Capítulo 4

Controle de Admissão de Conexão

4.1 Introdução

Umas das grandes vantagens na utilização das redes ATM está no ganho de multiplexação estatística que pode ser obtido através da superposição de diversas fontes em um mesmo canal de transmissão. Assim, quanto maior for o número de fontes superpostas, maior será este ganho. Entretanto, o aumento indiscriminado do número de fontes que compartilham o canal de transmissão pode levar a estados de congestionamento e à degradação da qualidade do serviço prestada pela rede a estas fontes devido a perda de células e introdução de atrasos. Faz-se, portanto, necessário o uso de mecanismos de controle de tráfego capazes de atender de forma eficiente ao maior número possível de conexões e de garantir a veiculação do tráfego contratado. Parâmetros de qualidade de serviço, como taxa de perda, atraso etc, e de descrição do tráfego, como taxa binária média, taxa binária de pico etc, devem ser especificados, monitorados e garantidos pela rede.

O conceito de controle de tráfego em redes ATM é bem simples: um usuário declara as características de tráfego de sua comunicação na fase de estabelecimento de conexão e a rede, uma vez aceitando a conexão, deve assegurar os recursos necessários para garantir as características negociadas. Caso seja impossível a alocação destes recursos, o pedido de conexão é rejeitado. Durante a comunicação, a rede monitora a conformidade entre as características de tráfego declaradas e as medidas. Se não houver a conformidade, uma penalidade é atribuída ao usuário.

O controle que aloca os recursos de rede e decide sobre a aceitação ou não da conexão é chamado de Controle de Admissão de Conexão (CAC - *Connection Admission Control*) e o controle que monitora o fluxo de células é chamado de Controle dos Parâmetros do Usuário (UPC - *Usage Parameter Control*). O controle de parâmetros do usuário foi objeto de estudo no GTA em duas outras teses de mestrado [25, 23]. Este trabalho enfoca prioritariamente o Controle de Admissão de Conexão.

O Controle de Admissão de Conexão (CAC) é definido pela ITU-T [3] como o conjunto de ações tomadas pela rede durante a fase de estabelecimento de conexão (ou durante a fase de renegociação da chamada) para estabelecer quando um pedido de conexão de canal virtual/caminho virtual pode ser aceito ou rejeitado (ou quando o pedido de realocação pode ser acomodado). O roteamento faz parte das ações do CAC, uma vez que a rede deve escolher o caminho que, ao mesmo tempo, permita a maior eficiência (maior número de conexões) e garanta a QoS desejada.

Para suportar toda a flexibilidade inerente às redes ATM, há vários requisitos que um mecanismo de CAC deve atender. Entre estes requisitos, podem ser citados [5]:

- uma resposta em tempo real é necessária para que se possa tomar uma decisão em um curto intervalo de tempo. Isto faz com que se tenha métodos simples, aproximados ou baseados em tabelas, pois não haveria tempo suficiente para a realização de cálculos numéricos exatos;

- deve existir uma margem de segurança de modo a garantir que a QoS será satisfeita quando todas as fontes estiverem se comportando como o declarado durante o contrato feito no estabelecimento da conexão;

- o mecanismo de admissão deve possuir associado a ele um mecanismo de policiamento (UPC) para verificar a conformidade entre o tráfego declarado e o real de modo a evitar que um tráfego excessivo na rede possa prejudicar a QoS da própria conexão e, o que seria ainda pior, das demais fontes;

- o mecanismo deve ser válido para casos de conexões de tipos variados (por exemplo, multimídia); e

- finalmente, é preciso utilizar eficientemente o ganho estatístico obtido através da multiplexação estatística. Como exemplo, um CAC baseado apenas na taxa de pico seria de simples implementação, mas, no entanto, implicaria num desperdício enorme de capacidades.

Como visto no capítulo 2, cada classe de serviço (CBR, VBR, ABR e UBR) possui as suas próprias características e o seu próprio conjunto de parâmetros descritores de tráfego.

Portanto, o procedimento de CAC deve ser específico para cada uma destas classes.

O CAC para chamadas da classe CBR é bastante simples e similar àquele feito em redes comutadas por circuito. Como a taxa binária da fonte é constante e o único parâmetro utilizado é a taxa binária de pico, a alocação pelo pico é adequada e bastante eficiente. Caso não haja recursos de rede disponíveis para suportar a taxa binária de pico requerida, o pedido de conexão é rejeitado.

No CAC para chamadas da classe ABR, vários parâmetros de tráfego são especificados e a taxa de transmissão a ser fornecida pela rede para a fonte pode ser variável, sendo apenas conhecidos a priori os seu valores mínimo e máximo. Um pedido de conexão será aceito caso haja banda passante disponível para prover a taxa binária mínima requerida, sendo este o valor previamente alocado para a fonte. No entanto, a taxa binária fornecida pela rede pode aumentar ou diminuir conforme o estado da rede.

No CAC para chamadas da classe UBR, nenhuma banda passante é alocada, estando o usuário desta conexão disposto a aceitar qualquer valor de taxa de transmissão e qualquer nível de qualidade de serviço que a rede possa prover.

O CAC para chamadas da classe VBR é bastante diferente dos anteriores, pois fontes com as mais diversas características são superpostas nesta classe. Para fontes cujas taxas médias de transmissão são muito próximas às suas taxas de pico, a adoção da alocação pelo pico é bastante aceitável. Contudo, para a maior parte das aplicações previstas nas RDSI-FL (tais como vídeo codificado e videofone), isto não é verdade e a alocação pelo pico pode resultar em um grande desperdício de banda passante e, conseqüentemente, em um baixo nível de utilização da rede. A multiplexação estatística, como mostrado no capítulo 3 para o caso de fontes em rajadas do tipo silêncio/atividade e VBR, pode ser bastante vantajosa, apesar de sua difícil avaliação.

Este capítulo tem por objetivo o estudo e análise da eficiência dos diversos mecanismos de controle de admissão de conexão encontrados na literatura. Na seção 4.2, é feita uma classificação destes mecanismos. Nas seções 4.3 à 4.8, cada uma destas classe é vista em detalhes. Na seção 4.9, é feita uma análise do impacto da variação do atraso de células (CDV) sobre os mecanismos de CAC. E, por último, na seção 4.10, é feito um resumo dos mecanismos de CAC apresentados.

4.2 Mecanismos de CAC

Tipicamente, o problema de decisão sobre a aceitação ou não de uma nova conexão pode ser formulado como um problema de teoria de filas envolvendo um multiplexador ATM, representado por um conjunto de fontes ATM alimentando uma fila de capacidade finita que é servida por um servidor com um tempo de serviço constante [7], conforme descrito na seção 3.3. Neste problema, procura-se encontrar o número máximo de conexões que podem ser aceitas enquanto estando de acordo com os objetivos de QoS contratados. Entretanto, a solução deste problema de filas é muito árdua e difícil de ser realizada em tempo de execução. Em vista disto, diversos mecanismos de CAC tem sido propostos baseados em diferentes aproximações ou em esquemas diferentes que não necessitam da solução de tal problema de filas.

Vários mecanismos de CAC podem ser encontrados na literatura. Alguns destes mecanismos requerem um modelo específico de tráfego, enquanto outros apenas necessitam dos parâmetros de tráfego, tais como as taxas binárias média e de pico. Devido a esta diversidade, aqui é feita uma classificação destes mecanismos para uma melhor compreensão dos mesmos. Esta classificação está baseada no princípio implícito que foi usado no desenvolvimento destes mecanismos, conforme feito em [7]. As classes de CAC estão representadas pelos seguintes grupos: método de alocação de capacidades, método de reservas rápidas, aproximação de tráfego intenso, aproximação não-paramétrica, método dinâmico e método híbrido.

O método de alocação de capacidades tem por objetivo determinar a capacidade efetiva de cada fonte de tráfego de acordo com a taxa de perda de células desejada. Um pedido de conexão só é aceito se a capacidade disponível do canal for maior do que a capacidade efetiva da fonte. Caso contrário, o pedido é rejeitado.

O método de reservas rápidas foi desenvolvido para fontes de tráfego em rajadas. Neste método, quando uma conexão é estabelecida, nenhum recurso de rede é previamente alocado, sendo apenas determinadas as rotas através da rede de cada conexão. No momento em que a fonte deseja transmitir, a rede tenta alocar os recursos necessários apenas para a duração da rajada.

A aproximação de tráfego intenso está baseada 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 de células, aproximada pela cauda da distribuição do tamanho da fila, seja menor que o valor desejado.

A aproximação não-paramétrica não assume nenhum modelo específico de tráfego e utiliza apenas os parâmetros de tráfego. A decisão de aceitação ou não de uma nova conexão é feita com base no limite superior que a taxa de perda de células pode assumir.

O método dinâmico é baseado na distribuição do número de células que chegam durante um determinado período de estimação. Cada período de estimação é composto por vários períodos de medição e é chamado de período de renovação. Um pedido de conexão só é aceito se sua capacidade equivalente for menor do que a capacidade equivalente estimada disponível no canal.

O método híbrido lida com fontes de tráfego com objetivos de qualidade de serviço (QoS) diferentes. São providas classes de alta e de baixa qualidades e um mecanismo de prioridades no *buffer* de saída. A decisão de aceitação de um pedido de conexão é feita com base apenas na sua taxa de pico e na medição da taxa de chegadas de células do tráfego agregado. Para uma conexão de alta qualidade é feita a alocação pelo pico. A banda passante que não é usada pelas conexões de alta qualidade pode, no entanto, ser usada pelas conexões de baixa qualidade.

4.3 Método de Alocação de Capacidades

O método de Alocação de Capacidades tem por objetivo alocar uma capacidade equivalente para cada conexão de modo que o objetivo de qualidade de serviço (QoS) - normalmente expresso pela taxa de perdas de células (CLR) - seja alcançado.

A figura 4.1 mostra o efeito da multiplexação estatística na alocação de capacidades. Nela são utilizadas fontes de tráfego em rajadas com taxas de pico (B_p) de 10 Mbps, taxas médias de 1 Mbps, comprimento médio de rajada (L) de 100 células e tamanho de *buffer* (K) de 50 células. Na alocação pelo pico, o número máximo de fontes que podem compartilhar um canal de W Mbps é $W/10$, apresentando uma taxa de probabilidade de perda de células quase nula. A perda só é nula para os casos onde o número total de fontes é menor que o tamanho do *buffer* de modo a não haver perda de células no momento em que todas as fontes estão ativas. Na alocação pela média, este número é máximo e igual a W, porém a taxa de perda de células é muito alta. Utilizando os resultados analíticos apresentados na seção 3.4, várias curvas de alocação de capacidades também são apresentadas nesta figura para diversas taxas de perda de células. Por exemplo, para suportar um total de 30 conexões com um objetivo de taxa de perda de células (CLR) de 10^{-6}, seria necessário uma banda passante de 115 Mbps. Já se o objetivo de CLR fosse

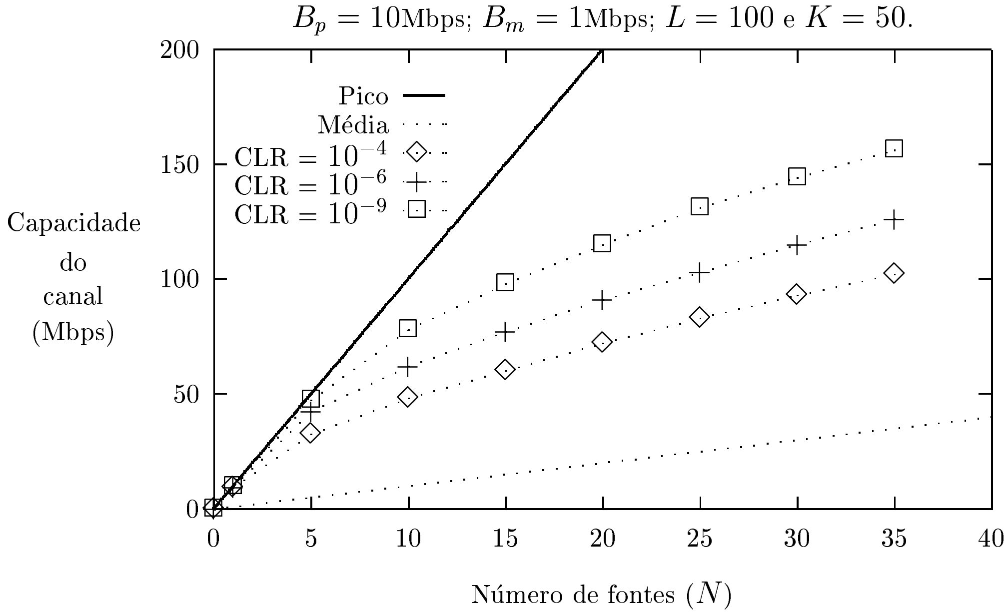

Figura 4.1: Alocação de capacidades para $B_p = 10$Mbps.

de 10^{-9}, seria necessário uma banda passante de 144 Mbps.

Sobre este gráfico duas importantes observações podem ser feitas. A primeira é que a medida que se aumenta o número de fontes multiplexadas, menor é a capacidade adicional que deve ser alocada para uma nova conexão. Isto pode ser comprovado pela curvatura dos gráficos em forma de "saturação". A segunda observação é que diferentes curvas podem ser traçadas para cada objetivo de taxa de probabilidade de perda. Quanto menor for esta taxa de probabilidade de perda, mais próxima será a curva daquela da alocação pelo pico. E quanto maior for a taxa, mais próxima será a curva daquela da alocação pela média.

A figura 4.2 faz uma análise semelhante a da figura 4.1, mas para uma taxa de pico de 3 Mbps. Comparando estas figuras, pode se observar que o decaimento das curvas para o tráfego com menor índice de rajada é mais lento, estando estas curvas bem mais próximas da curva de alocação pelo pico. Isto também mostra que o efeito de multiplexação estatística é menor e que a "saturação" é mais lenta para estes tráfegos, sendo necessário um número bem maior de fontes para que o efeito da multiplexação estatística seja significativo.

A figura 4.3 mostra o efeito da multiplexação estatística na alocação de capacidades para diferentes comprimentos médios de rajada (L) e taxas de perda de células (CLR). Como nos dois últimos exemplos, os dados desta figura foram obtidos utilizando os resultados analíticos apresentados na seção 3.4. Comparando as diversas curvas apresentadas nesta

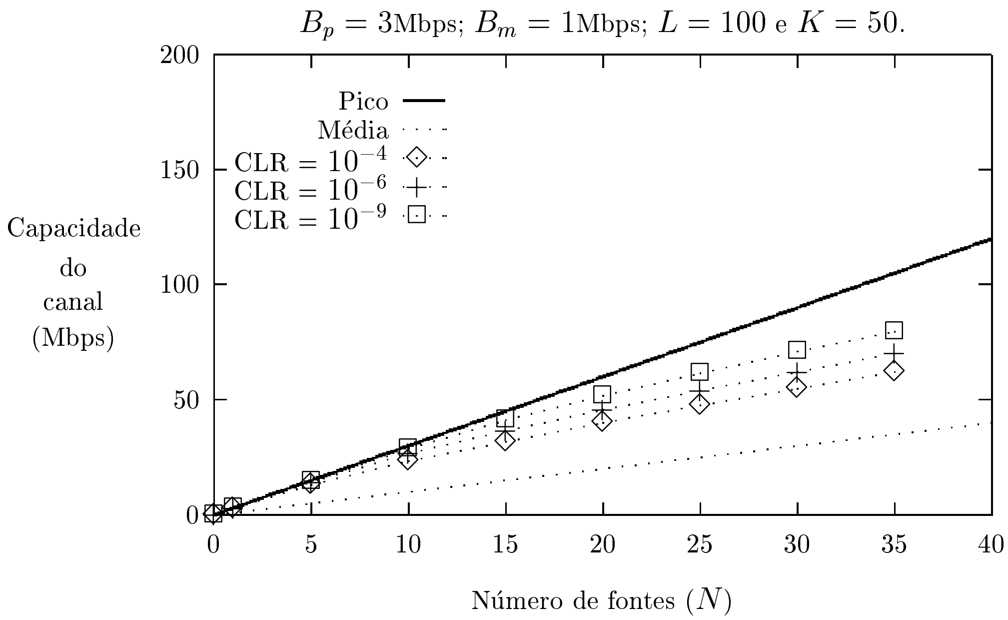

Figura 4.2: Alocação de capacidades para $B_p = 3$Mbps.

figura, observa-se que o decaimento da curva de alocação de capacidades é mais rápido para comprimentos menores de rajada, afastando-se bem mais rapidamente da curva de alocação pelo pico. Este efeito já era esperado, pois, como mostrado no capítulo 3, o efeito de multiplexação estatística é maior para comprimentos de rajadas menores e, portanto, a capacidade a ser alocada por fonte é menor.

Várias estratégias são adotadas no cálculo das capacidades efetivas de cada fonte. Aqui, algumas destas estratégias são abordadas, como os métodos de simulação, os métodos analíticos, a aproximação de estado estacionário e o método da capacidade equivalente.

4.3.1 Métodos de Simulação

O método de simulação usualmente adotado é o orientado a eventos discretos. Neste tipo de simulação, o tempo é gerado de modo descontínuo. A marcação de tempo é incrementada a cada etapa para o tempo correspondente ao próximo evento. Este evento é escolhido dentre todos aqueles possíveis de se realizar a partir do estado atual do sistema. Uma vez escolhido, o estado do sistema é então atualizado novamente e recomeça até que o critério de parada da simulação seja satisfeito.

No caso do ATM, a simulação deve ser feita a nível de célula, de modo a representar o mais fielmente possível o comportamento do multiplexador. Os eventos correspondem

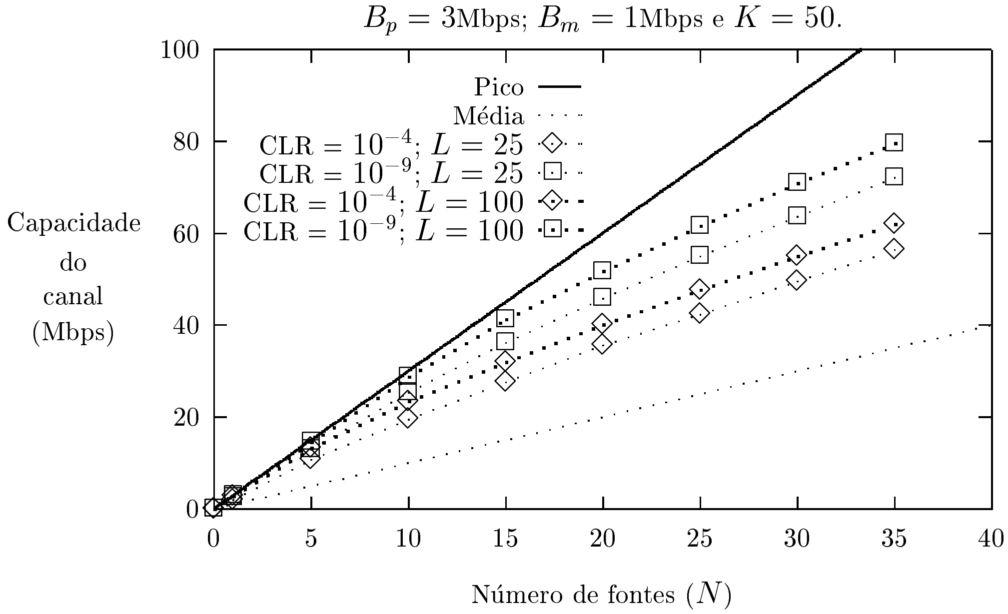

Figura 4.3: Alocação de capacidades para B_p = 3Mbps e diferentes comprimentos de rajada.

ou à chegada ou à partida de uma célula.

A tabela 4.1 mostra a variação do fator de expansão (R) com o número de fontes (N) para dois diferentes índices de rajada (b). Estes resultados foram obtidos usando o simulador QNAP2 [28] com fontes com taxas de pico de 10 Mbps. É importante ressaltar, porém, que, como já mencionado na seção 3.4, estes resultados continuam válidos para quaisquer outros valores de taxa de pico.

A grande vantagem na utilização da simulação é que ela permite a descrição de sistemas complexos, com processos arbitrários de chegada e com diferentes disciplinas de

Tabela 4.1: Variação do fator de expansão com o número de fontes ($CLR = 10^{-4}$).

Variação do Fator de Expansão (R)		
$CLR = 10^{-4}$ $K = 50$ $L = 100$		
Número de fontes	$b = 3$	$b = 10$
1	2.80	9.25
10	2.16	4.50
20	1.88	3.30
30	1.70	2.83

atendimento, sistemas estes difíceis de serem tratados analítica ou numericamente. Sua principal desvantagem é o fato de que para simular eventos raros como taxas de perdas de células de 10^{-9} é necessário um tempo enorme de execução.

Para resolver a desvantagem do tempo de execução, várias técnicas são adotadas. A mais simples delas é a de simular com taxas de perda de células de ordens mais altas e supor que, para taxas mais baixas, o resultado continue válido. Outra técnica possível é a simulação a nível de rajada, e não mais a nível de célula, sendo a probabilidade de perda de células estimada a partir do número de fontes ativas e do padrão de chegada das células destas fontes.

4.3.2 Métodos Analíticos

A solução por métodos analíticos está baseada na caracterização e modelagem das fontes de tráfego e na respectiva resolução dos modelos de filas que representam os multiplexadores.

Como mostrado no capítulo 2, vários modelos de tráfego foram propostos para os mais diversos tipos de tráfego. Estes modelos estão divididos basicamente em dois tipos: os métodos de matrizes geométricas, como o MMPP (seção 2.3.1), e os métodos de fluxo contínuo, como o UAS (seção 2.3.1).

Os métodos de matrizes geométricas são atribuídos a Neuts [16] e se aplicam a matrizes estocásticas com certos tipos de estruturas repetitivas de bloco [5]. Nestes métodos, o multiplexador é representado por um sistema de fila, cujas matrizes de probabilidade de transição são estocásticas do tipo geométrica. Dentre os sistemas com esta características, encontram-se aqueles com processos de chegada do tipo Markoviano ou do tipo MMPP. A capacidade efetiva do tráfego agregado pode ser obtida através da solução de sistemas de filas do tipo MMPP/G/1 ou MMPP/D/1/K. Ela será igual ao menor valor da capacidade do canal que forneça a taxa de perda de células desejada.

Os métodos de fluxo contínuo são aqueles baseados na equação de escoamento de fluidos. Estes métodos não consideram a divisão dos dados em células. Os *bits* são gerados continuamente pelas fontes que se encontram no seu estado ativo, e são servidas pelo multiplexador também a uma taxa contínua. No caso de redes ATM, esta taxa de serviço é constante.

O primeiro trabalho com este tipo de abordagem foi o UAS, já descrito na seção 2.3.1,

Tabela 4.2: Variação do fator de expansão com o número de fontes ($b = 3$).

Variação do Fator de Expansão (R)			
$K = 50$ $L = 100$			
Número de fontes	$CLR = 10^{-4}$	$CLR = 10^{-6}$	$CLR = 10^{-9}$
1	2.82	2.89	2.93
5	2.63	2.82	2.90
10	2.33	2.66	2.87
15	2.12	2.43	2.74
20	2.00	2.28	2.58
25	1.90	2.16	2.46
30	1.83	2.07	2.36
35	1.77	2.00	2.27

onde um número finito de fontes idênticas e independentes estatisticamente (fontes homogêneas) são multiplexadas. Como mostrado nas seções 3.4.1 e 3.4.2 para os casos de fontes em rajadas do tipo silêncio/atividade e VBR, respectivamente, este modelo apresentou uma ótima conformidade com os resultados obtidos por simulação. Como as condições assumidas pelo modelo permanecem válidas para quaisquer valores de taxa de perda de células (CLR), ele pode ser utilizado no cálculo das capacidades efetivas para valores de CLR mais baixos.

As tabelas 4.2 e 4.3 mostram, para índices de rajada (b) iguais a 3 e 10, respectivamente, a variação do fator de expansão (R) com o número de fontes (N) para diversos valores de CLR. Estes resultados foram obtidos analiticamente usando os resultados apresentados na seção 2.3.1 para fontes em rajadas do tipo silêncio/atividade com taxas de pico de 10 Mbps.

Por exemplo, para descobrir qual a capacidade (C) que deve ser alocada para suportar a multiplexação de 25 fontes com índices de rajada iguais a 10, taxas de pico (B_p) de 10 Mbps e um objetivo de CLR de 10^{-6}, basta fazer:

$$C = R(N, b) * N * \frac{B_p}{b} = 4.12 * 25 * \frac{10 \text{ Mbps}}{10} = 103 \text{ Mbps} .$$

A vantagem na utilização do modelo UAS está na sua rapidez e simplicidade na obtenção dos resultados, aliado ao fato do mesmo apresentar uma ótima conformidade com os resultados simulados. A sua principal desvantagem é o fato de apenas serem consi-

Tabela 4.3: Variação do fator de expansão com o número de fontes ($b = 10$).

Variação do Fator de Expansão (R)			
$K = 50 \quad L = 100$			
Número de fontes	$CLR = 10^{-4}$	$CLR = 10^{-6}$	$CLR = 10^{-9}$
1	9.29	9.57	9.73
5	6.52	8.44	9.48
10	4.80	6.20	7.80
15	4.00	5.13	6.53
20	3.60	4.55	5.75
25	3.32	4.12	5.32
30	3.10	3.83	4.80
35	2.91	3.60	4.49

deradas fontes homogêneas, isto é, fontes de tráfego do mesmo tipo e com os mesmos parâmetros. Para estender o método para o caso de fontes heterogêneas, foram propostas várias extensões, conhecidas como Fluxo Contínuo Estocástico (SFF - *Stochastic Fluid Flow*). No entanto, estas extensões não são aqui analisadas. Para uma melhor referência, ver [13, 5].

4.3.3 Aproximação de Estado Estacionário

Os métodos de aproximação de estado estacionário utilizam uma aproximação da distribuição das taxas de transmissão do tráfego gerado pelo agregado das conexões para o cálculo das capacidades efetivas. Nestes métodos, a taxa de perda de células é aproximada pela probabilidade da fonte estar transmitindo acima da capacidade do canal. Dentre os principais métodos propostos na literatura, destacam-se a aproximação gaussiana e a regra Sigma.

Aproximação Gaussiana

No método da aproximação gaussiana [2], a distribuição das taxas de transmissão do tráfego agregado no estado estacionário é aproximado por uma curva gaussiana. Esta suposição permite estimar a cauda da distribuição estacionária da taxa binária.

Para que haja perdas de células por transbordo no multiplexador, é necessário que a taxa de transmissão do tráfego agregado seja superior à do canal e que o *buffer* esteja cheio. Contudo, neste método, o tamanho do *buffer* não é levado em conta e a taxa de perda de células é aproximada pela probabilidade da taxa de transmissão do tráfego agregado ser superior a do canal.

A capacidade efetiva do agregado das conexões, C_s, é então aproximada por um valor tal que a probabilidade acumulativa da cauda além de C_s não exceda ε (o requisito de QoS desejado), ou seja,

$$\Pr(X > C_s) < \varepsilon \ , \tag{4.1}$$

onde X é taxa de transmissão do tráfego agregado.

Através da inversão da distribuição gaussiana, C_s pode ser aproximada por:

$$C_s = m + \alpha'\sigma \ , \tag{4.2}$$

onde

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

Sendo m_i a taxa binária média da fonte i e R_i a sua taxa binária de pico, os valores de m e σ, calculados de modo a satisfazer os requisitos de QoS, são :

$$\begin{aligned} m &= \sum_i m_i \\ \sigma &= \sqrt{\sum_i m_i(R_i - m_i)} \end{aligned}$$

A figura 4.4 mostra a variação do fator de expansão (definido na seção 3.4) com o número de fontes. O índice de rajada é 10 a taxa binária de pico é 10 Mbps. Além disto, também são apresentados os resultados obtidos através do modelo UAS, para um *buffer* de tamanho 50. Nela, pode-se observar que, a qualidade da aproximação gaussiana será melhor quanto maiores forem o número de fontes e a taxa de perda de células. Nos casos onde o número de fonte é pequeno e a taxa de perda de células é muito baixa, este método tende a superestimar o valor da capacidade necessária. É importante ressaltar que o tamanho do *buffer* usado nesta comparação foi escolhido de forma a melhor avaliar a acurácia desta aproximação, já que a mesma não leva em consideração este parâmetro. É óbvio que, para valores maiores de tamanho de *buffer*, os resultados desta aproximação tendem a ser bastante superestimados.

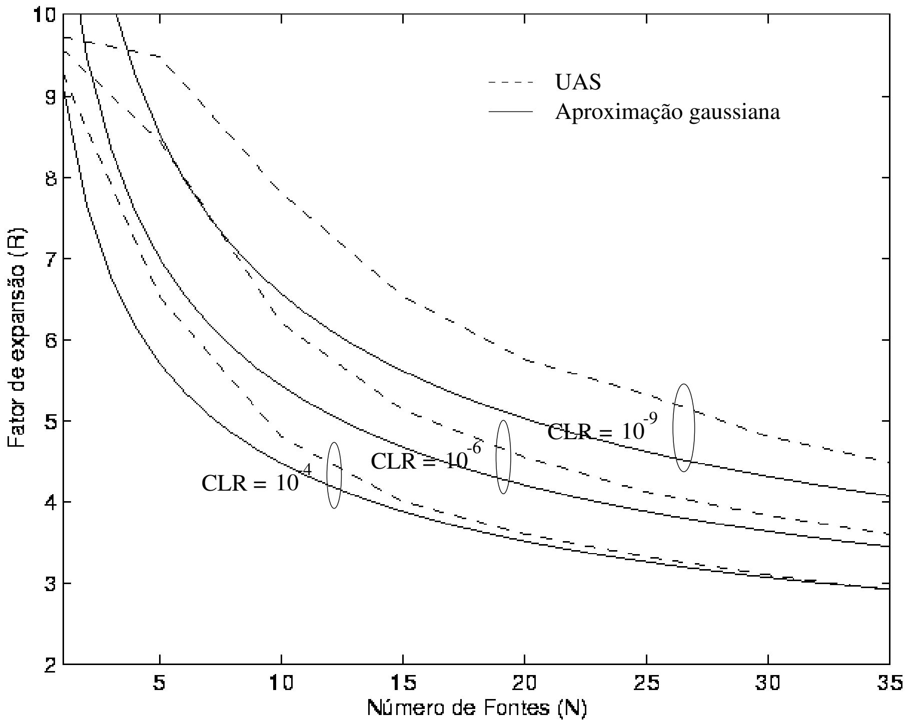

Figura 4.4: Comparação entre a aproximação gaussiana e o modelo UAS ($b = 10$).

A figura 4.5 faz uma análise semelhante a da figura 4.4, mas para o caso de fontes com índice de rajada igual a 3. O tamanho do *buffer* e a taxa de pico são os mesmos do caso anterior. Pode-se observar, nesta figura, que a qualidade da aproximação é pior do que no caso anterior, visto que a convergência entre os resultados é bem mais lenta. Isto se explica pelo fato de que o efeito de multiplexação estatística é menor para fontes com índices menores de rajada, principalmente quando o número de fontes é pequeno. Nestes casos, a aproximação do tráfego agregado por uma distribuição gaussiana tende a superestimar bastante a capacidade de canal necessária.

Como pode ser visto nestas figuras, este método é adequado para os casos onde o efeito de multiplexação estatística é significativo, ou seja, um grande número de fontes com altos índices de rajada e com objetivos de QoS pouco restrigentes sendo multiplexadas. Neste caso, o erro introduzido pela aproximação é pequeno e seus resultados são bem acurados. Porém, mesmo nestes casos, esta aproximação tende a superestimar o valor da capacidade necessária, pois em nenhum momento o tamanho do *buffer* é considerado. Já para os casos onde o número de fontes é pequeno, os objetivos desejados de QoS são mais restrigentes e os índices de rajada das fontes são baixos, a utilização deste método é muito pouco vantajosa, visto que os seus resultados são muito inacurados e tendem a produzir valores bastante superestimados.

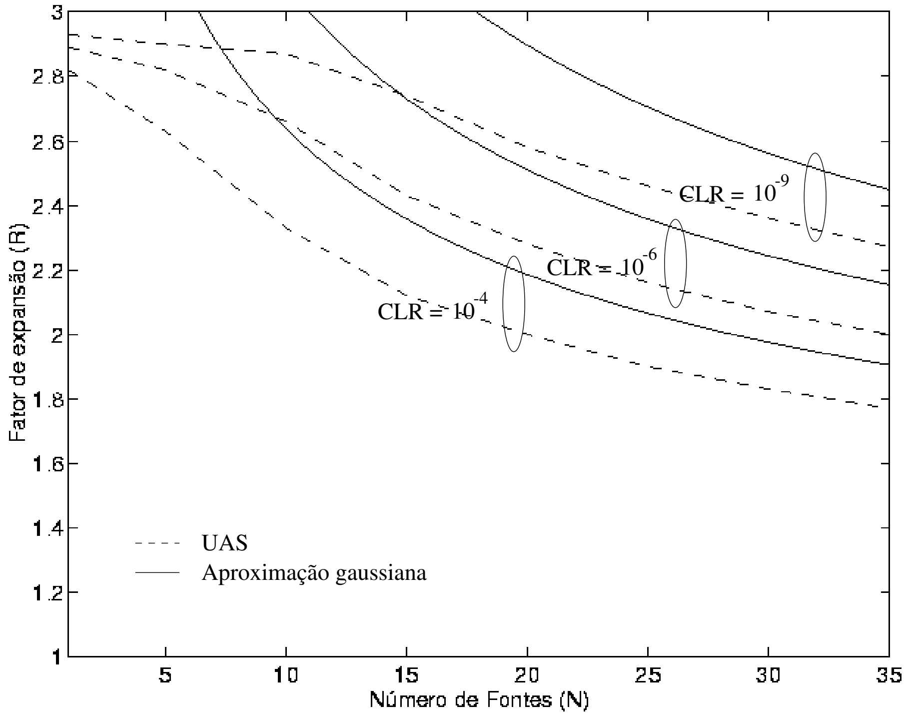

Figura 4.5: Comparação entre a aproximação gaussiana e o modelo UAS ($b = 3$).

Regra Sigma

A regra Sigma [2] é um método aproximado, mas que faz uso de uma classificação simples. Nesta regra são definidos dois tipos de classes: a de alto e a de baixo nível de rajadas. Esta classificação é feita com base nas taxas de transmissão média e de pico das fontes de tráfego.

Para a classe de baixo nível de rajadas, a alocação de banda passante é feita pelo pico, ficando a banda passante restante para ser usada na multiplexação estatística das fontes da classe de alto nível de rajadas.

Para a classe de alto nível de rajadas, o processo de admissão é equivalente ao da aproximação gaussiana. É definida uma função $q(\cdot)$ que representa o ganho estatístico da multiplexação das fontes, determinada anteriormente.

Quando a banda passante disponível para a classe de alto nível de rajada, c', é maior que um valor limite, então a capacidade equivalente, C_{eq}, para as chamadas destas classes é estimada por

$$C_{\text{eq}} = q(c')\sqrt{\sum_i \sigma_i^2} + \sum_i L\, a_i + \max_i L\, R_i \ , \tag{4.3}$$

onde

$$\begin{aligned} \sigma_i &= \text{desvio padrão da taxa binária da fonte } i; \\ L\,a_i &= \text{taxa binária média da fonte } i; \text{ e} \\ L\,R_i &= \text{taxa binária de pico da fonte } i. \end{aligned}$$

Caso a capacidade equivalente estimada seja menor que a banda passante disponível, a chamada é aceita.

4.3.4 Capacidade Equivalente

O método da capacidade equivalente é atribuído a Guérin et al. [29]. Seu objetivo é obter uma expressão unificada para a representação da banda passante efetiva, ou capacidade equivalente, de uma conexão, assim como para o agregado das conexões.

Duas aproximações diferentes são utilizadas no cálculo da banda passante efetiva: uma baseada em modelo de fluxo contínuo e outra, em uma aproximação da distribuição estacionária da taxa binária. Estas aproximações foram selecionadas de modo a se completarem, pois cada uma captura um diferente aspecto do comportamento das conexões multiplexadas, ao mesmo tempo que permanecem computacionalmente simples. A primeira aproximação estima de um modo acurado a capacidade equivalente quando o impacto das características de uma fonte individual é crítica. A segunda aproximação é representativa nos casos em que a multiplexação estatística das conexões é significativa.

Aproximação de Fluxo Contínuo

Na aproximação de fluxo contínuo, uma única fonte de tráfego é modelada por uma cadeia de Markov de 2 estados (fontes em rajadas do tipo silêncio/atividade), correspondentes aos períodos de atividade e silêncio. A caracterização de uma fonte é feita pela sua taxa binária de pico (R), pela utilização (ρ), e pelo tempo médio de duração da rajada (L). Note que a taxa de utilização é igual ao inverso do índice de rajada (b). A distribuição do tamanho da fila, quando tal fonte está alimentando uma fila com um servidor de taxa de serviço constante, pode ser derivada usando-se técnicas padrão para o caso de filas tanto de fontes idênticas quanto de fontes representadas por cadeias de Markov mais gerais [29]. Com esta distribuição, é possível determinar a taxa de serviço, ou capacidade equivalente, necessária para alcançar a QoS desejada. Assumindo uma fila de tamanho finito k, a equação satisfeita pela capacidade equivalente c para uma probabilidade de transbordo ε (o requisito de de QoS desejado) é da forma:

$$\varepsilon = \beta \ \exp\left(\frac{c - R\rho}{L(1-\rho)(R-c)c}\, k\right) , \tag{4.4}$$

onde

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

A capacidade equivalente para uma única fonte pode ser obtida resolvendo-se a equação 4.4 para c. Entretanto, como nenhuma expressão fechada pode ser obtida para a capacidade equivalente, a equação 4.4 deve ser resolvida numericamente. Devido a isto, Guérin et al. propuseram uma aproximação fixando o valor de β em 1. Neste caso, a capacidade equivalente $\hat{c}$ de uma única fonte isolada é dada por:

$$\hat{c} = \frac{f_i - k + \sqrt{(f_i - k)^2 + 4k\rho_i f_i}}{2\alpha\, b_i(1-\rho_i)} , \tag{4.5}$$

onde

$$\begin{aligned} \alpha &= \ln(1/\varepsilon) \ ; \\ f_i &= \alpha\, b_i(1-\rho_i)R_i \ ; \\ k &= \text{tamanho do } \textit{buffer}. \end{aligned}$$

No caso de múltiplas fontes, uma aproximação assintótica é utilizada para o cálculo da probabilidade de transbordo do *buffer*, onde apenas a contribuição do autovalor mais negativo associado ao sistema é considerada. Sob tais circunstâncias, a probabilidade de transbordo do *buffer* $G(x)$ é da forma

$$G(x) \approx \beta\ e^{z_0 x} , \tag{4.6}$$

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

Entretanto, enquanto z_0 pode ser facilmente obtido através de equações simples, a determinação da constante β é muito difícil. Em vista disto, novamente Guérin et al. aproximaram o valor de β por 1, o que equivale a não considerar o ganho de multiplexação estatística. Com isto, a capacidade equivalente do agregado de fontes, $\widehat{C_f}$, é dada pela soma das N capacidades equivalentes individuais das fontes.

$$\widehat{C_f} = \sum_{i=1}^{N} \widehat{c_i} \tag{4.7}$$

Esta linearização explícita indica que a capacidade equivalente do agregado é superestimada, a menos das situações em que as capacidades equivalentes individuais são próximas às suas taxas médias.

A figura 4.6 mostra a variação da capacidade equivalente com tamanho médio da rajada, para dois diferentes níveis de rajada. Para isto, foi utilizado o fator de expansão normalizado, definido na seção 3.4.1. Pode-se observar que, para comprimentos de rajadas menores que o tamanho do *buffer*, a capacidade equivalente da fonte é proxima da taxa média ($R_n \approx 1/b$). Também observa-se que, conforme o tamanho da rajada se aproxima do tamanho do *buffer*, a capacidade equivalente tende rapidamente para a taxa de pico ($R_n \rightarrow 1$). Nestes casos, a aproximação somente produz bons resultados quando há uma única fonte. Quando muitas fontes são multiplexadas, esta aproximação superestima bastante o valor da capacidade equivalente, pois ela não considera o efeito de multiplexação estatística, que é significativo principalmente para maiores índices de rajada. Esta mesma característica pode ser observada na figura 4.7, onde o comprimento médio da rajada é mantido fixo e o tamanho do *buffer* é variado. Pode-se observar que, quando o tamanho do *buffer* é menor ou da mesma ordem que o tamanho médio da rajada, a capacidade equivalente se aproxima da taxa de pico e diminui rapidamente para valores maiores que o tamanho médio da rajada.

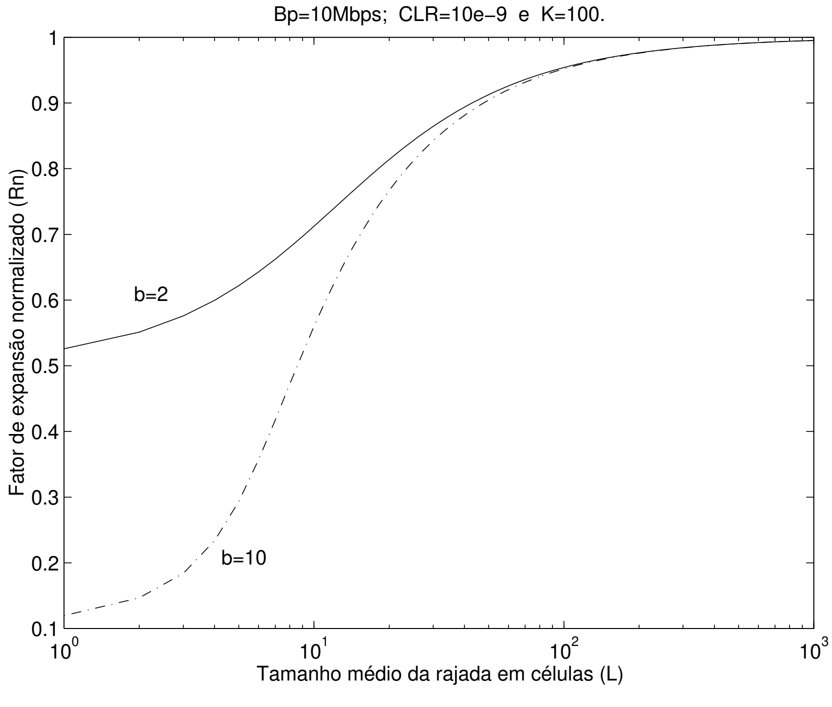

Figura 4.6: Variação da capacidade alocada com o tamanho médio da rajada na aproximação de fluxo contínuo.

A figura 4.8 mostra a variação da capacidade equivalente com o fator de utilização da fonte, para duas diferentes taxas de perda de células (CLR). O tamanho do *buffer* foi escolhido bem maior que o tamanho médio da rajada de modo que o efeito da multiple-

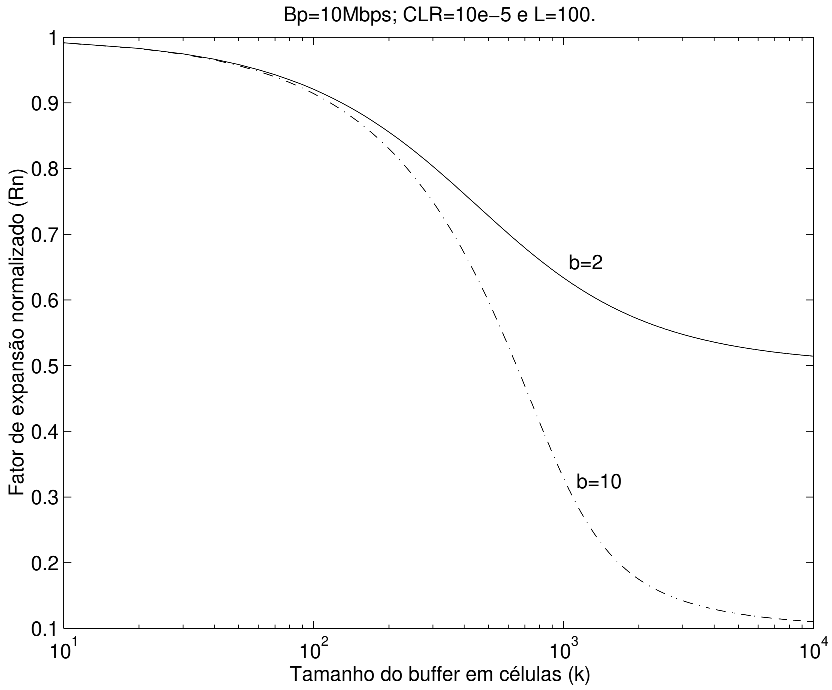

Figura 4.7: Variação da capacidade alocada com o tamanho do *buffer* na aproximação de fluxo contínuo.

xação estatística fosse mais significativo para fontes com níveis mais baixos de utilização (maiores índices de rajada), como visto nas figuras anteriores. Pode-se observar nesta figura que a capacidade equivalente varia bem menos com o fator de utilização para CLR mais baixas, estando bem mais próxima da taxa de pico. Quando o número de fontes multiplexadas é grande, isto pode resultar numa superestimação bem grande.

Aproximação Estacionária

Nos casos onde o efeito da multiplexação estatística é significativa, as capacidades equivalentes calculadas na aproximação de fluxo contínuo são superestimadas, com valores muito superiores às suas taxas médias. Este é essencialmente o caso de fontes com longos períodos de rajada e de baixa utilização. Entretanto, na maioria destes casos a distribuição estacionária da taxa binária do tráfego agregado pode ser aproximada por uma distribuição gaussiana [29].

Assim, a segunda aproximação utilizada é idêntica à apresentada no método da aproximação gaussiana. Portanto, a capacidade equivalente do agregado de fontes C_g é dada

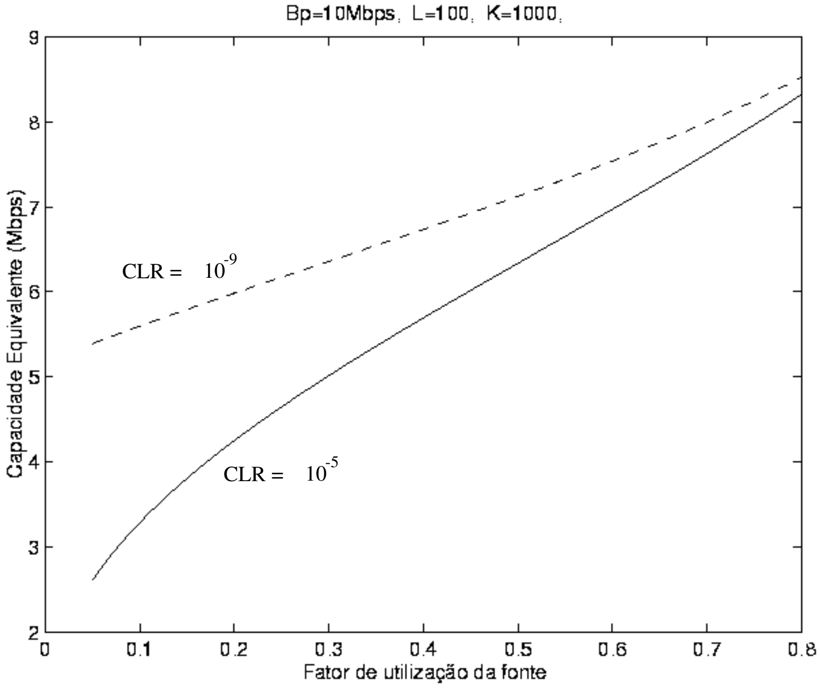

Figura 4.8: Variação da capacidade alocada com o fator de utilização na aproximação de fluxo contínuo.

por:

$$C_g = m + \alpha\sigma \ , \tag{4.8}$$

onde

$$\begin{aligned} \alpha &= \sqrt{-2\log(\varepsilon) - 2\log(2\pi)} \ ; \\ m &= \textstyle\sum_{i=1}^{N} \rho_i R_i \ ; \\ \sigma &= \sqrt{\textstyle\sum_{i=1}^{N} \rho_i R_i (R_i - \rho_i R_i)} \ . \end{aligned}$$

Como o tamanho do *buffer* não é considerado nesta aproximação, o valor da capacidade equivalente do agregado de fontes C_g é superestimado, visto que é permitido ao tráfego transmitir acima da capacidade do canal por um período curto de tempo, onde as células extras seriam armazenadas no *buffer*.

Capacidade Equivalente

Como ambas as aproximações utilizadas tendem a superestimar o valor real da capacidade necessária e são inacuradas para diferentes faixas de características das conexões, o valor da capacidade equivalente para o agregado das conexões será o mínimo entre os dois valores calculados, ou seja,

$$C_{\mathrm{eq}} = \min\left\{\widehat{C_f}\ , C_g\right\}\ . \tag{4.9}$$

Uma comparação entre os resultados de ambas as aproximações utilizadas neste método com os resultados obtidos com o modelo UAS (seção 2.3.1) é apresentada na figura 4.9 e 4.10 para taxas de perda de células iguais a 10^{-9} e 10^{-5}, respectivamente. Nelas, o comprimento médio da rajada é variado de um valor bem pequeno até um valor igual ao tamanho do *buffer*. Os valores do tamanho do *buffer* e do número de fontes foram escolhidos de modo que a variação da aproximação de fluxo contínuo pudesse ser notória. A aproximação gaussiana, como já visto, não varia com este parâmetro. Pode-se observar nestas figuras que, para comprimentos de rajadas pequenos, a aproximação de fluxo contínuo é muito boa. Isto se deve ao fato de que o impacto de uma única fonte é grande e que a capacidade equivalente de uma única fonte é próxima a sua taxa média. Conseqüentemente, quase não há ganho de multiplexação estatística. Entretanto, quando o comprimento médio da rajada aumenta, o impacto de uma única fonte deixa de ser significativo e o ganho de multiplexação estatística passa a ser maior. Com isto, a aproximação de fluxo contínuo tende a superestimar bastante o valor da capacidade equivalente. Nestes casos, a aproximação gaussiana passa a ser utilizada. Porém, como pode ser observado, ela também tende a superestimar o valor da capacidade necessária. E esta superestimação é tão maior quanto menor for a taxa de perda de células desejada.

Extensões

A noção de capacidade efetiva utilizada neste método é muito útil para fins de engenharia, pois a cada conexão é atribuído um valor de capacidade efetiva e um novo pedido de conexão só será aceito caso a soma das capacidades efetivas seja menor que a capacidade do canal de transmissão.

Os cálculos utilizados tanto na aproximação de fluxo contínuo quanto na aproximação gaussiana são bastante simples e passíveis de serem feitos em tempo de execução. Entretanto, como se estima que uma grande parte do tráfego em redes ATM seja gerado por aplicações com níveis altos de rajada, onde o ganho de multiplexação estatística é significativo, o método da capacidade equivalente acaba por utilizar a aproximação gaussiana, que produz resultados pouco acurados e bastante superestimados na maioria dos casos.

Em razão disto, diversas extensões foram propostas para um cálculo mais acurado da capacidade efetiva. Uma das mais importantantes é a presentada por Choudhury et

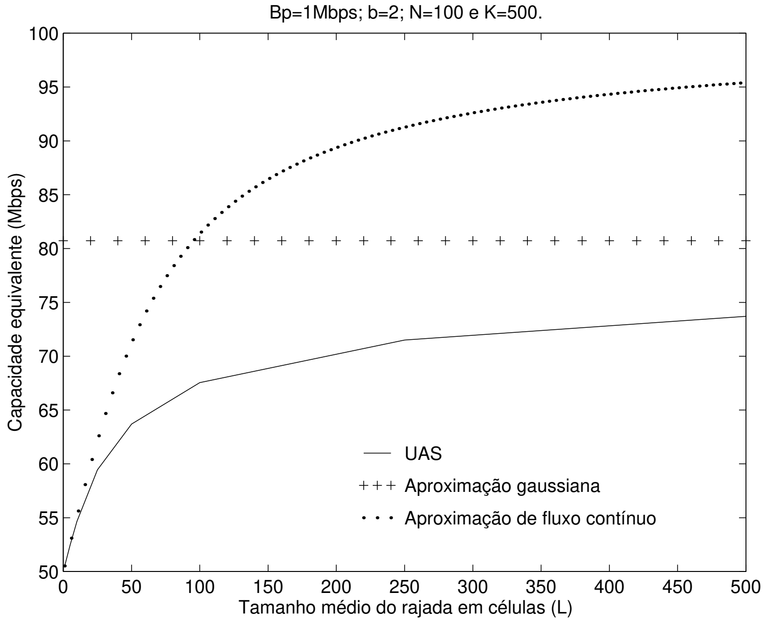

Figura 4.9: Comparação entre as aproximações de fluxo contínuo e gaussiana para um objetivo de taxa de perda de células de 10^{-9}.

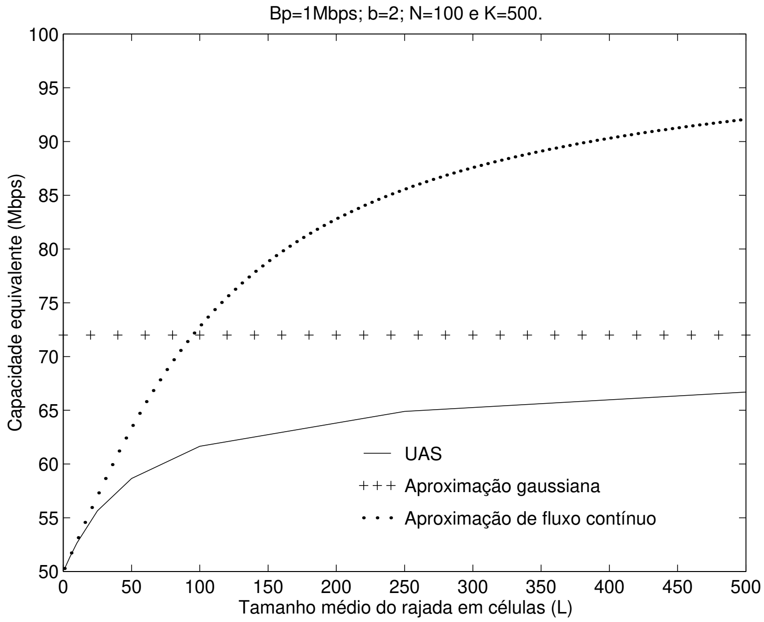

Figura 4.10: Comparação entre as aproximações de fluxo contínuo e gaussiana para um objetivo de taxa de perda de células de 10^{-5}.

al. [30]. Nesta extensão, a aproximação de fluxo contínuo é utilizada sem que o ganho de multiplexação estatística seja desconsiderado, ou seja, o valor de β na equação 4.6 não é aproximado por 1. Portanto, a probabilidade de transbordo do *buffer* $G(x)$ é da forma

$$G(x) \approx \beta\ e^{-\mu x}\ ; \quad x \to \infty \tag{4.10}$$

onde β e μ são constantes positivas chamadas de constante assintótica e taxa de decaimento assintótico, respectivamente.

Através de resultados numéricos, Choudhury et al. mostram que variando o número total de fontes e mantendo a taxa total fixa através do reescalamento das fontes, o valor de β varia exponencialmente em função do número total de fontes n, enquanto que a taxa de decaimento assintótico se mantem constante, ou seja,

$$\beta \approx \alpha\ e^{-n\phi}\ ; \tag{4.11}$$

onde $\alpha > 0$, $\phi > 0$ para fontes com um nível maior de rajadas do que a distribuição de Poisson (isto é, um coeficiente de variação maior que 1) e $\phi < 0$ para fontes com um nível menor de rajadas do que a distribuição de Poisson.

Baseado neste resultado, é proposto um método no qual os parâmetros α e ϕ são calculados de maneira exata para um número pequeno de fontes e por extrapolação logarítmica, através da equação 4.11, para um número grande de fontes.

A grande vantagem deste método em relação ao da capacidade equivalente está na sua abrangência, visto que ele apresenta resultados bastante acurados para uma gama bem maior de valores [30]. Além disto, uma outra vantagem apontada pelo autor para este método é que a qualidade de sua aproximação pode ser medida através da constante assintótica β. Caso o valor de β seja substancialmente maior que o objetivo de taxa de perda de células (CLR) desejado (como por exemplo, por mais de 4 ordens de grandeza), a aproximação é boa. Para valores não tão maiores que o objetivo de CLR, a qualidade da aproximação é mediana. Para valores inferiores ao objetivo de CLR, a aproximação é ruim.

Outras extensões para o cálculo da capacidade efetiva, baseadas na teoria dos grandes desvios, podem ser encontradas em Tse et al. [31], Simonian et al. [32], Veciana et al. [33] e Chang et al. [34]. Para uma introdução a esta teoria, ver Weiss [35].

4.3.5 Aproximações para Fontes Heterogêneas

Muitos dos resultados obtidos nesta seção são válidos apenas para o caso onde fontes homogêneas são multiplexadas. Para suprir esta deficiência, vários métodos de generalização para fontes heterogêneas foram propostos utilizando aproximações a partir dos resultados obtidos para fontes homogêneas. Dentre eles estão os métodos linear, o das misturas independentes, o da regra relacionada com a classe e o não-linear.

Método Linear

No método linear [36, 5], a capacidade equivalente a ser alocada para uma conexão da classe j, V_j, é aproximada pela capacidade total do canal, C, dividida pelo número máximo de conexões da classe j, x_j, que o canal possa suportar enquanto atende aos requisitos de QoS.

$$V_j = \frac{C}{x_j} \tag{4.12}$$

Dado que as capacidades equivalentes para cada classe são calculadas antes, uma conexão só será aceita se a capacidade residual do canal, R, for suficiente para a nova conexão.

$$R = C - \sum_j x_j Vj \tag{4.13}$$

Este método não considera a interferência entre as classes que estão sendo multiplexadas no canal, o que pode causar erros significativos. Por isto, o mesmo deve ser tratado como o limite superior para a utilização do canal ou como limite inferior para a capacidade equivalente de uma conexão.

Método das Misturas Independentes

O método das misturas independentes desconsidera o efeito da multiplexação estatística entre as classes. A capacidade equivalente de uma conexão é calculada como se a capacidade total do canal fosse subdividida em várias partes, cada qual exclusiva de uma determinada classe e alocada de acordo com o número de conexões desta classe. Por não levar em conta a multiplexação entre as classes, este método superestima a capacidade a ser alocada para uma conexão, devendo ser considerado como limite superior para a capacidade equivalente ou como limite inferior para a utilização do canal.

Uma observação deve ser feita, contudo, sobre o método das misturas independentes. O seu uso nos casos em que classes com diferentes requisitos de QoS estão sendo multiplexadas pode ser vantajoso, e resultar numa melhor utilização do canal. Isto se deve ao fato de que a aceitação das conexões de uma classe é feita de modo independente e que os diferentes requisitos de QoS podem ser tratados. Já numa estratégia de uso comum do canal pelas diferentes classes, o requisito de QoS a ser adotado deve ser o mais restringente, reduzindo, assim, o número de conexões das classes menos restringentes, pois a capacidade equivalente destas conexões seria superestimada.

Método Não-Linear

O método não-linear é atribuído a Dziong et al. [36]. Neste método, o processo de *background* é considerando no cálculo da capacidade equivalente de uma conexão, ou banda passante efetiva. Este processo é determinado pela superposição de todas as conexões, exceto aquela sob consideração, e pela capacidade do canal.

Dois conceitos são introduzidos para caracterizar o processo de *background*. São os conceitos de suavidade e de espalhamento. A suavidade representa a diferença entre a banda passante alocada e a soma das taxas de pico. O espalhamento representa a diferença entre a banda passante alocada e a taxa média total. A vantagem na utilização destes conceitos está no fato de que eles são aplicáveis num ambiente heterogêneo.

A suposição básica deste método é que na alocação da banda passante efetiva, a suavidade e o espalhamento do processo de *background* para uma conexão de uma certa classe devem ser pelo menos maior que a suavidade e o espalhamento num ambiente homogêneo para fontes desta mesma classe, resultando na mesma alocação de banda passante. Também é considerado adicionalmente que a banda passante efetiva calculada, ou capacidade equivalente, na aproximação linear é o limite inferior para a banda passante efetiva de uma classe.

Método da Regra Relacionada com a Classe (CRR)

O CRR, proposto por Gallassi et al. [37, 27], é um método baseado em tabelas, determinadas por simulação, que indicam as bandas passantes necessárias para acomodar conexões com as mesmas características (da mesma classe). O processo de chegada é modelado como um IPP (*Interrupted Poisson process*). O elemento (i, j) da tabela indica a banda passante necessária para acomodar i conexões da classe j de modo a satisfazer os requisitos de QoS.

O controle de admissão aceita uma nova conexão caso a capacidade total do canal seja maior que o mínimo entre a banda passante necessária sob o critério das misturas independentes e a banda passante máxima calculada considerando-se todas as conexões como sendo de uma única classe.

Esta regra se baseia no fato de que, numa estratégia de compartilhamento entre as classes, a banda passante alocada seria menor que na mistura independentes de classes.

Comparação entre os métodos

A figura 4.11 apresenta uma comparação entre os resultados obtidos por simulação e aqueles obtidos através dos métodos de aproximação para fontes heterogêneas. Nesta figura, são consideradas fontes de vídeo com taxas médias de 3.9 Mbps e taxas de pico de 10.6 Mbps e fontes em rajadas do tipo silêncio/atividade com taxas de pico de 10 Mbps, taxas médias de 1 Mbps e tamanho médio de rajada de 100 células. O tamanho de *buffer* é 35 a taxa de perda de células desejada é 10^{-4}. O método das misturas independentes e o método da regra relacionada com a classe (CRR), como mostrado na figura, subestimam bastante o número de conexões máximas que podem ser estabelecidas devido a superestimação dos valores das capacidades que devem ser alocadas para cada classe. Nesta figura, observar-se ainda que as curvas destes métodos quase não se distigüem. O método linear, como anteriormente previsto, superestima o número máximo de conexões que podem ser aceitas por classe. Porém, ao contrário do que poderia se imaginar, esta superestimação não é tão grande, obtendo resultados bem próximos dos simulados. O método não-linear, dentre todos aqui analisados, é o que melhor se aproxima dos resultados simulados, apresentando um erro bem pequeno. Entretanto, mais análises precisam ser feitas com um número maior de classes de características mais diversas para a validação deste modelo.

4.4 Método de Reservas Rápidas

O método de reservas rápidas foi desenvolvido para fontes de tráfego em rajadas. Neste método, quando uma conexão é estabelecida, nenhum recurso de rede é previamente alo-

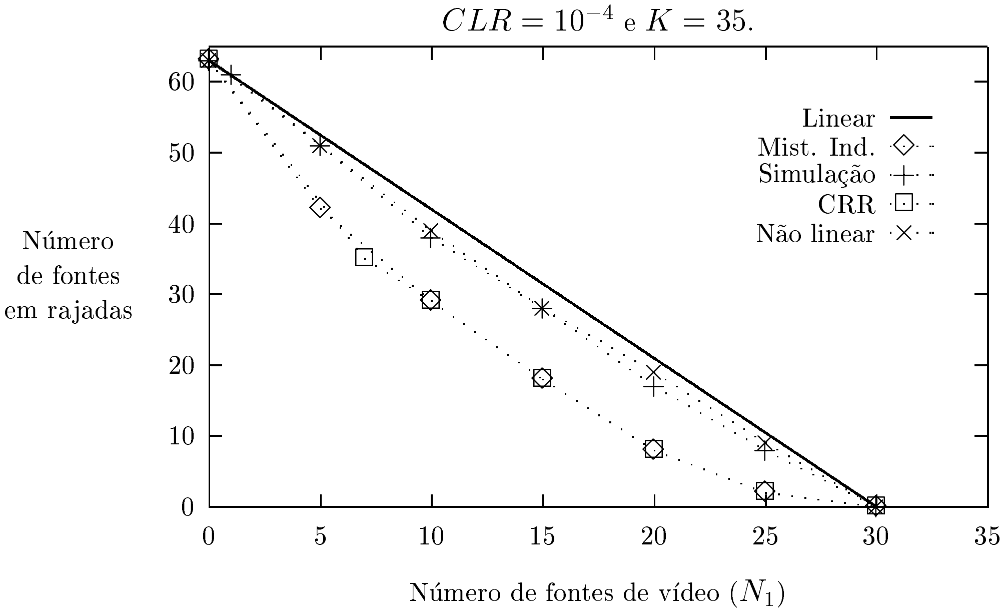

Figura 4.11: Aproximações para fontes heterogêneas.

cado, sendo apenas determinadas as rotas através da rede de cada conexão. No momento em que a fonte deseja transmitir uma rajada, a rede tenta alocar os recursos necessários apenas para a duração da rajada.

Visando a alocação destes recursos, diversos mecanismos foram propostos utilizando diferentes técnicas de reserva. Enquanto alguns se baseiam na reserva de banda passante, outros estão baseados na reserva de espaço no *buffer*. Aqui, cada um destes mecanismos é analisado.

4.4.1 Reserva Rápida de Banda Passante

O método de reserva rápida de banda passante foi proposto por Tranchier et al. para fontes em rajadas [7]. A alocação da banda passante 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 através de uma célula especial de pedido de reserva. 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. Os pedidos de decremento são feitos através de células de gerenciamento e são sempre aceitos.

Um protocolo de reservas rápidas (FRP - *Fast Reservation Protocol*) foi implementado para tratar as células especiais de gerenciamento. O protocolo utiliza diferentes tipos de temporizadores de modo a assegurar uma operação confiável. O terminal utiliza um temporizador para assegurar que as suas células, tal como um pedido de incremento de banda passante, enviadas para a sua unidade FRP local não sejam perdidas. Quando uma unidade FRP recebe um pedido de incremento, ela envia o pedido para o primeiro nó do caminho, que por sua vez envia para o próximo nó e assim sucessivamente. Caso o pedido seja aceito por todos os nós, o último nó envia um sinal de confirmação para a unidade FRP. A unidade FRP então informa ao terminal que o seu pedido foi aceito e envia uma célula de validação para todos os nós confirmando a reserva. Caso o pedido seja rejeitado por um nó, o nó simplesmente descarta o pedido. Os nós que já haviam reservado a banda passante, conseqüentemente, também descartarão o pedido pela ausência de uma célula de validação. No caso de uma rejeição de pedido, a unidade FRP voltará a fazer um novo pedido após um certo período determinado por um temporizador, sendo também limitado número de tentativas.

O método de reserva rápida de banda passante foi proposto, segundo o autor, para fontes em rajadas cujas taxas binárias de pico sejam bem inferiores a capacidade do canal, como, por exemplo, 2% [7].

4.4.2 Reserva Rápida de *Buffer*

O método de reserva rápida de *buffer* é atribuído a Turner [38]. Neste método, o espaço em *buffer* é que é alocado ao invés da banda passante, como no método anterior. Com isto, fontes cujas taxas de pico são próximas a capacidade do canal podem ser aceitas. Cada fonte é modelada por uma cadeia de Markov de 2 estados (fontes em rajada do tipo silêncio/atividade), alternando entre os estados ativos e inativos. Quando a fonte está no seu estado ativo, um número predeterminado de compartimentos é alocado no *buffer* por toda a duração do período ativo, sendo liberados logo após o final deste período. As transições entre estes estados se dão através de células especiais que determinam o início e o fim de uma rajada. Todas as outras células são tratadas como células intermediárias. Células simples podem ser transmitidas com prioridades mais baixas, sendo marcadas através do *bit* de prioridade de perda de células (CLP - *Cell Loss Priority*).

Sejam k' o número de compartimentos disponíveis no *buffer*, b_i o número de compartimentos reservados para a conexão i e u_i o número de compartimentos atualmente em uso pela conexão i. Então, o seguinte procedimento descreve o mecanismo de reserva rápida de *buffer*.

Quando uma fonte que se encontra no estado inativo deseja transmitir, ela envia a célula de início de rajada e as células subseqüentes. Caso a célula de início não seja enviada, todas as outras células são descartadas. Ao receber a célula de início, o nó verifica se é possível alocar o número predeterminado de espaço em *buffer* b_i. Caso b_i seja maior que k', então toda a rajada é descartada. Caso contrário, o nó então aceita a rajada, atualiza k' ($k' = k' - b_i$), muda o estado da conexão para ativo e dispara um temporizador. Se u_i for menor que b_i, então u_i é incrementado. Se u_i for igual a b_i, então a célula é marcada como de baixa prioridade. Sempre que uma célula é transmitida, u_i é decrementado. Caso o temporizador estoure antes da chegada de uma nova célula intermediária ou de fim de rajada, a conexão é passada para o estado inativo.

Como pode ser observado neste método, através da marcação de células, é permitida a transmissão de um número maior de células do que o previamente determinado. Com isto, aplicações tolerantes a perda podem ser facilmente tratadas. Além disto, este mecanismo também é próprio para aplicações não tolerantes a perda, pois normalmente estas aplicações forçam a retransmissão de parte ou de toda a rajada devido a perda de uma única célula. Neste caso, caso não haja espaço para a rajada no *buffer*, toda ela é descartada, evitando assim a retransmissão da mesma informação diversas vezes pela rede.

A decisão de aceitação ou não de uma nova conexão é feita com base na probabilidade de excesso de demanda de *buffer*, isto é, na probabilidade de se requerer um número maior de compartimentos do que o tamanho do *buffer*. Sendo R_i a taxa de pico da conexão i, a_i a taxa média da conexão i, C a capacidade do canal e K o tamanho do *buffer*, então $b_i = \left\lceil K \frac{R_i}{C} \right\rceil$. Sendo x_i a variável aleatória que denota a demanda de *buffer* da conexão i, então

$$p_i = \Pr(x_i = b_i) = \frac{a_i}{R_i} \quad \text{e} \quad \Pr(x_i = 0) = 1 - p_i \ .$$

Assim, a demanda de *buffer* X do agregado de conexões é dada pela soma das demandas de cada conexão, ou seja, $X = \sum_{i=1}^{N} x_i$. Assumindo que as demandas individuais são mutuamente independentes, a função de geração de probabilidade de X, $f_x(z)$, é

$$f_x(z) = \prod_{i=1}^{N} \left\{ (1 - p_i) + p_i \, z^{b_i} \right\} = C_0 + C_1 z + \cdots + C_L z^L \ , \tag{4.14}$$

onde $L = \sum_{i=1}^{N} b_i$ e $C_j = \Pr(X = j)$ é a probabilide da demanda de *buffer* X do agregado de conexões ser j. E, finalmente, a probabilidade de excesso de demanda de *buffer* é dada por

$$\Pr(X > K) = 1 - \sum_{i=1}^{K} C_j \ . \tag{4.15}$$

A figura 4.12 faz uma análise da variação da taxa de perda de células com o número de fontes, para diversos valores de índice de rajada e para uma taxa média de 2 Mbps. A capacidade do canal é de 155 Mbps e o tamanho do *buffer* é 365 células, o que corresponde a 100 μs de transmissão a taxa do canal. Como a taxa média é a mesma para cada fonte, a taxa média do tráfego agregado também a será para um mesmo número de fontes, independentemente do índice de rajada do tráfego. Portanto, como era esperado, as taxas de perda de células (CLR) serão maiores para os tráfegos com maiores índices de rajada. Além disto, pode-se observar que a variação da CLR com com o número de fontes é bem maior para índices de rajada menores.

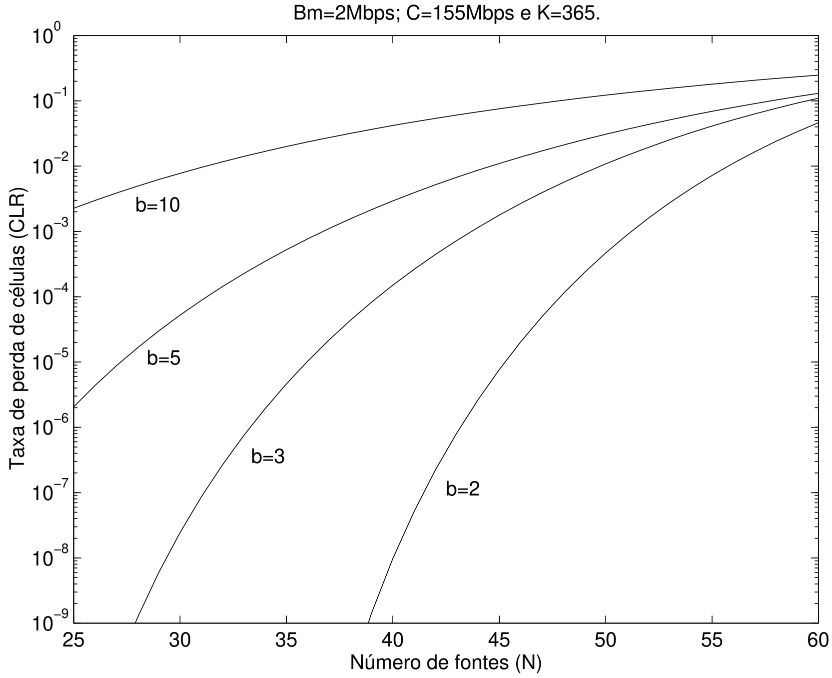

Figura 4.12: Variação da taxa de perda de células com o número de fontes (taxa média fixa).

A figura 4.13 faz uma análise semelhante a da figura anterior, fixando, porém, a taxa binária de pico de cada fonte em 10 Mbps. Com isto, a taxa média do tráfego agregado, para um mesmo número de fontes, é diferente para os índices de rajada utilizados. Os valores do tamanho do *buffer* e da capacidade do canal são os mesmos do exemplo anterior. Como esperado, tráfegos com menores índices de rajada e, portanto, maior taxa média agregada, possuem uma taxa de perda de células mais elevada.

Durante a análise deste método, observou-se que seus resultados se tornam mais inacurados para tamanhos de *buffer* menores. Isto se deve ao fato de que, nestes casos, a quantidade de compartimentos que são reservados para cada conexão passa a ser muito

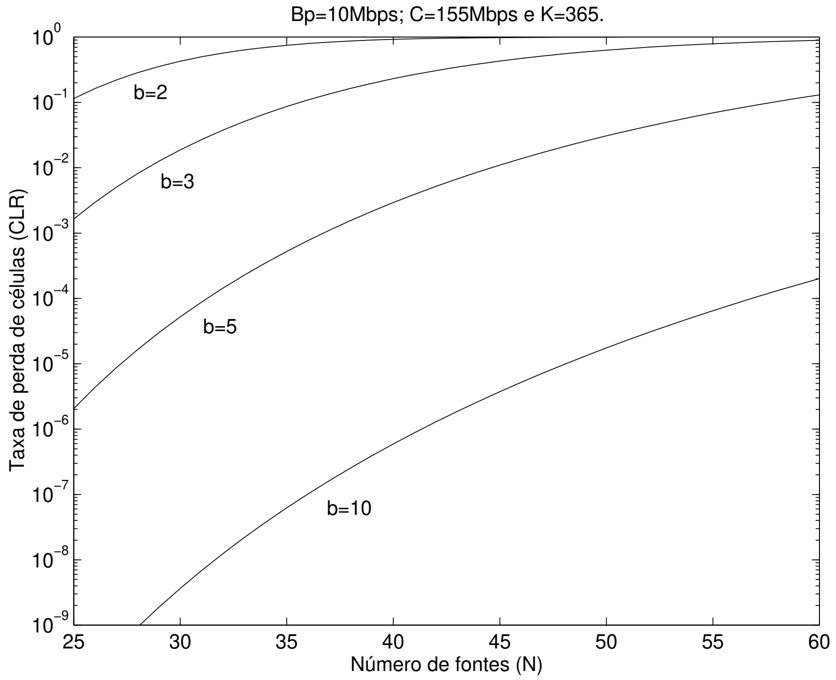

Figura 4.13: Variação da taxa de perda de células com o número de fontes (taxa de pico fixa).

pequena. Como a estimação da taxa de perda de células é feita com base neste valor, conseqüentemente, a sua acurácia piora. Além disto, como este método foi desenvolvido para fontes com períodos de rajada de longa duração, seus resultados são bastante superestimados quando fontes com períodos curtos de rajada são multiplexadas. Isto pode ser melhor observado na figura 4.14, onde os resultados deste método são comparados com os do modelo UAS para fontes com índice de rajada igual a 5, para diversos valores de comprimento médio de rajada (L). Resultados equivalentes podem ser obtidos para outros valores de índice de rajada.

4.5 Aproximação de Tráfego Intenso

A aproximação de tráfego intenso é atribuída a Sohraby [39] e é está baseada no comportamento assintótico da cauda da distribuição do tamanho de fila quando a intensidade de tráfego é alta. Nesta aproximação, o multiplexador ATM é modelado por um sistema de filas de capacidade infinita, serviço constante e com um processo de chegada modulado por uma cadeia de Markov, com uma matriz de geração de probabilidade $A(z)$.

A resolução deste sistema de filas pode ser feita através de métodos de geometria das

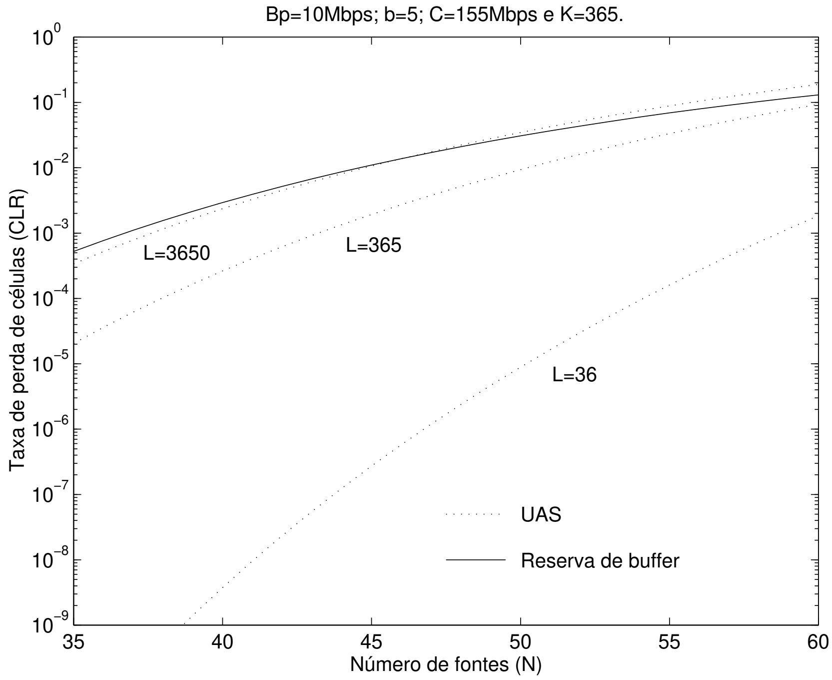

Figura 4.14: Comparação entre o método da reserva de *buffer* e o modelo UAS para diversos comprimentos médios de rajada.

matrizes, mas a computação exata exige muitos cálculos, especialmente para um grande número de fontes. Além disto, quando as cadeias de Markov que regem os processos de chegadas das fontes são finitas, aperiódicas e irredutíveis, a cauda da distribuição estacionária do tamanho da fila exibe um comportamento geométrico, que é caracterizado pela menor raiz fora do círculo unitário do determinante $|zI - A(z)|$. Denotando esta raiz por z^*, então para um valor de i suficientemente grande, tem-se

$$\Pr(q > i) \approx \alpha(z^*)^{-i}, \quad z^* > 1 \tag{4.16}$$

onde q denota a variável aleatória correspondente a distribuição estacionária do tamanho da fila.

Denotando por $\lambda_i(z)$, $1 \leq i \leq N$, os autovalores de $A(z)$, então o determinante pode ser fatorado por

$$\Delta(z) = |zI - A(z)| = \prod_{i=1}^{N} (z - \lambda_i(z)) \ . \tag{4.17}$$

Portanto, uma vez determinados os autovalores de $A(z)$, os zeros do determinante $\Delta(z)$ são facilmente obtidos. Entretanto, resta ainda determinar qual destas raízes é z^*. Sohraby mostra que a raiz z^* resolve a equação $z = \lambda(z)$, onde $\lambda(z)$ é o autovalor de Perron-Frobenious (PF) de $A(z)$.

Para um processo de chegada que é a superposição de processos de chegada independentes, os autovalores do processo de superposição são facilmente obtidos em termos daqueles dos processos individuais. E, o que é mais importante, o autovalor de PF do processo superposto é simplesmente o produto dos autovalores de PF das fontes individuais.

A resolução de tal sistema, no entanto, tem que ser feita numericamente. Assim, para uma maior eficiência no cálculo de z^* , Sohraby propôs a seguinte aproximação, válida nos casos onde a intensidade do tráfego é grande:

$$z^* \approx 1 + \frac{2(1-\rho)}{\sum_{i=1}^{N} \rho_i \left(1 - R_i \rho_i\right)^2 \left(C_{on}^2(i) + C_{off}^2(i)\right) T_i} \tag{4.18}$$

onde T_i é o tempo médio do período ativo, ρ_i é a utilização da fonte, R_i é a razão entre a taxa do canal e a taxa de pico da fonte e $C_{on}^2(i)$ e $C_{off}^2(i)$ são os coeficientes quadráticos de variação dos períodos ativo e inativo, respectivamente. Note que $R_i \rho_i$ deve ser menor que 1. Então, a cauda da distribuição do tamanho da fila pode ser aproximada por $\Pr(q > i) \approx \rho\,(z^*)^{-i}$, onde ρ é a utilização total. Logo, para um dado tamanho de *buffer* K, a probabilidade de perda de células é aproximada por

$$\Pr(q > K) \approx \rho\,(z^*)^{-K} \ . \tag{4.19}$$

Portanto, um pedido de conexão só é aceito se a taxa de perda de células, aproximada pela equação 4.19, for menor que o valor desejado ε, ou equivalentemente,

$$\ln\left[\rho\,(z^*)^{-K}\right] < \ln(\varepsilon) \ . \tag{4.20}$$

Segundo o autor, a aproximação utilizada neste mecanismo é muito boa quando o nível de utilização do canal é alto ($\rho \approx 1$). Entretanto, como se estima que grande parte do tráfego em redes ATM seja gerado por fontes com períodos de rajada bem variados e com baixos níveis de utilização [30], este mecanismo pode ser pouco adequado.

4.6 Aproximação Não-paramétrica

O objetivo do Controle de Admissão de Conexão é o de alocar a banda passante necessária para uma conexão de modo a garantir a Qualidade de Serviço desejada. E isto deve ser feito para qualquer conjunto de parâmetros de tráfego especificados pelo usuário e para qualquer processo de chegada de células que satisfaça os valores destes parâmetros. O método de avaliação de desempenho que não assume um modelo para o processo de chegada de células é chamado de aproximação não-paramétrica [2].

A idéia básica deste método está em calcular o valor máximo que a taxa de perda de células pode assumir para um dado conjunto de parâmetros de tráfego e utilizá-lo como critério de aceitação para um pedido de uma nova conexão. Para tanto, torna-se necessária a obtenção do tráfego do pior caso, que é aquele que produz a maior taxa de perdas de células. Para cada conjunto de parâmetros de tráfego haverá um tráfego de pior caso.

Em [40] são apresentados resultados para dois conjuntos de parâmetros de tráfego: taxa binária média (ABR - *Average Bit Rate*) e taxa binária de pico (PBR - *Peak Bit Rate*); e taxa binária média (ABR) e variância da taxa binária. Por se tratarem de casos equivalentes, aqui somente o primeiro é abordado.

Sejam ANA e MNA, que representam respectivamente o número médio e o número máximo de células que chegam durante um intervalo de tempo fixo, os parâmetros de tráfego utilizados. O comprimento deste intervalo é escolhido como o tempo necessário para a transmissão de $K/2$ células pelo canal, onde K representa o tamanho do *buffer* em células.

Dado que em um canal de transmissão células de N conexões estão sendo multiplexadas e que a probabilidade de chegada de j células da conexão i seja $p_i^*(j)$, $i = 1, 2, \ldots, N$, $j = 0, 1, \ldots$, então o limite superior da probabilidade de perdas de células (CLP - *Cell Loss Probability*) é:

$$\begin{aligned} CLP &\leq B^*(p_1^*, \cdots, p_N^*; K/2) \\ &\triangleq \frac{\displaystyle\sum_{k=0}^{\infty} [k - K/2]^+ p_1^* \star \cdots \star p_N^*(k)}{\displaystyle\sum_{k=0}^{\infty} k \cdot p_1^* \star \cdots \star p_N^*(k)} \end{aligned} \tag{4.21}$$

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

Seja $\theta_i = \{\theta_i(j)\}$, descrita na equação 4.22, a distribuição do tipo Bernoulli obtida através dos parâmetros de tráfego (a_i, R_i) da conexão i, onde a_i é o ANA e R_i é o MNA.

$$\theta_i(j) = \begin{cases} a_i/R_i & ,j = R_i \ ; \\ 1 - a_i/Ri & ,j = 0 \ ; \\ 0 & \text{, caso contrário.} \end{cases} \tag{4.22}$$

Saito [40, 2] mostra que esta distribuição representa o tráfego do pior caso para a conexão i. E, aplicando-se este resultado recursivamente para todas as conexões, pode-se obter, então, o limite superior para a probabilidade de perda de células (CLP).

$$\begin{aligned} CLP & \leq B^*(p_1^*, \cdots, p_N^*; K/2) \\ & \leq B^*(p_1^*, \cdots, p_{N-1}^*, \theta_N; K/2) \\ & \leq B^*(p_1^*, \cdots, \theta_{N-1}, \theta_N; K/2) \\ & \quad \vdots \\ & \leq B^*(\theta_1, \cdots, \theta_N; K/2) \\ & = \frac{\sum_{k=0}^{\infty} [k - K/2]^+ \theta_1^* \star \cdots \star \theta_N^*(k)}{\sum_{k=0}^{\infty} k \cdot \theta_1^* \star \cdots \star \theta_N^*(k)} \end{aligned} \tag{4.23}$$

Este limite superior representa o caso em que cada conexão i, que se encontra no seu período ativo, esteja transmitindo R_i células em uma única rajada a cada intervalo fixo de tempo, e permanecendo sem transmitir até o final deste intervalo. Logo, como mostrado neste caso, existe um processo que alcança o limite superior da CLP e, deste modo, este valor não pode ser mais melhorado.

O controle de admissão que se baseia neste limite superior da CLP é bastante simples. Uma nova conexão $N+1$, que faz um pedido de admissão, será aceito se

$$B^*(\theta_1, \cdots, \theta_{N+1}; K/2) \leq \tilde{B} \tag{4.24}$$

e rejeitado em caso contrário, onde N é o número de conexões já conectadas e $\tilde{B}$ é a probabilidade padrão de perdas de células.

A figura 4.15 faz uma análise da variação da taxa de perda de células (CLR) com o número de fontes, para diversos valores de índice de rajada e para uma taxa média de 2 Mbps. A capacidade do canal é de 155 Mbps e o tamanho do *buffer* é 365 células, o que corresponde a 100 μs de transmissão a taxa do canal. Como a taxa média do tráfego agregado é a mesma para diferentes índices de rajada, a perda se mostra bem maior para as fontes com índices de rajada mais altos, já que nos seus períodos de atividade, a taxa com que elas transmitem é bem mais alta. Além disto, a variação da CLR é bem maior para índices de rajada menores.

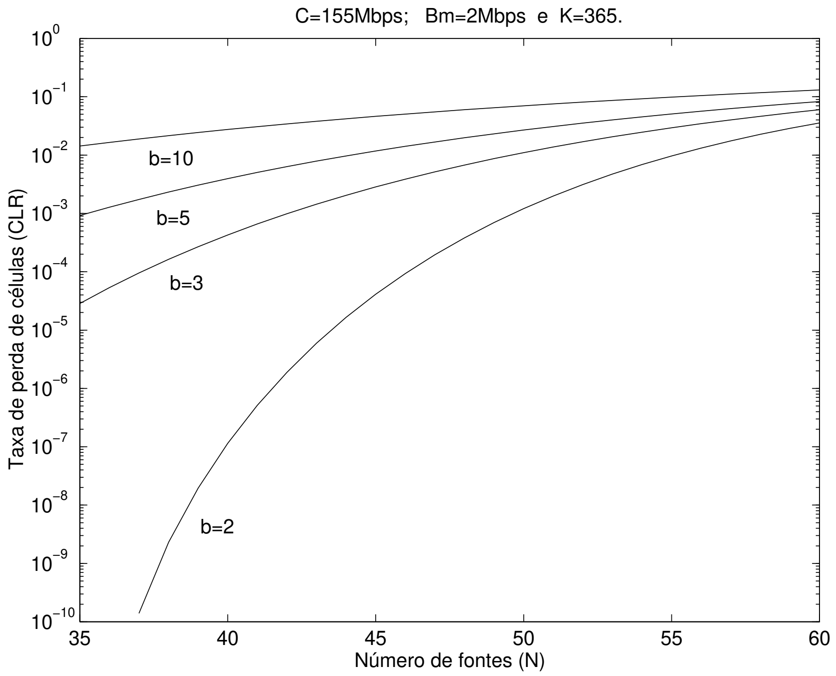

Figura 4.15: Variação da taxa de perda de células com o número de fontes (taxa média fixa).

A figura 4.16 faz uma análise semelhante a da figura anterior, fixando, porém, a taxa binária de pico de cada fonte em 10 Mbps. Como a taxa média do tráfego agregado é menor para fontes com maiores índices de rajada, estas apresentam um taxa de perda de células (CLR) bem menor. Entretanto, pode-se notar que a variação da CLR com o número de fontes não muda drasticamente com o valor do índice de rajada. Isto pode ser explicado observando-se a distribuição utilizada para o cálculo da CLR (equação 4.22). Nesta distribuição, para um mesmo valor de taxa de pico, o número máximo de células geradas por fontes é o mesmo, mudando apenas a probabilidade de sua ocorrência, o que implica apenas em valores diferentes de CLR, mas não em variações muito diferentes.

Durante a análise deste método, observou-se que o mesmo apresenta resultados pouco acurados para tamanhos pequenos de *buffer*. Isto ocorre devido a uma perda de acurácia na computação dos números médio e máximo de células que podem ser transmitidos durante um período de definição. Como este período toma como base o tamanho do *buffer* para o seu cálculo, para tamanhos pequenos de *buffer*, a faixa de variação destes parâmetros é mais estreita, perdendo a sua acurácia. Além disto, como nenhum modelo específico de tráfego é assumido, também se observou que seus resultados tendem a ser

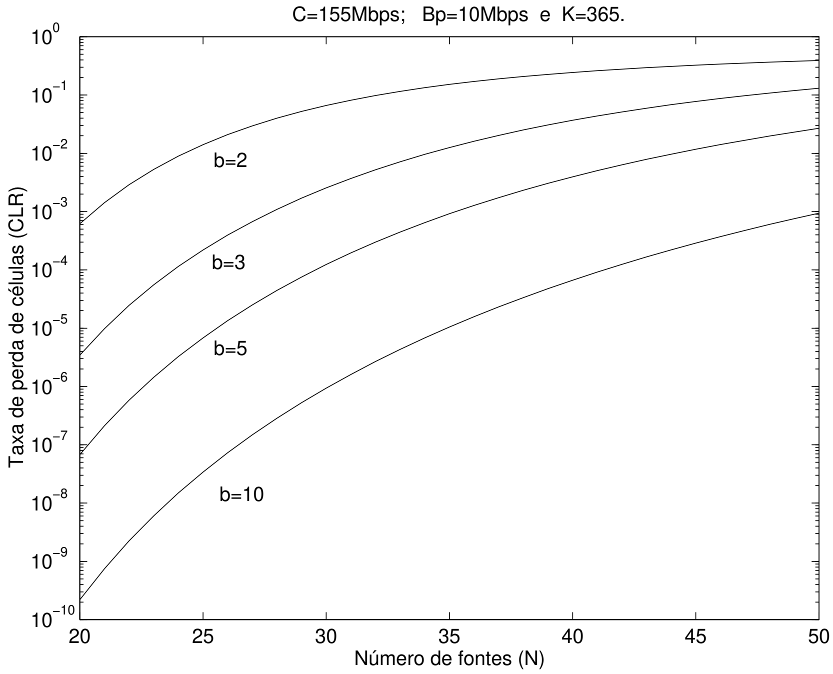

Figura 4.16: Variação da taxa de perda de células com o número de fontes (taxa de pico fixa).

bastante superestimados, como pode ser visto na figura 4.17. Nesta figura, os resultados deste método são comparados com os do modelo UAS para fontes com índice de rajada igual a 5, para diversos valores de comprimento médio de rajada (L). Esta comparação mostra que, para tráfegos com comprimentos médios de rajada menores, a aproximação superestima bastante a taxa de perda de células. Entretanto, para comprimentos maiores de rajada, esta aproximação passa a apresentar bons resultados.

4.7 Métodos Dinâmicos

A base do controle de admissão de conexão em redes ATM está no fato de que o usuário será capaz de especificar os parâmetros de seu tráfego. Porém, nem sempre o usuário tem o conhecimento exato, ou mesmo aproximado, de todos os seus parâmetros. E ao tentar estimá-los, dois problemas podem surgir: a subestimação ou superestimação dos valores dos parâmetros. O primeiro caso implica numa subestimação da taxa de perda de células (CLR), podendo eventualmente não atender aos requisitos de QoS desejados. O último caso implica na superestimação da CLR e num conseqüente desperdício na utilização da rede, apesar de garantir os requisitos de QoS.

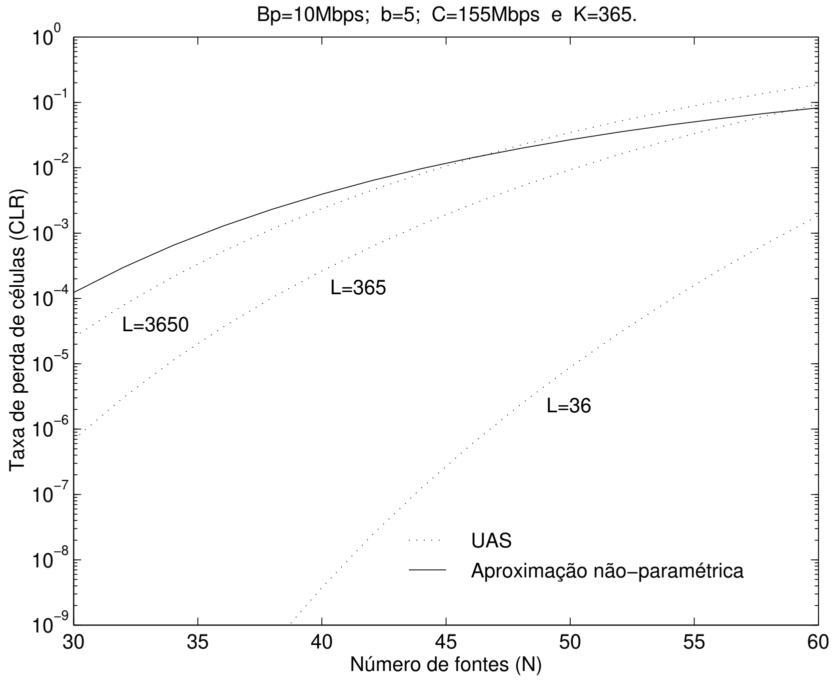

Figura 4.17: Comparação entre a aproximação não-paramétrica e o modelo UAS para diversos comprimentos médios de rajada.

Com o intuito de alcançar um boa utilização da rede e simplificar a interface usuário-rede (UNI - *User Network Interface*), Saito [41, 2] propôs o CAC dinâmico. Este método é baseado na distribuição do número de células que chegam durante um determinado período de estimação ou de renovação. Cada período de estimação é composto por vários períodos de medição.

Dado que tenham N conexões sendo multiplexadas, e que a probabilidade de chegarem j células da conexão i durante um intervalo fixo de tempo S seja $p_i^*(j)$, então, como já mencionado anteriormente (seção 4.6), o limite superior da CLR é dada por

$$P_{\mathrm{CLR}} \leq \frac{\displaystyle\sum_{k=0}^{\infty}[k-K/2]^{+}p_1^* \star \cdots \star p_N^*(k)}{\displaystyle\sum_{k=0}^{\infty} k \cdot p_1^* \star \cdots \star p_N^*(k)} \tag{4.25}$$

onde S corresponde ao período de transmissão de $K/2$ células a taxa do canal.

Quando uma nova conexão com MNA $= R_{N+1}$ e ANA $= a_{N+1}$ é aceita no canal, então o novo limite superior da CLR é dado por

$$P_{\mathrm{CLR}} \leq \frac{\sum_{k=0}^{\infty}[k-K/2]^{+}p_1^{*}\star\cdots\star p_N^{*}\star\theta_{N+1}(k)}{\left(\sum_{k=0}^{\infty}k\cdot p_1^{*}\star\cdots\star p_N^{*}(k)+a_{N+1}\right)} \tag{4.26}$$

O princípio básico deste CAC dinâmico está na estimativa da distribuição do número de células que chegam dos usuários já conectados através da medição do fluxo de células. Com base nesta distribuição estimada $\hat{p}=[\hat{p}(0),\hat{p}(1),\ldots]$ e nos parâmetros da nova conexão, o limite superior para o CLR, U_B', pode ser estimado por

$$U_B' = \frac{\sum_{k=0}^{\infty}[k-K/2]^{+}\hat{p}\star\theta_{N+1}(k)}{\left(\sum_{k=0}^{\infty}k\hat{p}(k)+a_{N+1}\right)} \tag{4.27}$$

Assim, um pedido de conexão será aceito caso o limite superior estimado para o CLR atenda aos requisitos de QoS desejados.

Seja um período de estimação ou período de renovação um conjunto de M períodos de medição de comprimento fixo. Assim, a distribuição de frequências medidas neste t-ésimo período de renovação será $q(k;t)$, $k=0,1,2,\ldots$, com média $b(t)=\sum_{k=0}^{\infty}k\cdot q(k;t)$.

Caso o número de conexões existentes não se altere durante um período de renovação, então $\hat{p}$ e $\hat{a}$ podem ser estimados por

$$\hat{p}(t+1) = \alpha q(t)+(1-\alpha)\hat{p}(t); \tag{4.28}$$

$$\hat{a}(t+1) = \alpha b(t)+(1-\alpha)\hat{a}(t); \tag{4.29}$$

onde $q(t)=\{q(k;t)\}$ e α representa o quão rápido o valor estimado vai se atualizando com os atuais valores medidos.

Quando um novo pedido de conexão é adicionado durante um período de renovação, o procedimento de atualização da distribuição é dado por

$$\hat{p}(k;t+1)=\hat{p}(\cdot;t)\star\theta_{N+1}(k); \qquad k=0,1,2,\ldots \tag{4.30}$$

$$\hat{a}(t+1)=\hat{a}(t)+a_{N+1}; \tag{4.31}$$

Portanto, a nova conexão é adicionada considerando-se para o processo de chegada desta conexão o tráfego do pior caso θ_{N+1}, garantindo assim os requisitos de QoS. Porém, conforme novas estimativas vão sendo feitas, o efeito de θ_{N+1} vai se reduzindo por um fator $1-\alpha$, enquanto que o efeito das medições aumenta por um fator α.

Quando ocorre o término de uma conexão durante um período de renovação, o procedimento de atualização da distribuição é dado por

$$\hat{p}(k;t+1) = \hat{p}(k+a;t) \ ; \quad (4.32)$$
$$\hat{a}(t+1) = \hat{a}(t) - a' \ ; \quad (4.33)$$

onde a' representa a média real medida da conexão encerrada.

A razão mostrada por Saito [2] é que o maior limite superior de CLR para as conexões restantes ocorre quando a conexão encerrada tem seu valor mínimo de CLR. A distribuição que resulta no menor limite superior para o CLR da conexão sainte é

$$p(k) = \begin{cases} 1 & , k = a' \ ; \\ 0 & , \text{caso contrário.} \end{cases} \quad (4.34)$$

E, conseqüentemente, a equação 4.32 resulta no limite superior para o CLR das conexões restantes.

4.8 Método Híbrido

A maior parte dos métodos de CAC já apresentados é capaz de lidar com diferentes tipos de tráfego, mas, no entanto, nenhum deles lida com diferentes objetivos de QoS. Objetivando suprir esta deficiência, Saito [42, 2] propõe um método híbrido para o CAC que trata requisitos de QoS heterogêneos. O CAC híbrido provê duas classes de CLR e introduz um mecanismo de prioridades no *buffer* de saída. O único parâmetro de tráfego utilizado pelo CAC híbrido é a taxa de pico, porém, como no CAC dinâmico, ele faz uso de medidas de tráfego. A alocação para a classe de alta qualidade é feita pela taxa de pico, enquanto que o CAC dinâmico é utilizado para a classe de baixa qualidade.

A alocação de banda passante pela taxa de pico para cada conexão provê um alta qualidade, porém com baixa utilização do canal. Já o CAC dinâmico fornece uma ótima utilização de canal através das medicões do tráfego de células, porém não lida com diferentes objetivos de QoS. O CAC híbrido foi criado combinando-se as vantagens destes dois métodos da seguinte forma: (1) somente a taxa de pico é especificada; (2) a taxa de utilização do canal é alta; e (3) a alta qualidade pode ser garantida para células de alta prioridade.

A reserva de banda passante para a classe de baixa prioridade (baixa qualidade) é feita através do CAC dinâmico com a especificação da taxa de pico. Para a classe de alta prioridade (alta qualidade), a reserva de banda passante é feita baseando-se somente na taxa de pico. Qualquer banda passante reservada para a classe de alta prioridade que não esteja sendo utilizada, assim como àquela não utilizada pela classe de baixa prioridade, é gerenciada pelo CAC dinâmico para a classe de baixa prioridade (figura 4.18)

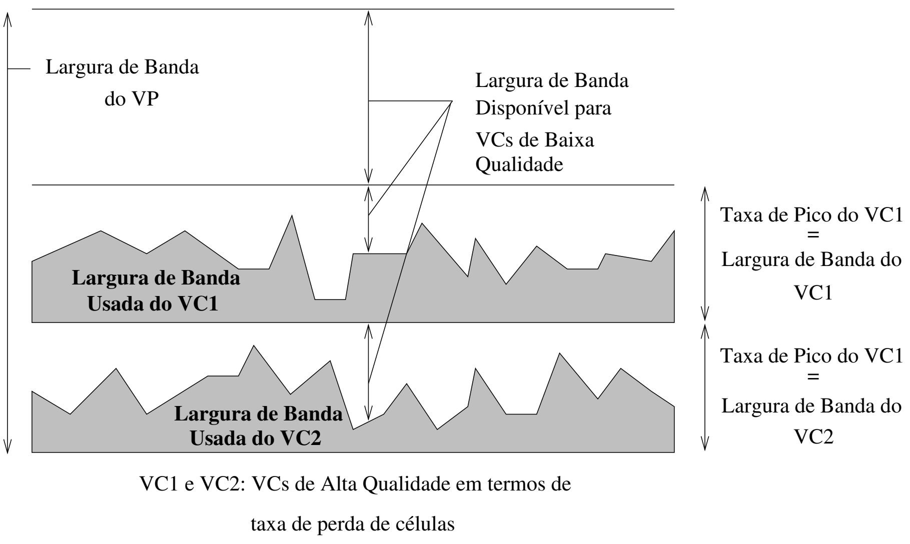

Figura 4.18: CAC híbrido.

Células de diferentes classes de prioridades compartilham um único VP sob um esquema de reserva de *buffer*. Neste esquema, uma célula de baixa prioridade é descartada caso o número de células no *buffer* de saída seja maior que um limiar especificado K' (figura 4.19).

4.9 Impacto do CDV sobre os mecanismos de CAC

O CDV (*Cell Delay Variation*), como já definido na seção 2.6.3, é um parâmetro que descreve a variação no atraso fim-a-fim de transferência de células. Ele é causado principalmente devido aos atrasos de fila nos nós de entrada e de comutação através da rede, podendo assumir valor positivos e negativos.

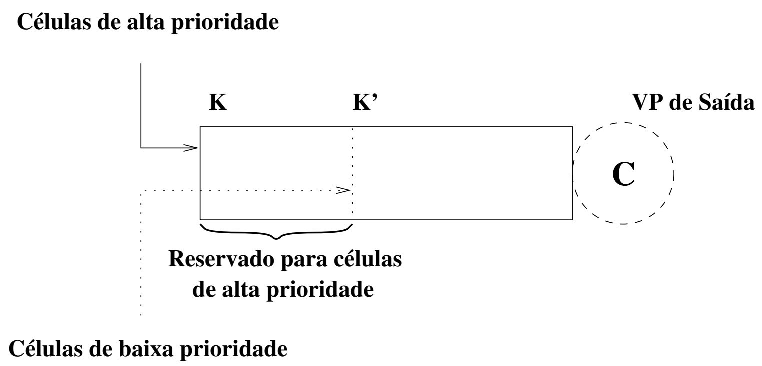

Figura 4.19: Esquema de reserva de *buffer* para controle da taxa de perda de células.

A ocorrência do CDV no nó de entrada da rede dificulta tanto os mecanismos de policiamento quanto as decisões de CAC visto que este CDV impede a estimação pela rede das características do tráfego real, mesmo que as características de fluxo de células de uma fonte estejam corretamente declaradas.

A ocorrência do CDV dentro de uma rede também dificulta as as decisões de CAC e interfere no controle de parâmetros de policiamento do tráfego entre redes. Este tipo de CDV ocorre principalmente devido ao armazenamento de células nos nós ATM e altera o padrão de chegadas do fluxo de células no nó seguinte. Como conseqüencia, a banda passante a ser alocada deve levar em consideração este impacto.

Existem três formas de se considerar o CDV no controle de tráfego [2]: avaliar o aumento do CDV em cada nó e adicionar uma banda passante extra; prover conformadores (*shapers*) de tráfego; e, por último, não considerar o impacto do CDV.

O primeiro método, apesar de intuitivamente atrativo, pode ser impraticável para o suporte dos serviços de telecomunicações a um preço baixo, pois a quantidade de banda passante extra pode ser tornar enorme e diminuir drasticamente o nível de utilização da rede. A razão para isto é que a banda passante extra deve ser dada para cada conexão. No entanto, esta banda passante adicional nem sempre é necessária, visto que o CDV também pode aumentar o espaçameno entre as células, causando a suavização do tráfego. Porém, como não é possível determinar quando a banda passante adicional será necessária, ela deve ser sempre provida para cada conexão.

O segundo método procura alterar a características do tráfego de entrada de modo a conformá-lo as características contratadas. A grande vantagem deste método é o fato dele não requerer uma banda passante adicional, apesar de causar atrasos devido ao armazenamento de células e perdas devido ao transbordo no conformador. O grande problema na sua implementação está no número de conformadores necessários. Por exemplo, numa rede com 100 de nós de interface entre redes de 600 Mbps, este número pode chegar a milhões caso fossem utilizadas conexões de 16 kbps.

O terceiro, e último, método não considera o impacto do CDV sobre os mecanismos de CAC. A base para esta decisão está na análise do processo de partida de um fluxo de células de um multiplexador. Nesta análise, é mostrado que, quando os processos de chegada são em rajadas, a banda passante inclui uma margem que se adapta a flutuação do tráfego mesmo quando o CDV não é considerado, isto é, tráfegos com níveis de rajada maiores do que o da distribuição de Poisson são suavizados ao passar por um nó de comutação ATM.

4.10 Resumo

Como foi visto neste capítulo, a diversidade de mecanismos de CAC propostos na literatura é muito grande. Muitas das vezes, estes mecanismos diferem por apenas alguns detalhes de implementação. Outras vezes, apresentam abordagens e métodos bem diferentes. Em vista disto, foi feita uma classificação destes mecanismos na tentativa de obter um melhor conhecimento sobre os mesmos e melhor compreender as suas diferenças. A comparação entre a eficiência destes mecanismos, entretanto, é muito difícil, pois cada um deles tenta capturar um aspecto particular do tráfego que é submetido à rede e apresenta uma caracterização própria do problema. Assim, torna-se mais importante ressaltar as vantagens e desvantagens de cada um destes métodos e os casos para os quais estes métodos são mais indicados.

Portanto, a avaliação de cada um dos mecanismos apresentados deve considerar a relação custo-benefício que existe entre a simplicidade de sua implementação e a eficiência por ele obtida, como já mencionado nos requisitos aos quais um mecanismo de CAC deve atender, bem como a abrangência alcançada por este mecanismo.

No grupo de métodos de alocação de capacidades, a cada fonte é atribuído um valor de capacidade efetiva, que varia entre a sua taxa média e a de pico, de acordo com o objetivo de QoS desejado. Vários métodos foram utilizados no cálculo destes valores.

Os métodos de simulação permitem a descrição de sistemas bem complexos e produzem resultados batante acurados. Porém, devido ao excessivo tempo de execução necessário, seu uso é apenas indicado para fins de validação de modelos teóricos.

O uso de métodos analíticos também produz resultados bastante acurados, mas difíceis de serem obtidos em tempo de execução. Assim, para que estes métodos possam ser aplicados, torna-se necessário a criação de classes de tráfego e de tabelas com os valores pré-calculados das capacidades efetivas de cada classe. A vantagem dos mecanismos que requerem uma classificação está na sua simplicidade, visto que a maior parte dos cálculos é feita previamente e não em tempo de execução. A sua desvantagem é que ela pode causar desperdícios de banda passante. Por exemplo, caso seja criada uma classe para conexões com taxas binárias de pico entre 2 Mbps e 6 Mbps, uma conexão de 2 Mbps já causaria um desperdício de banda passante de 4 Mbps. Uma alternativa para diminuir esta perda seria a criação de um maior número de classes com bandas passantes mais estreitas. No entanto, esta solução pode aumentar bastante a complexidades destes modelos.

As aproximações de estado estacionário são métodos bem simples e facilmente implementáveis. Seu uso é bastante indicado quando um grande número de fontes é multiplexado e o impacto de cada uma destas fontes individualmente não é crítico sobre o ganho de multiplexação estatística. Uma de suas desvantagens é que ela não considera o tamanho do *buffer* e, conseqüentemente, superestima os valores das capacidades efetivas. Esta superestimação também ocorre quando poucas fontes com baixos índices de rajada são multiplexadas.

O método da capacidade equivalente utiliza duas aproximações no cálculo das capacidades efetivas. A aproximação de estado estacionário é usada quando o efeito da multiplexação estatística é significativo, enquanto que a aproximação de fluxo contínuo é usada quando o impacto de cada fonte indivual é crítica. A aproximação de fluxo contínuo produz bons resultados quando um número pequeno de fontes com alto nível de utilização é multiplexado. Entretanto, quando aumentam-se tanto o número de fontes quanto o nível de utilização destas, seus resultados superestimam demasiadamente os valores das capacidades efetivas. Nestes casos, mesmo a aproximação de estado estacionário superestima os valores destas capacidades, principalmente quando o tamanho do *buffer* não é desprezível. Uma solução para este problema foi proposta em uma das extensões deste método. Nela, os valores dos parâmetros do modelo na aproximação de fluxo contínuo são calculados de maneira exata para um pequeno número de fontes e extrapolados logaritmicamente para um número maior de fontes. Assim, o ganho de multiplexação estatística é considerado e resultados bastante acurados podem ser obtidos para uma gama bem maior de valores.

As aproximações para fontes heterogêneas são bastante simples e produzem resultados relativamente acurados, como por exemplo no método não-linear. Entretanto, sua aplicação está condicionada a criação de tabelas com os valores pré-calculados para cada classe no caso homogêneo. Esta classificação, como já salientado, pode causar desperdícios de banda passante e uma conseqüente redução no ganho de multiplexação estatística.

O grupo de métodos de reservas rápidas é bastante adequado para fontes de tráfego em rajadas do tipo silêncio/atividade com níveis baixos de utilização. Nestas fontes, a chegada de uma nova rajada é um evento raro, mas normalmente de longa duração. Assim, como estes métodos fazem a alocação de recursos somente quando ocorre a chegada de uma rajada, uma grande economia de banda passante é conseguida. Entretanto, a sua utilização não traz bons resultados quando são multiplexadas fontes com níveis altos de utilização e rajadas de curta duração ou fontes que possuam uma variação contínua de taxas, como no caso do vídeo.

A aproximação não-paramétrica é um método simples e facilmente implementável. Sua grande vantagem está na não utilização de modelos específicos de tráfego, baseando no tráfego de pior caso para os parâmetros especificados. Entretanto, na prática, seus resultados superestiman demasiadamente a taxa de perda de células. Como conseqüência, um número bem menor de conexões pode ser aceito. A sua implementação se torna vantajosa quando feita em conjunto com medidas de tráfego, como no método dinâmico. Neste caso, o efeito da superestimação inicial da taxa de perda de células é reduzido ao longo do tempo, conforme novas medidas vão sendo feitas.

Os métodos híbridos são bastante vantajosos por permitirem que fontes com diferentes objetivos de QoS compartilhem o mesmo canal, o que aumenta o efeito da multiplexação estatística. Contudo, maiores estudos precisam ser feitos para verificar a viabilidade destes mecanismos.

De um modo geral, os mecanismos de CAC que fazem uso do comprimento médio da rajada tendem a produzir melhores resultados em relação aos que não os usam. Todavia, esta vantagem é menor nos casos onde o comprimento médio da rajada é muito maior do que o tamanho do *buffer*, como nas fontes de vídeo.

Por sua vez, os métodos dinâmicos de CAC, em princípio, tendem a produzir melhores resultados do que os outros mecanismos, visto que o seu conhecimento sobre a distribuição do processo de chegada das fontes é bem melhor. Porém, algumas questões continuam em aberto como a estabilidade destes mecanismos e a escolha ideal do grau de atualização das medidas de tráfego. Além disto, não se sabe ao certo se estas medidas podem ser feitas em tempo hábil para que possam ser utilizadas.