alocação de recursos ao se realizar a chamada. Por isso, deve-se respeitar as taxas contratadas de modo a garantir a QoS da conexão bem como das outras conexões que compartilham o mesmo canal com recursos reservados. Para isso, é importante a caracterização do tráfego gerado pelas fontes. Será importante da mesma forma determinar o nível de QoS exigido pela aplicação que faz a chamada (voz, vídeo etc.).

Uma forma simples de se realizar a reserva de recursos seria a alocação da taxa de pico, que consiste em alocar os recursos pelo máximo que é exigido pelo tráfego gerado. Contudo, esta intensidade de tráfego ocorre apenas por um período de tempo, caracterizando-se o que se chama rajada. Mesmo que este período de tempo seja longo, haverá um espaço de tempo em que o canal não estará sendo utilizado por esta conexão, de modo que estes recursos podem ser destinados dinamicamente a outra conexão. Logo, pode-se realizar uma multiplexação das várias conexões, trabalhando-se com a distribuição estatística dos tráfegos que possuem. Este efeito pode ser visualizado na figura 1, em que observa-se que somando-se os recursos utilizados pelas várias fontes, o total exigido é bem inferior ao total que seria alocado se fosse efetuada alocação pelo pico.

O problema da alocação de recursos e aceitação da chamada é tipicamente um problema de teoria de filas. As várias fontes injetam fluxos em uma fila de capacidade finita que é servida por um servidor com tempo de serviço constante, igual ao tempo necessário

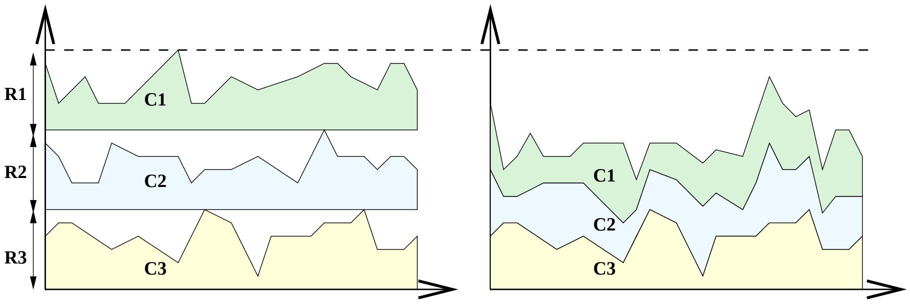

Figura 1: Multiplexação Estatística.

para transmitir uma célula.

3 Caracterização de Tráfego

Para que se projete com eficiência um sistema de comunicações, faz-se importante uma caracterização do tráfego que impõe-se ao canal utilizado. Por isso, estudam-se modelos que o descrevam o mais próximo possível da realidade.

Com intuito de especificar classes de serviço segundo a taxa binária do tráfego e nível de QoS proporcionado, foram definidas as seguintes classes: Taxa Binária Constante (*Constant Bit Rate* - CBR), Taxa Binária Váriavel (*Variable Bit Rate* - VBR), Taxa Binária Disponível (*Available Bit Rate* - ABR) e Taxa Binária Indefinida (*Unspecified Bit Rate* - UBR).

O serviço CBR foi definido para aplicações que geram informação a taxa constante. Equivaleria à situação em que a fonte transmitiria ininterruptamente. São exemplos: emulação de circuito, vídeo e voz não comprimida. De uma maneira geral, em CBR, deve-se utilizar a taxa de pico, especificada pelo usuário, que é igual a taxa média como parâmetro de descrição.

O serviço ABR define para as aplicações a utilização da faixa disponível, com um valor mínimo reservado. Oferece uma realimentação com o objetivo de permitir controle de fluxo e de congestionamento na rede [11].

O serviço UBR não oferece garantia de qualidade de serviço nenhuma. Permite a transmissão caso possa atender o pedido (*Best Effort*), de modo que o usuário deve tolerar qualquer capacidade que a rede ofereça. Um exemplo possível é uma transferência de arquivos executada em *background*. Por tudo isto, não requer conhecimento de características do tráfego gerado pela fonte.

O serviço a que a propriedade de multiplexação estatística mostra-se mais adequada é o VBR. Neste serviço, as características do tráfego gerado pelas fontes variam muito. Por isso, torna-se mais complexa sua modelagem. Entretanto, a possibilidade de maior ganho de multiplexação estatística, com garantia de QoS para as conexões, o tornam mais atrativo para aplicações como vídeo e voz comprimidos.

Nas seções seguintes, serão detalhadas as características importantes para a caracterização de fontes de voz e de vídeo e os principais modelos que as descrevem.

3.1 Fontes de Voz

O comportamento típico de uma fonte de voz é a alternância entre intervalos de tempo com rajada e intervalos com silêncio (figura 2). Portanto, há dois períodos característicos que podem ser associados a estes intervalos: ativo e inativo. Para uma conversa normal, a variável aleatória que descreve a duração dos períodos ativos e inativos podem ser aproximados por uma distribuição exponencial. Para o período inativo, esta aproximação não é perfeita, pois dever-se-ia fazer uma distinção entre os períodos inativos relativos à pausa e os relativos à escuta. No entanto, com um número grande de fontes multiplexadas em um canal de transmissão, a aproximação por distribuição exponencial é muito boa [12].

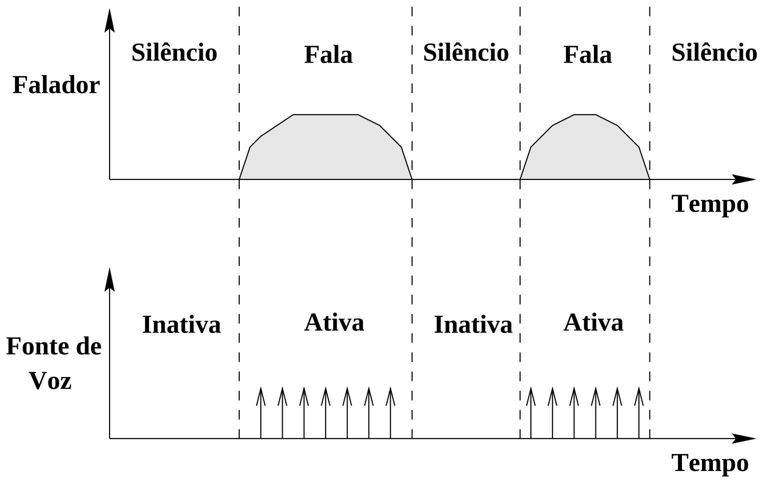

Figura 2: Comportamento típico de uma fonte de voz.

Pode-se dizer que a aproximação por distribuição exponencial (sem memória) equivale a dizer que o processo de chegadas é um processo de renovação. Desta forma, não existe correlação entre as chegadas sucessivas. Entretanto, como este tráfego possui alto índice de rajadas, o agregado de várias fontes não será um processo de renovação.

3.2 Fontes de Vídeo

As fontes de vídeo assumem importância destacada nas aplicações multimídias atuais e nas que estão por vir. Necessitam, em geral, de larga faixa para transmissão. Por isto, um correto dimensionamento é essencial para a modelagem dos parâmetros de transmissão.

Deve-se, inicialmente, comentar algumas características intrínsecas ao vídeo que estão presentes nos algoritmos de compressão que tornam sua taxa variável. Em primeiro lugar, a correlação entre quadros é alta, normalmente. Uma opção vantajosa é o envio das diferenças entre quadros, pois economiza-se largura de faixa necessária e, ao mesmo tempo, gera tráfego variável. Os modelos desenvolvidos podem utilizar variações intra-quadro, introduzidas pela manipulação de blocos da imagem, procedimento comum em algoritmos de compressão.

Há, também, variações de longo intervalo. Como o próprio nome já diz, estão relacionadas às variações que ocorrem em pontos bem espaçados em uma escala de tempo. É o caso das mudanças de cenário, por exemplo. A mudança na taxa binária sofre uma alteração descrita como um degrau de variação.

Há, em contrapartida, as variações de curto intervalo. Estão presentes nos detalhes que aparecem em um vídeo, como mudanças de contexto da imagem em uma mesma cena. As variações de taxa binária são mais suaves, sendo que a correlação entre variações decai exponencialmente com o tempo.

Para o estudo de modelagem de fontes de vídeo, alguns modelos têm sido propostos. Alguns deles são bem similares aos que foram apresentados anteriormente no estudo de fontes de voz.

3.3 Modelo de Chegadas e Serviços Uniformes

O modelo de chegadas e serviços uniformes (*Uniform Arrivals and Services* - UAS) assume um fluxo contínuo para fontes homogêneas e é utilizado para fontes de voz. É chamado processo de fluido. São considerados dois estados possíveis para as fontes: ON e OFF, correspondentes às fases ativa e passiva, respectivamente. Supõe-se que estes estados possuem uma distribuição exponencial, com médias α^{-1} e β^{-1} (figura 3). Este modelo trata o processo de chegada e de remoção da fila uniformemente. Utiliza-se, então, uma cadeia de Markov de tempo contínuo para modelar o número de fontes ativas no sistema, com um sistema composto por N fontes de voz. O diagrama que representa esta cadeia de Markov pode ser observado na figura 4.

Para o caso de k fontes estarem ativas, a taxa de probabilidade com que o sistema passa para $k-1$ fontes ativas é igual a $k\beta$ e a taxa com que passa para o estado de $k+1$

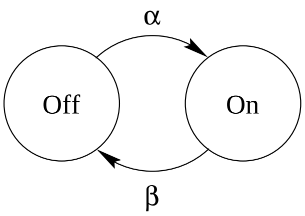

Figura 3: Dois estados para uma fonte homogênea.

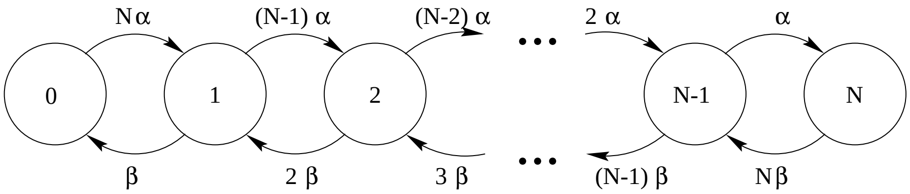

Figura 4: Diagrama de Estados para o modelo UAS.

fontes ativas, $(N-k)\alpha$. Já observando-se as taxas de probabilidade, em que o sistema vai para este estado, pode-se observar (figura 5): transição de $k-1$ para k ocorre com taxa $(N-(k-1))\alpha$ e a transição de $k+1$ para k com taxa $(k+1)\alpha$. Para obter-se uma solução analítica para este problema, deve-se escrever as equações diferenciais que regem este sistema:

$$\begin{aligned}(k-C)\frac{dF_k(b)}{db} &= (N-k+1)\alpha F_{k-1}(b) - \{(N-1)\alpha + \beta k\} F_k(b) + \beta(k+1)F_{k+1}(b) \ , \\ & \quad 0 \leq k \leq N. \end{aligned} \tag{1}$$

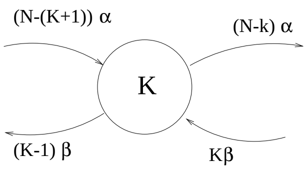

Figura 5: Fluxo de probabilidade para um estado genérico k.

3.4 Cadeias de Markov de Tempo Contínuo

O modelo de Cadeias de Markov de Tempo Contínuo (CMTC) é semelhante ao descrito anteriormente. Possui caráter geral e permite modelagem de fontes de voz e de vídeo.

Também trabalha com cadeia de Markov e é considerado um processo de renovação. Para modelagem de fontes de foz, os estados da fonte são ativa e inativa. Assume-se distribuição exponencial com médias α^{-1} e β^{-1}. No entanto, em vez de supor uma geração de pacotes uniforme com taxa constante, considera-se um processo de Poisson, a uma taxa λ. Portanto, a superposicão de várias fontes acarretará em uma distribuição de Poisson com taxa dada pela soma das taxas individuais. Para o processo de serviço, assume-se uma distribuição exponencial, com parâmetro υ.

O diagrama de estados que representa este modelo encontra-se na figura 6. Cada estado será representado por um par $(s, \wp)$, onde s é o número de pacotes na fila e $\wp$, o número de fontes ativas no sistema. As equações de equilíbrio são obtidas do diagrama.

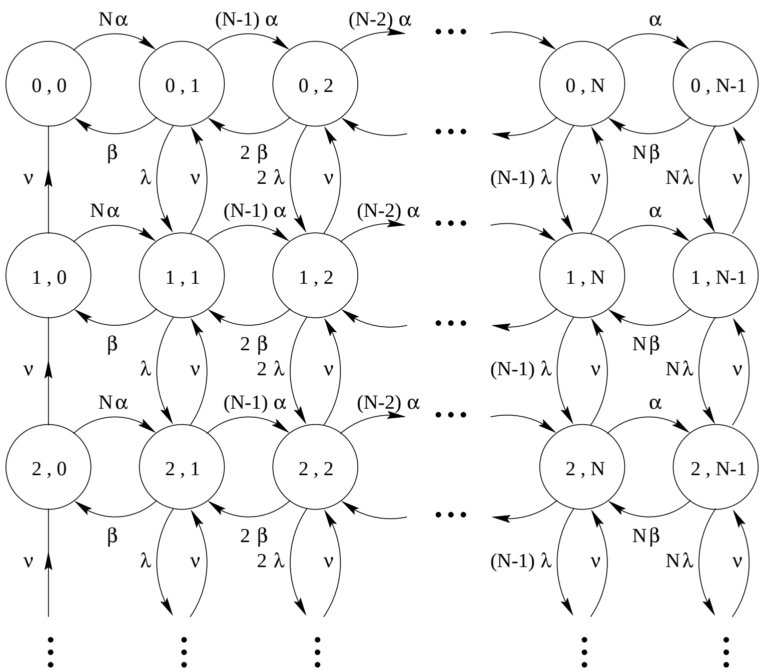

Figura 6: Diagrama de estados para o modelo CMTC.

Então, a probabilidade de do tamanho da fila ser igual a i, será dado por $\pi_i = \Pr(s = i+1) = \sum_{j=0}^{N} p_{i+1,j}$ e a probabilidade de que o tamanho do *buffer* exceda o valor b será:

$$G(b) \triangleq \Pr(\text{tamanho} > b) = 1 - \sum_{i=0}^{b} \pi_i, \qquad b \geq 0\ . \tag{2}$$

Há, para fontes de vídeo, semelhante ao que se usa para fontes de voz, uma modelagem por cadeias de Markov de tempo contínuo (CMTC), em que o espaço de estados discretos representa a taxa binária em seu níveis quantizados. Portanto, a quantidade de níveis utilizados estará intimamente ligada à qualidade e complexidade obtida neste modelo.

Com a suposição de apenas mudanças suaves correspondentes às variações de curto intervalo ou mudanças de contexto, é razoável a inexistência de saltos na cadeia de Markov entre estados que não sejam os mais próximos. Em outras palavras, permite-se apenas transições entre estados vizinhos, ou seja, o processo é dito processo de nascimento e morte (o diagrama de estados apresentado na figura 7 representa este modelo). As taxas de probabilidade de transição decrescem com os valores para as taxas binárias dos estados, ou seja, para estados cuja taxa binária é um valor mais elevado são menores as taxas de transição. Por outro lado, para estados com taxa binária mais baixa o inverso acontece pois as taxas de transição são maiores.

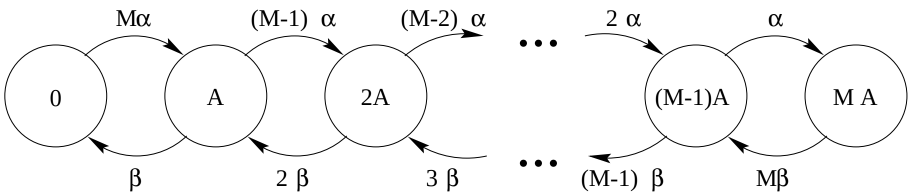

Figura 7: Diagrama de estados para o modelo CMTC aplicado a fontes de vídeo.

Com base nisso, pode-se alcançar alguns valores necessários para esta caracterização como a variância e função de autocovariância. Sejam M o número de níveis de quantização, a média $E\left(\overline{\lambda_n}\right)$, a variância $\overline{C_n}(0)$ e a função de autocovariância $\overline{C_n}(\tau)$ e do processo, tem-se:

$$
\begin{aligned}
&E(\overline{\lambda_n}) = MA\tfrac{\alpha}{\alpha+\beta}\ , \qquad\qquad \overline{C_n}(0) = MA^2\tfrac{\alpha}{\alpha+\beta}\left(1-\tfrac{\alpha}{\alpha+\beta}\right)\ , \\
&\overline{C_n}(\tau) = \overline{C_n}(0)e^{-(\alpha+\beta)\tau}\ .
\end{aligned}
\tag{3}
$$

Outro modelo, também baseado em cadeias de Markov, procura considerar adicionalmente as chamadas variações de longo intervalo. Com isto, obtém-se resultados mais precisos do que os anteriores. A representação é feita utilizando-se duas taxas binárias referentes à taxa alta e baixa, denotadas por A_h e A_l, respectivamente. O diagrama apresentado na figura 8 ilustra este modelo.

3.5 Processos de Poisson modulado por Markov

Os Processos de Poisson modulados por Markov (MMPP) são modelos que possuem como característica principal uma aproximação para o tráfego agregado por um processo de fluxo mais simples, correlacionado e de não-renovação. O modelo MMPP de m estados

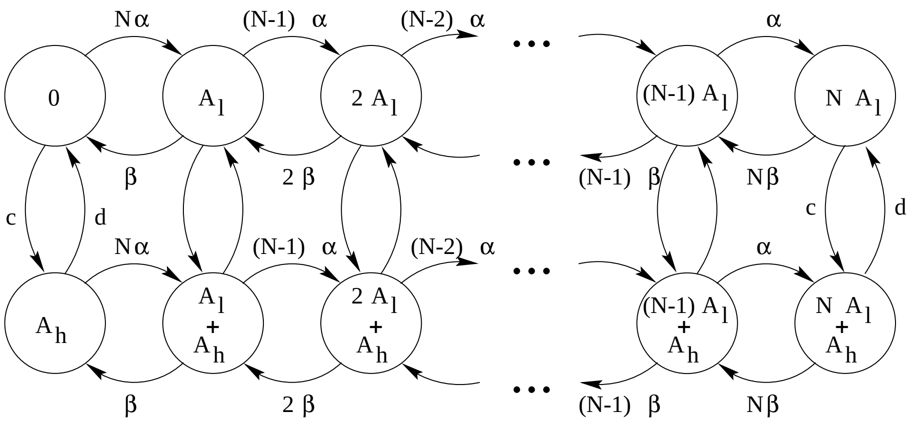

Figura 8: Diagrama de estados com possibilidade de variação de longo intervalo.

constitui um processo de Poisson duplamente estocástico, pois a taxa de chegada do processo de Poisson varia com o estado em que o sistema se encontra. Mais detalhadamente, neste ponto, encontra-se o conceito de modulação a que se refere o nome do modelo. A taxa de chegada será um processo de Poisson, mas não de taxa constante. Esta será determinada de acordo com o estado de uma cadeia de Markov de m estados. Por isso, diz-se que o processo é Poisson modulado por Markov. Portanto, a cada estado k do sistema pode-se associar uma taxa de chegada de Poisson λ_k. Para a cadeia de Markov de tempo contínuo de m estados, pode-se obter o gerador infinitesimal Q, a partir das taxas de probabilidade de transição. Assim como para as taxas de chegada λ_k, define-se a matriz: $\Lambda = \mathrm{diag}(\lambda_1, \lambda_2, \cdots, \lambda_m)$ e também o vetor $\lambda = (\lambda_1, \lambda_2, \cdots, \lambda_m)^T$. Pode-se destacar como propriedade do MMPP o fato de que uma superposição de MMPP's resulta em um novo MMPP. Matematicamente, pode-se descrever isto através dos geradores infinitesimais Q_i e pela matriz de taxas Λ_i:

$$\begin{aligned} Q &= Q_1 \oplus Q_2 \oplus \cdots \oplus Q_m \quad \text{e} \\ \Lambda &= \Lambda_1 \oplus \Lambda_2 \oplus \cdots \oplus \Lambda_m \ , \end{aligned}$$

onde $\oplus$ denota a soma de Kronecker.

Deve-se considerar que o número de estados m estará intimamente relacionado à complexidade do modelo. Um valor elevado para m acarretará em alta complexidade, o que em uma implementação computacional será altamente indesejável.

Este modelo possui variações de acordo com as premissas assumidas. Em um caso discreto, as taxas de chegadas poderiam ser do tipo Bernoulli, moduladas por uma cadeia de Markov e obtendo-se o modelo MMBP (*Markov Modulated Bernoulli Processes*). Outra possibilidade é considerar similarmente um modelo de fluido, acarretando pela mesma idéia em uma modelagem MMSF (*Markov Modulated Source Fluid*).

Particularmente, é interessante a utilização dos MMPP's para fontes de vídeo. A base para esta consideração fundamenta-se nas hipóteses de que uma cena única gera pacotes segundo um processo de Poisson, as taxas de transição entre cenas constituem uma cadeia de Markov e as altas taxas binárias seguintes a uma mudança de cena são causadas por um estado de taxa alta ou devido a chegadas em grupo. Neste último ponto, está a principal vantagem da utilização deste método, pois busca representar o aumento rápido das taxas binárias, decorrentes de uma mudança de cena.

4 Controle de Admissão de Chamada

O tráfego em redes ATM deverá possuir uma particularidade: o usuário poderá especificar a qualidade do serviço esperada em função das características do seu tráfego gerado e a rede de comunicação irá alocar os recursos necessários à garantia desta QoS. Um dos possíveis problemas que pode colocar esta garantia a perder é um eventual congestionamento na rede. Este controle é dificultado devido à alta velocidade do enlace, dos diversos requisitos de serviço e das diversas características de tráfego que as redes ATM se propõem a suportar [13, 14]. Há duas categorias básicas de classificação de controle de congestionamento: preventivo e reativo [15].

Supõe-se, quando ocorre um pedido de conexão na rede, que é possível alocar recursos para esta chamada, caso contrário o pedido é negado. Quem realiza esta tarefa de alocação de recursos e de aceitação (ou não) da chamada desempenha um controle de congestionamento preventivo. Trata-se do Controle de Admissão de Chamada (CAC - *Call Admission Control*) [16].

Alguns requisitos para este mecanismo são importantes:

- resposta em tempo real;
- eficiência na multiplexação estatística;
- margem de segurança;
- mecanismo de policiamento associado;
- validade para conexões de tipos variados.

Há, relativamente, maior facilidade para implementação de mecanismos de CAC nos serviços ABR, CBR e UBR. Para o CBR, como o parâmetro especificado é a taxa de pico, somente será aceita uma chamada caso seja possível alocar recursos da rede para esta taxa especificada. Para a classe ABR, há maior flexibilidade tanto nos parâmetros especificados como no gerenciamento do tráfego, por causa do mecanismo de realimentação que produzirá uma taxa binária com possibilidade de variação de acordo com a carga ou disponibilidade do canal [11]. Neste caso, a chamada será aceita se puder suportar a taxa binária mínima, sendo este o valor inicial assumido. Já para a classe UBR, não há garantia de QoS, por conseguinte perde-se um pouco do sentido do mecanismo. Não há alocação de recursos e qualquer QoS é tolerada.

Para o serviço VBR, há grande possibilidade de ganho na utilização da multiplexação estatística. Por isso, o estudo de mecanismos que permitam uma alta eficiência neste sentido e garantia de QoS para a chamada que será atendida e também para as outras conexões é vantajoso. Para resolução deste problema, é necessário o conhecimento da caracterização do tráfego, pois conhecendo-o será possível criar mecanismos que adotem os parâmetros que o descrevem para uma alocação de recursos ótima.

Portanto, vários mecanismos de CAC têm sido propostos. Alguns são baseados em um modelo de tráfego específico [17, 18], outros se utilizam de apenas alguns parâmetros de tráfego como taxa média, por exemplo. Com base nisso, elaborou-se uma classificação para estes mecanismos [19, 12]. Alguns mecanismos são apresentados segundo esta classificação, a saber: método de alocação de capacidades, aproximação não-paramétrica, método dinâmico.

4.1 Método de Alocação de Capacidades

Um dos mais importantes mecanismos propostos é o método de alocação de capacidades. Este método pretende avaliar a capacidade necessária a cada chamada para sua aceitação, a partir dos requisitos de QoS especificados. Este objetivo de QoS normalmente é expressado como um valor para a taxa de perda de células (CLR - *Cell Loss Rate*).

Neste contexto, encaixam-se os métodos analíticos que são baseados na caracterização e modelagem das fontes de tráfego como um modelo de problema de teoria de filas bem formulado e sua conseqüente resolução. Com este intuito, há alguns modelos do tipo MMPP/G/1 ou MMPP/D/1/K.

Uma opção a ser utilizada é o modelo UAS, considerando-se uma multiplexação de um número finito de fontes idênticas e independentes, ou seja, fontes homogêneas. A partir da CLR, do número de fontes e do tamanho do *buffer*, poder-se-ia calcular o fator de expansão e, com isso, facilmente obter a capacidade a ser alocada para a chamada. A vantagem deste modelo é exatamente a sua rapidez e simplicidade, porém sua desvantagem consiste no fato de que apenas considera fontes homogêneas.

Com o objetivo de especificar uma capacidade equivalente para as conexões, assim como o agregado das conexões, foi proposto por Guérin et al. [20] um método que se encontra nesta classe de mecanismos, denominado Capacidade Equivalente. Este método supõe duas situações distintas e apresenta para cada uma uma aproximação que melhor se adapta a alcançar os objetivos desejados, ao mesmo tempo em que constituem métodos de baixa complexidade. Para avaliar a capacidade equivalente, no caso em que o impacto das características de uma fonte individual é crítica, utiliza-se uma aproximação baseada em modelo de fluxo contínuo. Outra situação possível ocorre quando a multiplexação estatística das conexões é significativa. Neste caso, utiliza-se uma aproximação da distribuição estacionária da taxa binária.

4.1.1 Aproximação de Fluxo Contínuo

Esta aproximação utiliza um modelo de cadeia de Markov de 2 estados para cada fonte, descrito anteriormente na seção 3, correspondente a possuir fontes gerando tráfego em rajadas do tipo silêncio/atividade. Para caracterização de uma fonte, são necessários a taxa binária de pico (R), sua utilização (ρ) e o tempo médio de duração da rajada (L). Para a sua solução, deve-se considerar a distribuição do tamanho da fila, alimentada pela fonte com um servidor operando a uma taxa de serviço constante.

A partir disso, calcula-se a capacidade equivalente para a garantia da QoS desejada. O requisito de QoS especificado é a probabilidade de transbordo de *buffer* (ϵ). Com uma fila de tamanho finito (k), a relação entre a capacidade equivalente e esta probabilidade de transbordo de *buffer* pode ser expressa por

$$\varepsilon = \beta \ \exp\left(\frac{c - R\rho}{L(1-\rho)(R-c)c} \ k\right) , \tag{4}$$

onde

$$\beta = \frac{(c - \rho R) + \varepsilon\rho(R-c)}{(1-\rho)R} .$$

Para obter o valor de c, contudo não há solução analítica, devendo-se usar métodos numéricos. Por isso, Guérin et al. propuseram uma aproximação para o valor de β em 1. Analisando-se a expressão de β, observa-se que é uma aproximação válida quando o valor de c tende a R. Isto significa dizer que a capacidade equivalente seria muito próxima a taxa binária de pico. Conclui-se que o ganho de multiplexação estatística neste caso é baixo.

Para múltiplas fontes, o método propõe a utilização de uma aproximação assintótica para o cálculo da probabilidade de transbordo de *buffer*. Será apenas considerada a contribuição devida ao autovalor mais negativo associado ao sistema. A expressão para esta probabilidade é da forma:

$$G(x) \approx \beta \ e^{z_0 x} , \tag{5}$$

onde z_0 é o autovalor negativo de menor módulo e β é independente de x.

O valor de z_0 é calculado de maneira simples, todavia a constante β possui determinação complexa o que levou a novamente aproximar seu valor para 1.

Finalmente, a capacidade equivalente do agregado de conexões $\widehat{C_f}$, será dada pela soma das N capacidades equivalentes individuais das fontes ($\widehat{c_i}$):

$$\widehat{C_f} = \sum_{i=1}^{N} \widehat{c_i} \ . \tag{6}$$

4.1.2 Aproximação Estacionária

Quando considera-se o efeito de multiplexação estatística, as capacidades equivalentes obtidas por meio da aproximação de fluxo contínuo tendem a ser superestimadas. Por isso, a melhor aproximação a ser utilizada é uma distribuição estacionária da taxa binária. A capacidade equivalente do agregado de fontes será

$$C_g = m + \alpha\sigma \ , \tag{7}$$

onde

$$\begin{aligned} \alpha &= \sqrt{-2\log(\varepsilon) - 2\log(2\pi)} \ , \\ m &= \text{taxa média do agregado das conexões e} \\ \sigma &= \text{desvio padrão do agregado das conexões} \ . \end{aligned}$$

Entretanto, como esta aproximação não considera o tamanho do *buffer*, o valor obtido para a capacidade equivalente do agregado de fontes é superestimado, pois seria possível o tráfego exceder a capacidade do canal em intervalo curto de tempo, quando as células excedentes seriam armazenadas em *buffer*.

Finalmente, como nos dois casos o valor obtido tende a ser superestimado e representa situações bem distintas, o valor da capacidade equivalente para o agregado das conexões será o valor mínimo dentre os dois calculados.

4.2 Extensão proposta por Choudhury *et al.*

A proposta de mecanismo de CAC de Choudhury *et al.* [21] encontra-se inserida nesta classe de alocação de capacidades e, mais do que isso, constitui uma extensão do mecanismo Capacidade Equivalente. Em seus estudos, foram apontados casos em que a aproximação assintótica exponencial fornece valores superestimados para as capacidades equivalentes. Com isso, há um conseqüente subaproveitamento dos recursos alocados em um canal de transmissão. Esta proposta visa ampliar o espectro de situações com soluções mais acuradas.

Foi desenvolvida uma aproximação mais acurada para a expressão assintótica para a probabilidade de transbordo de *buffer*, denominada aproximação de três termos. De fato, forma-se por uma soma de três parcelas:

$$Pr(W > x) \approx \alpha_1 e^{-\eta_1 x} + \alpha_2 e^{-\eta_2 x} + \alpha_3 e^{-\eta_3 x}, \tag{8}$$

onde os valores α_1 e η_1 correspondem aos valores anteriormente discutidos para a constante assintótica e a taxa de decaimento do método Capacidade Equivalente na aproximação de fluxo contínuo. Já os valores de α_2, η_2, α_3 e η_3 são calculados em função da probabilidade de atraso $Pr(W > 0)$ e dos três primeiros momentos $E(W)$, $E(W^2)$ e $E(W^3)$, sendo que η_2 e η_3 satisfazem a relação $\eta_1 \leq min(\eta_2, \eta_3)$.

Foram evidenciadas eventuais situações em que a aproximação de fluxo contínuo simples tende a se distanciar dos valores reais. Pode-se destacar casos em que o número de fontes aumenta, a utilização do canal diminui ou o tamanho do *buffer* diminui. A aproximação de três termos alarga o campo de abrangência do método sem afetar a eficiência por uma possível superestimação da capacidade equivalente.

4.3 Aproximação Não-Paramétrica

A aproximação não-paramétrica constitui um método que não se baseia em um modelo do processo de chegadas. Este método consiste em calcular o valor máximo para a taxa de células, dado o conjunto de parâmetros de QoS especificados. O valor obtido será referente ao pior caso e será utilizado para a aceitação (ou não) da chamada [22].

Sejam ANA e MNA, respectivamente, o número médio e o número máximo de células que chegam em um intervalo de tempo fixo. Seja K o tamanho do *buffer*, $K/2$ o número de células transmitidas em tal intervalo e N o número de fontes multiplexadas em um canal. O limite superior de probabilidade de perda de células (CLP - *Cell Loss Probability*), obtido em função das probabilidades de chegada de j células da i-ésima conexão, $p_i^*(j)$ ($i = 1, 2, \cdots, N$ e $j = 0, 1, 2, \cdots$), será:

$$\begin{aligned} CLP &\leq B^*(p_1^*, \cdots, p_N^*; K/2) \\ &\triangleq \frac{\sum_{k=0}^{\infty} [k - K/2]^+ p_1^* \star \cdots \star p_N^*(k)}{\sum_{k=0}^{\infty} k \cdot p_1^* \star \cdots \star p_N^*(k)}, \end{aligned} \tag{9}$$

onde $\star$ denota a convolução e $[x]^+ = \begin{cases} x & , x \geq 0 ; \\ 0 & , x < 0 . \end{cases}$

A distribuição que representa o pior caso será do tipo Bernoulli, $\theta_i(j)$. Poderá ser obtida através dos parâmetros ANA e MNA:

$$\theta_i(j) = \begin{cases} ANA_i/MNA_i & , j = MNA_i \ ; \\ 1 - ANA_i/MNA_i & , j = 0 \ ; \\ 0 & , \text{caso contrário.} \end{cases} \tag{10}$$

Pode-se mostrar que o limite superior para a probabilidade de perda de células será:

$$\begin{aligned} CLP &\leq B^*(p_1^*, \cdots, p_N^*; K/2) \ , \\ &\leq B^*(p_1^*, \cdots, p_{N-1}^*, \theta_N; K/2) \ , \\ &\leq B^*(p_1^*, \cdots, \theta_{N-1}, \theta_N; K/2) \ , \\ &\quad \vdots \\ &\leq B^*(\theta_1, \cdots, \theta_N; K/2) \ , \\ &= \frac{\sum_{k=0}^{\infty} [k - K/2]^+ \theta_1^* \star \cdots \star \theta_N^*(k)}{\sum_{k=0}^{\infty} k \cdot \theta_1^* \star \cdots \star \theta_N^*(k)} \ . \end{aligned} \tag{11}$$

Finalmente, a chamada será aceita caso o valor obtido em $B(\theta_1, \cdots, \theta_{N+1}; K/2)$ for menor do que a probabilidade admissível de perda de células.

4.3.1 Método Dinâmico

Com a hipótese de que nem sempre o usuário conhece os parâmetros de QoS pretendidos para sua aplicação, surgiu a idéia de métodos dinâmicos. Por razões mais diversas, o usuário possivelmente não saberá especificar seus requisitos de QoS, nem mesmo estimá-los. Um mecanismo adaptável dinamicamente seria altamente aconselhável neste caso. Com este intuito surgiu o CAC dinâmico [4].

O método dinâmico apóia-se na aproximação não-paramétrica, uma vez que muitas características do tráfego gerado pela fontes não são conhecidas. Baseia-se na distribuição do número de células que chegam em um intervalo de tempo, usado como período de estimação, que, por sua vez é composto por vários períodos de medição.

Para multiplexação de N conexões, com período de transmissão $K/2$, seja $p_i^*(j)$ a probabilidade de j chegadas na conexão i. Para aceitar uma nova conexão com $MNA = R_{N+1}$ e $ANA = a_{N+1}$, o limite superior para a CLR pode ser calculado a partir da equação 11 obtida para aproximação não-paramétrica:

$$P_{\mathrm{CLR}} \leq \frac{\sum_{k=0}^{\infty}[k-K/2]^{+}p_1^* \star \cdots \star p_N^* \star \theta_{N+1}(k)}{\left(\sum_{k=0}^{\infty} k \cdot p_1^* \star \cdots \star p_N^*(k) + a_{N+1}\right)} \,. \tag{12}$$

Com base na medição do fluxo de células e obtenção de distribuição do número de chegadas, pode-se estimar para as probabilidades p_i^* por $\hat{p}(i)$. Denotando-se o vetor de probabilidades estimadas por $\hat{p} = [\hat{p}(0), \hat{p}(1), \ldots]$ e tomando-se os valores dos parâmetros da nova conexão pode-se estimar o limite superior da nova CLP, $U_B^{'}$, por

$$U_B^{'} = \frac{\sum_{k=0}^{\infty}[k-K/2]^{+}\hat{p} \star \theta_{N+1}(k)}{\left(\sum_{k=0}^{\infty} k\hat{p}(k) + a_{N+1}\right)} \,. \tag{13}$$

Então, a chamada será aceita se o limite superior calculado com a estimativa para a CLR atender os requisitos de QoS pretendidos.

5 Políticas de Escalonamento

Um novo fator a se considerar em estimativa para o controle de admissão é a política de escalonamento a ser utilizada. Uma vez que o problema relacionado ao CAC constitui um problema de filas, uma possibilidade adicional é estudar a disciplina de serviço aplicada à rede em questão. É necessário conhecê-la para a correta estimativa do número de conexões a serem permitidas. Além disso, pode-se dispensar um tratamento diferenciado às classes de serviço. Trata-se de uma vantagem, visto que as diversas classes existentes possuem variadas restrições. Por exemplo, existem serviços com menor tolerância ao atraso, por exemplo, ao passo que outros toleram-no bem, porém são bem sensíveis ao erro.

Uma política de escalonamento de pacotes possível e simples é a bem conhecida FCFS (*First Come First Served*). Seu funcionamento está bem resumido na sua denominação, pois os pacotes serão escalonados em ordem de chegada de tal modo que o que o primeiro a chegar será o primeiro a ser servido. Como se observa, esta disciplina não permite qualquer prioridade entre os pacotes de acordo com a QoS, pois a única condição a ser respeitada é, efetivamente, a ordem de chegada.

Algumas novas políticas de escalonamento vêm sendo propostas como forma de possibilitar uma melhor disponibilidade de atendimento às classes de serviço. De uma maneira geral, as disciplinas associam prioridades aos pacotes para que se atinja os requisitos e eficiência pretendidos.

A disciplina *Static Priority* (SP) associa a cada conexão um determinado nível de prioridade diferente e sempre seleciona para transmissão o pacote de mais alta prioridade com o tempo de chegada mais recente. Este esquema é relativamente simples, pois pode ser implementado com um número fixo de filas FCFS com os diferentes níveis de prioridade associados (figura 9). Liebeherr et al.[10] apresentam este escalonador para o caso de

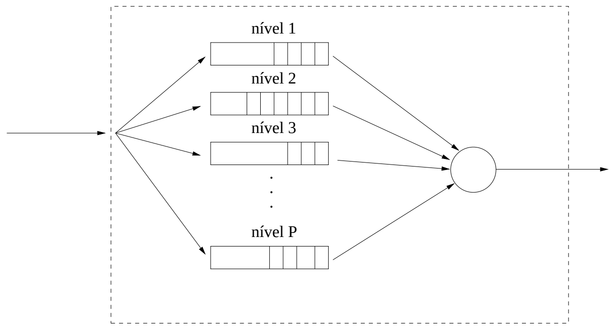

Figura 9: Esquema geral para um escalonador SP.

serviços com diferentes restrições ao atraso e modelo de tráfego (σ, ρ), onde σ é o índice de rajada e ρ o índice de vazão. Considera-se P níveis de prioridade, sendo que cada conexão possuirá um nível p de prioridade, com $1 \leq p \leq P$. Os menores índices de prioridade significam maior prioridade. Ao iniciar um período de ocupação ou ao terminar a transmissão de um pacote, o escalonador SP procurará uma fila FCFS não vazia de maior prioridade. O limite para o atraso será tanto menor quanto for maior a prioridade da fila. Denota-se C_p o conjunto de conexões com prioridade p, M_p o tempo máximo de transmissão de pacotes para estas conexões, $m_p \geq 0$ o tempo mínimo e d_p o seu limite para o atraso. $A_j(t)$ é uma função que traduz os limites do tráfego de chegadas em uma conexão j e está relacionada ao modelo de tráfego utilizado. A condição de escalonabilidade [10] é expressa por:

"Um conjunto N de conexões dado por $\{A_j, d_j\}_{j \epsilon N}$ é escalonável se e somente se para todas as prioridades p e para todo $t \geq 0$ existe τ com $\tau \leq d_p - m_p$ tal que

$$t + \tau \geq \sum_{j \epsilon C_p} A_j(t) + \sum_{q=1}^{p-1} \sum_{j \epsilon C_q} A_j((t+\tau)^-) - m_p + max_{r>p} M_r \text{ .''} \quad (14)$$

A política *Earliest Deadline First* (EDF) [9] irá associar a cada pacote que recebe um tempo de expiração (*deadline*) que é calculado como função do somatório dos tempos de chegada no escalonador e do limite de atraso d_j. O algoritmo determina que seja selecionado sempre o pacote com o prazo mais próximo de expiração. O teste de escalonabilidade para este algoritmo será mais simples do que o necessário para o SP [10]. Entretanto, há necessidade de uma ordenação dos pacotes na fila de acordo com os deadlines associados e conseqüente busca de posição para o encaminhamento de novo pacote para transmissão.

Uma variação da disciplina EDF é a chamada *Rotating Priority Queues* (RPQ) proposta por Liebeherr *et al.* [10] para garantia de atraso máximo. Este algoritmo dispensa a ordenação dos pacotes na fila, representando uma vantagem e podendo ser implementado como o escalonador SP por meio de um conjunto de filas FCFS. Tal como o escalonador SP, estas filas estarão associadas a diferentes prioridades e os pacotes de uma mesma fila FCFS possuem o mesmo limite para o atraso. O algoritmo determina a colocação dos pacotes de acordo com a prioridade desejada. Utiliza um parâmetro fixo Δ, denominado intervalo de rotação. Para cada $n \geq 1$, a fila de índice n (n-fila) será renomeada como (n-1)-fila ao fim de um intervalo de rotação. A fila de índice 0 assumirá índice n_p, dado que o escalonador RPQ comporta $n_p + 1$ filas FCFS (figura 10). Esta rotação desempenhada pelo escalonador garante que os pacotes não irão exceder o atraso máximo tolerado [10].

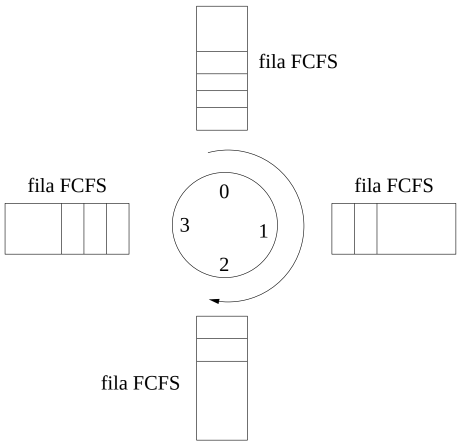

Figura 10: Esquema geral para um escalonador RPQ com $n_p = 3$.

O algoritmo *Generalized Processor Sharing* (GPS) [8] é também conhecido como *Weighted Fair Queueing*. Trata-se de uma disciplina de escalonamento dita conservativa. O tráfego das sessões é tratado como fluido infinitamente divisível. Em [23] é realizada a análise estatística desta disciplina. Para n sessões compartilhando um servidor GPS com taxa r, associa-se a cada sessão um parâmetro ϕ_i que determina o compartilhamento mínimo de largura de faixa. Cada sessão possuirá uma taxa mínima de serviço expressa por $g_i = \frac{\phi_i}{\sum_{j=1}^{n} \phi_j} r$. Presti *et al.* [24] apresentaram este escalonador para servir Fontes de Fluido Moduladas por Markov (*Markov Modulated Fluid Sources* - MMSF), podendo-se estender os resultados para os MMPP's. A partir de um servidor GPS com duas filas, pode-se realizar aproximações para expressar o compartilhamento de canal e determinação de alocação de capacidades às sessões [25]. Zhang *et al.* [26] apresentam um controle de admissão sob esta disciplina com a probabilidade de perda como requisito de QoS e em duas abordagens: baseada em sessões e baseada em classes de serviço. Evidenciou-se que, para duas classes e probabilidade de perda no *buffer* como único requisito de QoS, o escalonador FCFS supera o desempenho do servidor GPS. Ao considerar-se dois requisitos de QoS, perda e garantia de atraso máximo, o escalonador GPS fornece maior flexibilidade no compartilhamento de largura de faixa.

6 Conclusão

O controle de congestionamento preventivo é uma necessidade atualmente por causa das altas velocidades existentes nos enlaces que geram sobrecarga nas redes se efetuar apenas um controle de congestionamento reativo. Ao mesmo tempo, o avanço crescente nas aplicações multimídias está intimamente relacionado à garantia de QoS que as redes ATM oferecem. Neste trabalho, foi descrita a caracterização de tráfego gerado pelas fontes, em particular, fontes de voz e de vídeo.

Nos modelos UAS e CMTC para fontes de voz, as fontes são consideradas independentes e os processos de chegadas são representados por um processo de renovação. Por outro lado, o modelo MMPP não considera o processo de chegada de uma única fonte. Em vez disso, o processo de chegadas está associado ao tráfego agregado das fontes.

O grau de complexidade destes modelos pode ser analisado. O UAS utiliza filas infinitas e apresenta fórmulas fechadas para o cálculo dos parâmetros pretendidos. Já o CMTC e o MMPP são resolvidos por métodos numéricos, sendo que o MMPP, de acordo com o grau de complexidade, pode se tornar computacionalmente extremamente custoso.

Para fontes de vídeo, de uma forma geral, os modelos consideram as fontes independentes e sempre colocam alta correlação entre as chegadas sucessivas. Neste aspecto, o modelo MMPP busca representar variações de curto e de longo intervalo. A principal implicação disto está na complexidade da análise para obtenção dos parâmetros desejados.

O estudo dos mecanismos de CAC está relacionado à alocação de recursos para as conexões que compartilham um mesmo canal em uma rede. Os mecanismos propostos até hoje são numerosos e diferem nas abordagens. Com base nisso, apresenta-se uma classificação.

Os métodos de alocação de capacidades buscam associar a cada fonte uma capacidade efetiva de acordo com os requisitos de QoS pretendidos. Um dos principais mecanismos nesta classe é o método da capacidade equivalente. Utiliza duas aproximações para situações distintas, porém tende a superestimar a capacidade o que pode influir negativamente em sua eficiência. Contudo, existem extensões que alargam o espectro de possibilidades deste método.

A aproximação não-paramétrica possui a vantagem de não necessitar dos modelos de tráfego gerado pelas fontes. Entretanto, para não ocorrerem efeitos indesejáveis na qualidade de serviço que se oferece à conexão, faz-se uma estimativa considerando-se o pior caso. Por isso, os resultados tendem a ser superestimados. O chamado CAC dinâmico constitui uma extensão a este método. Baseia-se nesta aproximação e em medidas do tráfego para conferir dinamismo ao mecanismo. Em virtude disto, o efeito da superestimação tende a ser reduzido, trazendo eficiência ao método. Resta a dúvida, se estas medidas podem ser feitas em tempo hábil para se realizar as ações necessárias.

O estudo de políticas de escalonamento oferece a perspectiva atraente de maior grau de diferenciação entre classes de serviço quanto aos seus requisitos de QoS. A análise estatística do escalonador GPS permite que se implemente um controle de admissão sob suas condições, embora seja complexo seu tratamento. A utilização da disciplina RPQ pode ser aplicada de maneira similar pois sua implementação é relativamente simples e eficiente em termos de garantia de atraso máximo. Como o mecanismo de aproximação não-paramétrica apenas considera os requisitos de QoS, uma alternativa possível é utilizá-lo juntamente com um escalonador que ofereça prioridade a fim de que as diferentes conexões sejam reunidas em classes comuns de serviço. O mesmo princípio pode ser estendido para a alocação de capacidades.

Finalmente, a escolha do mecanismo de CAC pressupõe o conhecimento do tráfego na rede. Deve ser observada a faixa de parâmetros de tráfego e verificar qual mecanismo se adequa melhor à situação em que se encontra a rede. Essencialmente, os mecanismos deves ser rápidos e abrangentes e possibilitar eficácia na multiplexação estatística sem denegrir a QoS das conexões estabelecidas na rede.

Referências

[1] Rec. I.371, *Traffic Control and Congestion Control in B-ISDN*. ITU-T, março 1993.

[2] D. Hong e T. Suda, "Congestion control and prevention in ATM networks", *IEEE Network*, pp. 10–15, julho 1991.

[3] R. Guérin, H. Ahmadi e M. Naghshineh, "Equivalent capacity and its application to bandwidth allocation in high-speed networks", *IEEE Journal on Selected Areas in Communications*, vol. 9, no. 7, pp. 968–981, setembro 1991.

[4] H. Saito e K. Shiomoto, “Dynamic call admission control in ATM networks”, *IEEE Journal on Selected Areas in Communications*, vol. 9, no. 7, pp. 982–989, setembro 1991.

[5] K. Sohraby, “Flow and admission control of ON-OFF sources in high speed networks”, in *IEEE GLOBECOM'94 - Tutorial 7*, San Francisco, novembro 1994.

[6] W. Fischer e K. Meier-Hellstern, “The markov-modulated poisson process (MMPP) cookbook”, *Performance Evaluation*, vol. 18, pp. 149–171, março 1991.

[7] H. Heffes e D. M. Lucantoni, “A markov modulated characterization of packetized voice and data traffic and related statistical multiplexer performance”, *IEEE Journal on Selected Areas in Communications*, vol. 4, no. 6, pp. 856–868, setembro 1986.

[8] A. K. Parekh e R. G. Gallager, “A generalized processor sharing approach to flow control in integrated services networks”, *IEEE/ACM Trans. on Networking*, vol. 1, no. 3, pp. 344–357, junho 1993.

[9] V. Firoiu, J. Kurose e D. Towsley, “Efficient admission control for EDF schedulers”, in *Proc. IEEE INFOCOM'97*, 1997.

[10] J. Liebeherr, D. E. Wrege e D. Ferrari, “Exact admission control for networks with a bounded delay service”, *IEEE/ACM Transactions on Networking*, pp. 885–901, dezembro 1996.

[11] R. Jain, S. Kalyanaraman e R. Goyal, “Source behavior for ATM ABR traffic management: an explanation”, *IEEE Communications Magazine*, pp. 50–57, novembro 1996.

[12] M. F. Abdalla, “Análise de mecanismos de controle de admissão de conexão para redes ATM”, Tese de Mestrado, Universidade Federal do Rio de Janeiro, PEE/COPPE/UFRJ, setembro 1996.

[13] D. Hong e T. Suda, “Congestion control and prevention in ATM networks”, *IEEE Network*, pp. 10–15, julho 1991.

[14] L. Tratjković e J. Golestani, “Congestion control for multimedia services”, *IEEE Network*, pp. 20–26, setembro 1992.

[15] A. E. Eckberg, “B-ISDN/ATM traffic and congestion control”, *IEEE Network*, pp. 28–37, setembro 1992.

[16] ITU-T, *Recommendation I-371: Traffic Control and Congestion Control in B-ISDN*, março 1993.

[17] S. Abe e T. Soumiya, “A traffic control method for service quality assurance in an atm network”, *IEEE Journal on Selected Areas in Communications*, vol. 12, no. 2, pp. 322–331, fevereiro 1994.

[18] C. Shim, I. Ryoo, J. Lee e S. Lee, “Modeling and call admission control algorithm of variable bit rate video in atm networks”, *IEEE Journal on Selected Areas in Communications*, vol. 12, no. 2, pp. 332–344, fevereiro 1994.

[19] H. G. Perros e K. M. Elsayed, “Call admission control schemes: a review”, *IEEE Communications Magazine*, pp. 82–91, novembro 1996.

[20] R. Guérin, H. Ahmadi e M. Naghshineh, “Equivalent capacity and its application to bandwidth allocation in high-speed networks”, *IEEE Journal on Selected Areas in Communications*, vol. 9, no. 7, pp. 968–981, setembro 1991.

[21] G. L. Choudhury, D. M. Lucantoni e W. Whitt, “Squeezing the most out of ATM”, *IEEE Transactions on Communications*, vol. 44, no. 2, pp. 203–217, fevereiro 1996.

[22] H. Saito, “Call admission control in ATM networks using upper bound of cell loss probability”, *IEEE Transactions on Communications*, vol. 40, no. 9, pp. 1512–1521, setembro 1992.

[23] Z.-L. Zhang, J. Kurose e D. Towsley, “Statistical analysis of generalized processor sharing scheduling discipline”, *IEEE Journal on Selected Areas in Communications*, vol. 13, no. 6, pp. 1071–1080, agosto 1995.

[24] F. lo Presti, Z.-L. Zhang e D. Towsley, “Bounds, approximations and applications for a two-queue gps system”, in *Proc. IEEE INFOCOM'96*, 1996.

[25] V. Firoiu e D. Towsley, “Call admission and resource reservation for multicast sessions”, in *Proc. IEEE INFOCOM'96*, 1996.

[26] Z.-L. Zhang, , Z. Liu, J. Kurose e D. Towsley, “Call admission control schemes under the generalized processor sharing scheduling discipline”, in *Telecommunications Systems*, 1996.