Universidade Estadual de Campinas
FACULDADE DE ENGENHARIA ELÉTRICA E DE COMPUTAÇÃO
DEPARTAMENTO DE TELEMÁTICA

Análise e Dimensionamento de Enlaces em Redes de Telecomunicações Multisserviços

Dissertação apresentada à Faculdade de Engenharia Elétrica e de Computação da Universidade Estadual de Campinas, como parte dos requisitos exigidos para a obtenção do título de Mestre em Engenharia Elétrica.

Luís Renato Bercho
Graduado em Engenharia Elétrica pela Escola Politécnica da Universidade de São Paulo

Orientador:

Prof. Dr. Ivanil S. Bonatti

Banca Examinadora:

Prof. Dr. Ivanil Sebastião Bonatti (presidente)	**FEEC – Unicamp**
Prof. Dr. Michel Daoud Yacoub	**FEEC – Unicamp**
Dr. Paulo Cardieri	**Fundação CPqD**
Prof. Dr. Shusaburo Motoyama	**FEEC – Unicamp**

Agosto de 2002

Resumo

Neste trabalho são abordados os procedimentos de análise e síntese de redes multisserviços baseados em três aspectos complementares. No primeiro, são considerados enlaces individuais transportando tráfegos de fontes distintas. Métodos de cálculo de banda efetiva são aplicados aos modelos de fontes para perdas e para atraso de pacotes presentes no nó de comutação, tendo em vista os parâmetros descritores de tráfego disponíveis. Comparações analíticas e numéricas são feitas para diferentes fontes de tráfego. No segundo, probabilidades de bloqueio de chamadas com requisitos distintos são obtidas para sistemas multiclasses comutados. Algumas políticas de reserva de recursos são comparadas para uso em dimensionamento de enlaces e proteção de classes. No terceiro, é proposta uma formulação matricial estendendo os resultados de bloqueio em enlaces para as redes, por meio da resolução das equações de ponto fixo para sistemas multiclasses.

Abstract

This work presents the analysis and synthesis procedures for multiservice loss networks based on three complementary aspects. Firstly, individual links carrying traffic from different sources are considered. Distinctive methods of calculation for effective bandwidth are applied to the source models for loss and delay of packets on the switching nodes, taking into account the number of traffic description parameters available. Numerical and analytical comparisons are made for several traffic sources. Secondly, blocking probabilities of call requests having distinct requirements are obtained for switched multiclass systems by different methods. Some resource reservation policies are compared, for both link dimensioning procedure and service class protection. Finally, a matrix notation extending the blocking results from links to networks is proposed to solve fixed point non-linear equations when applied to multiclass systems.

para Cíntia

Non omnis possumus omnes
(Virgílio 70-19 a.C.)

Agradecimentos

- Ao professor Ivanil Sebastião Bonatti, não só pelo seu incondicional apoio e dedicação na tarefa de orientação, mas sobretudo pela amizade e atenção dispendida durante toda esta trajetória dos últimos três anos;
- Ao professor Pedro Peres, outro exemplo de competência e solicitude, pelas valiosas dicas no uso do LaTeX e demais sugestões para o enriquecimento dos trabalhos desenvolvidos;
- Aos membros da banca, professores e doutores Cardieri, Motoyama e Yacoub pela atenção e paciência;
- A todos os amigos e amigas do DT, que nos últimos anos foram uma presença constante e inestimável do dia-a-dia na universidade;
- À minha esposa Cíntia, por todo o carinho e apoio, especialmente naqueles momentos difíceis durante a elaboração deste trabalho;
- Ao CNPq, pelo apoio financeiro;
- Finalmente, mas não menos importante, aos meus pais, sem os quais nada disto teria sido possível.

Acrônimos e Abreviaturas

ATM	Asynchronous Transfer Mode
CAC	Connection Admission Control
CBR	Constant Bit Rate
CDV	Cell Delay Variation
CLR	Cell Loss Ratio
COST	European Cooperation in the field of Scientific and Technical Research
CTD	Cell Transfer Delay
FIFO	First In First Out
IP	Internet Protocol
ITU	International Telecommunication Union
ITU-T	ITU Telecommunication Standardization Sector
MaxCTD	Maximum Cell Transfer Delay
MBS	Maximum Burst Size
MMPP	Markov-Modulated Poisson Process
MPLS	Multi-Protocol Label Switching
PCR	Peak Cell Rate
PNNI	Private Network to Network Interface
PVC	Permanent Virtual Channel
QoS	Quality of Service
RR	Round Robin
SCR	Sustainable Cell Rate
SVC	Switched Virtual Channel
UBR	Unspecified Bit Rate
VBR	Variable Bit Rate
VBR-rt	VBR real time
VBR-nrt	VBR non-real time
WRR	Weighted Round Robin

Conteúdo

Capítulo 1

Introdução

As redes de telecomunicações multisserviços são apresentadas para a contextualização do trabalho, bem como dos métodos utilizados nos capítulos subseqüentes.

1.1 Redes de dados multisserviços

As redes multisserviços de banda larga foram previstas há mais de vinte anos, numa época em que a comutação de pacotes era tão somente um método alternativo para transmitir dados. Hoje tais previsões já fazem parte da realidade, na qual os sistemas de telecomunicações desempenham um papel de grande relevância. Prediz-se que o crescimento exponencial das redes de comutação de pacotes, tais como a Internet e/ou o modo de transferência assíncrono (ATM) continuará ainda por algum tempo, e que seu volume de tráfego excederá em breve àquele das redes tradicionais de telefonia. Tão logo isto ocorra, ou mesmo antes, haverá uma pressão muito forte para que todos os serviços de comunicação façam uso do serviço mais barato e eficiente de comunicação disponível. A maneira mais adequada de se atingir este objetivo é fazendo com que todos os serviços utilizem uma rede multisserviços integrada. De fato, isto já é um processo em curso, e tanto os órgãos regulamentadores, como a ITU-T, quanto as indústrias de comunicação, computadores e entretenimento têm se reorganizado de modo a antecipar este desenvolvimento, que é algumas vezes referido como *convergência* das redes de comunicação. Para tanto, existem pelo menos duas razões relevantes [1]:

i) Uma rede é mais barata do que duas (ou mais) para os usuários;

ii) Serviços freqüentemente necessitam de acesso a outros serviços, o que se torna difícil caso estejam em redes diferentes.

1.1.1 Caracterização

Nas redes de comutação de pacotes os dados são segmentados em blocos, que podem ser pacotes Internet ou células ATM. Para o propósito deste trabalho, as especificidades da segmentação de tráfego não são importantes; a palavra pacote é indistintamente utilizada em referência à unidade básica de tráfego, a menos que explicitamente mencionado.

Tais pacotes são comutados entre nós de comutação (*switches*) de acordo com a informação contida em seus cabeçalhos. Quando células chegam a um nó de comutação, ou

quando estão prontas para deixá-lo, caso um número excessivo de pacotes necessite utilizar o mesmo enlace ao mesmo tempo, estas deverão ser armazenadas em um *buffer* aguardando transmissão.

Uma rede multisserviços é caracterizada por possuir diversas classes de serviço com características distintas entre si, e que são definidas como um denominador comum para serviços semelhantes (voz, vídeo, dados não tempo-real, etc.). Serviços agrupados na mesma classe de serviço são tratados de maneira idêntica e homogênea. Todas as classes compartilham os mesmos recursos de rede, podendo haver ou não privilégios de acesso de alguma delas ao sistema – como por exemplo reserva de recursos ou rotas exclusivas através dos enlaces – desde que seja especificado *a priori* na concepção do projeto e provisionado pelo gerenciamento da rede.

O processo de ingresso de elementos individuais no sistema é suposto estocástico, com uma distribuição cujos parâmetros são conhecidos para cada uma das classes definidas. Como a tecnologia de transmissão é baseada no uso de pacotes ou células, o tempo de atendimento do servidor depende tanto da taxa de transmissão utilizada como do tamanho dos segmentos de dados. Pacotes que possuam tamanho estatisticamente variável implicam num tempo de atendimento aleatório, com uma distribuição conhecida. Alguns exemplos possíveis [1] são as distribuições de Weibull, Pareto e Gaussiana. Em geral, em razão da facilidade de tratamento, a distribuição exponencial negativa é uma das mais adotadas. Pacotes com tamanho fixo – células – têm por sua vez o tempo de atendimento constante, função apenas de seu comprimento.

1.1.2 Tipos de redes e QoS

As redes multisserviços podem ser divididas em dois grandes grupos, de acordo com o caminho seguido pelos pacotes através da rede:

- **Sem conexão ou 'connectionless'**: Cada pacote é encaminhado individualmente da fonte para o destino, podendo inclusive cada um deles seguir percursos diferentes. Exemplo: redes IP clássico.

- **Orientada a conexão**: O caminho lógico através da rede é estabelecido antes da transmissão e todos os pacotes seguem o mesmo percurso. As conexões virtuais estabelecidas podem ser do tipo permanente (PVC) ou comutado (SVC), dependendo se são realizadas pelo provisionamento do gerente de rede ou se são conectadas e desconectadas conforme a dinâmica das chamadas (requisições de conexão) individuais. Exemplo: redes ATM, sistemas MPLS.

Um conceito de grande importância para as redes de telecomunicações é o conceito de Qualidade de Serviço (*Quality of Service* ou simplesmente QoS), que pode ser enunciado da seguinte forma:

> *Parâmetros estabelecidos de comum acordo entre a rede e o usuário para assegurar que o serviço de transporte apresente um desempenho mínimo aceitável.*

Os parâmetros especificados de QoS podem ser, por exemplo, largura de banda, atraso fim-a-fim, taxa de perda de pacotes/células, probabilidade de bloqueio de acesso à rede. O *contrato de serviço* entre usuário e rede pode ser especificado antes da operação do

serviço propriamente dito para toda uma classe. Pode também ser especificado de maneira automática, para cada requisição de conexão de uma chamada individual pertencente a uma dada classe. Neste caso, um procedimento conhecido como *Controle de Admissão de Chamada* (Call Admission Control ou CAC) é executado para cada requisição a fim de verificar se a rede possui recursos suficientes para atender às especificações de QoS da classe de serviço a qual a requisição pertence. Caso contrário, ou a chamada é bloqueada e deixa o sistema sem ser atendida ou é renegociada para uma outra QoS.

Redes orientadas à conexão têm melhores condições de garantir a QoS, uma vez que por estabelecerem caminhos fixos podem controlar melhor a variação dos parâmetros (perda, bloqueio, atraso, etc.) através dos enlaces individuais que compõem a conexão lógica. Redes sem conexão, por outro lado, apresentam uma dispersão muito grande destes parâmetros para cada pacote individual que atravessa a rede devido às diferentes trajetórias percorridas.

1.2 Abordagem

Duas outras conceituações específicas para a abordagem das redes de telecomunicações são utilizadas ao longo do desenvolvimento deste trabalho. A primeira é a diferença existente entre os processos de *análise* e de *síntese* de uma rede, que podem ser expressos como:

> **Análise**: Avaliação do desempenho de redes, planejadas ou já em operação.
>
> **Síntese**: Dimensionamento de redes a serem implantadas segundo uma especificação de QoS, custo, etc.

A segunda é a divisão – para redes orientadas a conexão – entre os tempos de pré-conexão e os de pós-conexão, definidos da seguinte forma:

> **Pré-conexão**: Os procedimentos que ocorrem antes do estabelecimento da conexão e que afetam o desempenho da rede percebido pelos usuários devido ao bloqueio de requisições por insuficiência de recursos disponíveis.
>
> **Pós-conexão**: Os processos de transporte que ocorrem enquanto durar a conexão e que afetam o desempenho das chamadas em progresso (QoS) devido à natureza estocástica do tráfego de pacotes.

A escolha de caminhos diferentes através da rede (i.e. o encaminhamento) pode afetar tanto a QoS de conexões estabelecidas quanto a probabilidade de bloqueio das conexões ingressantes – chamadas – devido às características particulares de cada um dos enlaces que compõem o caminho escolhido.

O escopo deste trabalho é a análise e a síntese de redes orientadas à conexão do tipo PVC ou SVC, cujos tamanhos de pacote sejam variáveis ou fixos, para ambos os tempos de pré e de pós-conexão.

Este trabalho está dividido em três partes – banda efetiva, bloqueio em enlaces e bloqueio em rede – com cada capítulo focalizando um aspecto distinto, além de uma introdução e uma conclusão geral.

Banda efetiva

Neste capítulo é descrita a metodologia de banda efetiva. Esta ferramenta, que pode ser utilizada tanto na análise como na síntese de redes, estabelece requisitos de banda necessários para o atendimento de uma certa QoS de cada conexão em particular pertencente a uma classe de serviço. É considerada essencial como método de cálculo para conexões já estabelecidas através da rede – pós-conexão – a fim de assegurar uma perda máxima aceitável devido ao congestionamento de *buffer* dos equipamentos presentes nos nós de comutação. Também os atrasos de transmissão associados a uma banda efetiva calculada podem ser especificados e/ou verificados como parte da QoS.

Em tempo de estabelecimento de chamada – pré-conexão – pode ser importante para verificar se a banda necessária para uma requisição com QoS dada pode ser admitida pela rede, durante o processo de CAC.

Bloqueio em enlaces

Este capítulo trata dos métodos de cálculo utilizados para estimar a porcentagem de bloqueio de acesso a um enlace compartilhado por diferentes classes de serviço durante o processo de requisição de estabelecimento de conexão. É portanto um procedimento válido para conexões do tipo SVC, em tempo de pré-conexão apenas, e que pode ser utilizado tanto para análise como para a síntese de redes.

Classes com diferentes requisitos de banda – ou circuitos – presentes em um mesmo enlace interagem mutuamente nas respectivas probabilidades de bloqueio. A reserva de recursos para determinadas classes pode ser usada para melhorar o desempenho de bloqueio de algumas delas em detrimento de outras.

Bloqueio em rede

Neste capítulo é discutido o método de cálculo conhecido como equações de ponto fixo, que estende os resultados da probabilidade de bloqueio verificadas nos enlaces para uma rede. É um método geralmente utilizado em análise, embora com algumas adições possa ser adaptado para síntese de redes.

Este procedimento possibilita, além da determinação de bloqueio fim-a-fim numa rede, a estimativa do atraso de um fluxo de dados através dos caminhos virtuais existentes.

Capítulo 2

Banda Efetiva

A definição de banda efetiva, bem como alguns métodos multi-paramétricos de cálculo, são apresentados neste capítulo para modelos comuns de fontes de tráfego. Vantagens e desvantagens das diversas formulações são também discutidas e um modelo uni-paramétrico é proposto.

2.1 Introdução

Banda efetiva é uma metodologia de cálculo bastante difundida dentre as técnicas empregadas na análise, síntese e operação das redes multisserviços. Dentre as possíveis aplicações deste conceito duas delas se destacam. Uma é no dimensionamento das redes multisserviços onde coexistam diferentes classes com requisitos distintos de largura de banda, atraso, *jitter*, taxa de erro máximo admissível, etc. Além disto, parâmetros descritores que são particulares de cada classe, tais como as respectivas taxa de pico, taxa média e duração de surtos devem ser convertidos de preferência a um mesmo denominador comum para facilitar o tratamento pelos algoritmos de dimensionamento. Assim, a transformação de todos estes requisitos em uma métrica única para cada classe – a banda efetiva – simplifica o procedimento de cálculo quando do planejamento da rede. A segunda aplicação importante é nos algoritmos de controle de admissão de conexões (CAC), onde a rede deve tomar uma série de ações de forma a determinar se uma chamada (requisição de conexão) pode ser aceita ou não. Para isso, uma estimativa tanto dos requisitos da nova chamada bem como dos recursos e conexões presentes na rede deve ser feita, e um método eficaz de cálculo é obtido pelo uso do conceito de banda efetiva e de suas propriedades de aditividade e independência.

2.1.1 Definições

A *banda efetiva* de uma requisição de chamada (no caso de SVC's) ou de conexão já estabelecida (para PVC's) é definida como sendo o mínimo de banda requerida para uma classe de serviço – caracterizada por um conjunto de descritores de tráfego – que deve necessariamente atender a uma qualidade de serviço (QoS) previamente especificada. Com freqüência, estes requisitos de QoS são reduzidos a duas condições essenciais que servem de base para todos os métodos de cálculo empregados: a primeira, de que não seja excedida uma dada probabilidade de congestionamento (ou *overflow*) de *buffer* em um nó qualquer

da rede e, portanto, uma certa probabilidade de perda de células ou pacotes [33], [39], [43]; a segunda, de que o atraso máximo devido ao tempo de permanência nos *buffers* não exceda um valor limite especificado [9], [39].

Para a utilização eficiente do conceito de banda efetiva duas propriedades são desejáveis:

i) *Independência*: A banda efetiva de um tráfego deve ser independente dos demais tráfegos com os quais venha a ser eventualmente agregado;

ii) *Aditividade*: A banda efetiva da soma de tráfegos deve ser igual a soma das bandas efetivas de cada um dos constituintes.

Essas propriedades, se satisfeitas, são úteis para técnicas de alocação de banda, pois permitem que tanto fluxos de taxa constante (CBR) como fluxos com taxa variável (VBR) sejam tratados como fluxos equivalentes de taxa constante e cujo requisito de banda é o valor assim calculado. Nessas condições, o atendimento do tráfego com banda equivalente ao longo da rede implica que os requisitos de QoS especificados para o tráfego original sejam satisfeitos.

2.1.2 Tipos de modelos

Dentre as possíveis formulações do conceito banda efetiva destacam-se duas abordagens distintas do problema.

Modelo de perdas

Seja um processo genérico de chegada $\mathcal{A}(t)$ cujo tráfego é oferecido a uma fila Q de capacidade infinita e com um certo limiar b, conforme a Figura 2.1:

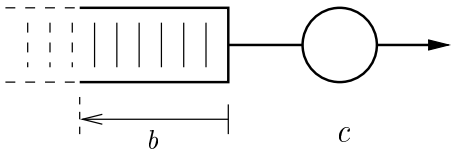

Figura 2.1: Fila infinita com limiar

A banda efetiva associada é denotada por c. Este valor é definido como a mínima capacidade necessária (taxa de transmissão) tal que a probabilidade de que seja excedido o limiar dado b é menor que um limite ε pré-definido tal que

$$Prob\{Q \geq b\} \leq \varepsilon \tag{2.1}$$

Esta definição é a base para a modelagem conhecido como *restrição devido a perdas*, onde o limite ε é tomado como uma probabilidade de perda ou descarte de elementos, células ou pacotes, devido ao congestionamento (por exemplo, o parâmetro de perda de células CLR das redes ATM). O limiar b é definido pela capacidade máxima do *buffer* de armazenamento de células ou pacotes. Nesta definição de banda efetiva assume-se que o tamanho b do *buffer* seja conhecido *a priori*, e cujo valor não seja re-dimensionado quando a capacidade calculada c é alterada.

Modelo de atraso

O requisito principal neste modelo passa a ser que o atraso verificado no *buffer* não exceda um valor máximo admissível d. O último elemento a entrar no *buffer*, imediatamente antes do preenchimento de sua capacidade, está sujeito a um atraso que é proporcional ao produto do tamanho do *buffer* pelo inverso da taxa de atendimento do servidor. Esta restrição é dada por

$$\left(\frac{b}{c}\right) \leq d \tag{2.2}$$

para a taxa de serviço c. Esta condição serve de base para a modelagem conhecido por *restrição pelo atraso*, no qual o limite para o atendimento de uma dada QoS é imposto pelo tempo máximo de latência aceitável (MaxCTD nas redes ATM). Outra variante possível é considerar que o tamanho do *buffer* também está sujeito ao dimensionamento.

O dimensionamento simultâneo dos parâmetros c e b, quando possível, permite um melhor aproveitamento do ganho estatístico de multiplexação para diversas fontes [39].

2.1.3 Parâmetros de modelagem

Diversas propostas têm sido apresentadas com diferentes formulações para o cálculo de banda efetiva [33], [9], [43], [8] e capacidade de enlace [29], [35], bem como para o tamanho de *buffer* associado [37].

Para os cálculos da banda efetiva, os modelos valem-se de características extraídas das classes de serviço apresentadas ao sistema. Três, dois ou mesmo apenas um dos parâmetros descritores de tráfego são utilizados neste processo. São eles, em geral, a taxa média de transmissão da fonte (obtida do SCR), a taxa de pico (PCR) e alguma medida associada à duração do surto (MBS ou equivalente).

Nem sempre todos os parâmetros estão disponíveis para serem utilizados na determinação da banda efetiva, especialmente no dimensionamento (síntese) de redes. Com freqüência, muitos deles são apenas estimativas grosseiras, o que leva a imprecisões cumulativas nos resultados obtidos. Essa dificuldade estimula o uso de modelos com menor número de parâmetros.

Embora não tão significativo na análise e síntese de redes, outro motivo para o uso de modelos com poucos parâmetros é a simplificação dos cálculos. Uma vez que nas aplicações de controle de admissão de chamadas (CAC) a estimativa de banda deve ser feita em tempo real, a eficiência é imprescindível, ainda que às custas de uma menor precisão dos valores obtidos.

2.2 Modelos de três parâmetros

Nessas formulações os parâmetros especificados são a taxa de pico, a taxa média e o comprimento máximo (ou médio) de surto. Tais modelos se aplicam a fontes que possam ser descritas segundo estes parâmetros, notadamente as do tipo *on-off*. Fontes poissonianas não se enquadram nestas formulações.

2.2.1 Método de momento logarítmico

O modelo proposto por Kesidis et al. [33] é um modelo de perdas, baseado no cálculo de uma função $H(\cdot)$ geradora de momentos logarítmicos assintóticos. Seja uma fila de capacidade infinita e taxa de serviço C. A situação de congestionamento do *buffer* é modelada como a probabilidade de encontrar-se mais que b elementos na fila, de modo que para uma certa probabilidade máxima de perda admitida ε tem-se

$$Prob\{Q > b\} \leq e^{-b\delta} \tag{2.3}$$

onde $e^{-b\delta} = \varepsilon$. Para N fontes de tráfego multiplexadas no *buffer*, a banda efetiva calculada para cada fonte c_i depende exclusivamente da respectiva fonte e do parâmetro real δ. Para atingir o grau de serviço requisitado é necessário que

$$\sum_{i=1}^{N} c_i \leq C \tag{2.4}$$

Seja $H_i(\delta)$ a função geradora de momentos logarítmicos (neperiano) para um modelo de fonte, suposta ergódica, definida por

$$H_i(\delta) = \lim_{t \to +\infty} \frac{\log E\{e^{\mathcal{A}_i(t)\delta}\}}{t} \tag{2.5}$$

com $\mathcal{A}_i(t)$ representando o processo (tempo) de chegada de elementos gerados pela fonte i no intervalo de tempo $(0, t)$. Caso a função $H_i(\delta)$ seja convexa e crescente para todo δ real então a banda efetiva para esta fonte pode ser calculada como $c_i = H_i(\delta)/\delta$, ou seja [33]

$$c_i = \frac{1}{\delta} \lim_{t \to +\infty} \frac{\log E\{e^{\mathcal{A}_i(t)\delta}\}}{t} \tag{2.6}$$

As propriedades de aditividade e independência também se preservam [40]. Se $\mathcal{A}_1$ e $\mathcal{A}_2$ são variáveis independentes, com o auxílio da transformada Z tem-se

$$E\{z^{(\mathcal{A}_1+\mathcal{A}_2)}\} = E\{z^{\mathcal{A}_1}\}E\{z^{\mathcal{A}_2}\} \tag{2.7}$$

implicando

$$\log E\{z^{(\mathcal{A}_1+\mathcal{A}_2)}\} = \log E\{z^{\mathcal{A}_1}\} + \log E\{z^{\mathcal{A}_2}\} \tag{2.8}$$

o que equivale dizer que $c_{1+2} = c_1 + c_2$.

A principal característica deste método é tratar os diversos modelos de fontes no cálculo da equação (2.6) numa forma analítica.

Fontes MMPP

Uma fonte é do tipo MMPP (*Markov-Modulated Poisson Process*) se a geração de elementos de tráfego segue uma distribuição poissoniana cuja taxa $\lambda(t)$ é uma função variante no tempo definida por uma cadeia de Markov. O modelo consiste de uma fonte de tráfego com um número Φ finito de fases cujas transições dão-se conforme a cadeia de Markov. Em uma i-ésima fase a fonte emite segundo uma distribuição poissoniana de taxa média λ_i associada àquela fase. Duas matrizes, Λ e R, caracterizam a fonte MMPP. A matriz

$\Lambda(\Phi \times \Phi)$ é uma matriz diagonal com as taxas médias de geração de elementos em cada fase e a matriz $R(\Phi \times \Phi)$ representa a taxa de transição entre as fases.

O vetor linha q de probabilidades estacionárias de fase satisfaz as condições

$$\begin{cases} q \cdot R &= \mathbf{0}; \\ q \cdot u &= 1; \end{cases} \tag{2.9}$$

onde $\mathbf{0}$ é o vetor linha nulo e u é o vetor coluna unitário. A *matriz de probabilidades* $P(n,t)$, cujas componentes $P_{x,y}(n,t)$ representam a probabilidade de que a fonte gere n elementos no intervalo de tempo t, dado que a fonte encontra-se numa certa fase x no início do intervalo e numa fase y ao final, é dada conforme Baiochi et al. por [5]

$$P(n,t) = \frac{e^{(R-\Lambda)t}(\Lambda t)^n}{n!} \tag{2.10}$$

e portanto

$$E\{\exp[\mathcal{A}(t)\delta]\} = q\sum_{n=0}^{+\infty} e^{n\delta}\frac{e^{(R-\Lambda)t}(\Lambda t)^n}{n!}u = qe^{[R+(e^{\delta}-1)\Lambda]t}u \tag{2.11}$$

e com a banda efetiva c da fonte é expressa após cálculos como

$$c = \frac{1}{\delta}eigmax[R + (e^{\delta} - 1)\Lambda] \tag{2.12}$$

onde $eigmax[\cdot]$ é o maior autovalor da matriz entre os colchetes.

No caso particular de uma fonte MMPP de duas fases do tipo *on-off*, a fonte emite segundo um processo poissoniano de taxa média λ quando no estado *on* e com taxa zero quando no estado *off*, sendo descrita pelas matrizes

$$\Lambda = \begin{bmatrix} \lambda & 0 \\ 0 & 0 \end{bmatrix}; \qquad R = \begin{bmatrix} -1/T_{\text{on}} & 1/T_{\text{on}} \\ 1/T_{\text{off}} & -1/T_{\text{off}} \end{bmatrix}; \tag{2.13}$$

onde T_{on} e T_{off} são os tempos médios de permanência em cada fase.

A banda efetiva calculada vale [33], [40]

$$c_i = \alpha + \sqrt{\alpha^2 + \beta^2} \tag{2.14}$$

$$\alpha = \frac{1}{2\delta}\left((e^{\delta} - 1)\lambda - \frac{1}{T_{\text{on}}} - \frac{1}{T_{\text{off}}}\right) \tag{2.15}$$

$$\beta^2 = \frac{\lambda}{T_{\text{off}}}\frac{(e^{\delta} - 1)}{\delta^2} \tag{2.16}$$

Fluido markoviano on-off

As fontes do tipo fluido markoviano *on-off* são similares às fontes MMPP, com a fonte emitindo a uma taxa constante em cada uma das fases. A banda efetiva em função das matrizes Λ e R é dada por [5]

$$c_i = \frac{1}{\delta}eigmax[R + \delta\Lambda] \tag{2.17}$$

No caso das fontes de duas fases, onde a fonte emite com taxa λ constante na fase *on* e com taxa zero na fase *off*, a banda efetiva é calculada por [33], [40]

$$c_i = \alpha + \sqrt{\alpha^2 + \beta} \tag{2.18}$$

$$\alpha = \frac{1}{2\delta}\left(\lambda\delta - \frac{1}{T_{\text{on}}} - \frac{1}{T_{\text{off}}}\right) \tag{2.19}$$

$$\beta = \frac{\lambda}{T_{\text{off}}}\frac{1}{\delta} \tag{2.20}$$

Para valores idênticos de λ, T_{on} e T_{off} as fontes tipo MMPP têm bandas efetivas maiores que as fontes tipo fluido markoviano.

Do ponto de vista do dimensionamento, ainda que um cálculo individual seja necessário para cada fonte modelada, o método de Kesidis é um procedimento flexível e que produz bons resultados.

Diversos outros métodos valem-se do momento logarítmico para o cálculo de banda efetiva. O modelo proposto por Chang e Thomas [11] é uma variante do modelo proposto por Kesidis, sob uma óptica distinta. Por meio de um raciocínio similar ao utilizado na mecânica estatística para o teorema H de Boltzmann, o momento logarítmico é interpretado como uma função de energia. Já no trabalho de Elwalid e Mitra [21], uma extensão do método de Kesidis é proposta para o tratamento de fontes do tipo MMPP, além do *on-off*, notadamente para o caso de múltiplas fases. Segundo os autores, esta aproximação é mais adequada para a modelagem de tráfego de vídeoconferência.

2.2.2 Modelo de Guérin

Um dos primeiros modelos de três parâmetros descritores proposto é o apresentado por Guérin et al. em [27], baseado em três parâmetros descritores. Trata-se de um modelo de perdas cujo principal objetivo é a eficiência de cálculo pelo uso de equações algébricas simples. A *capacidade equivalente* C para as conexões de fluxos a serem multiplexados é apresentada como a quantidade de banda *total* necessária (agregada) para que seja atendido um grau de serviço dado em um enlace da rede. O método é a combinação de duas métricas distintas.

Seja um tráfego do tipo fluido markoviano *on-off* com geração dada por $\mathcal{A}_i(t)$, descrito pelos parâmetros taxa de pico de transmissão $\hat{\lambda}_i$, taxa média λ_i e tamanho médio do período de surto denotado por $\bar{p}_i$. Um conjunto de N fontes agregadas é oferecido a um *buffer* de tamanho b que é esvaziado a uma taxa de transmissão constante. Define-se a primeira métrica *capacidade equivalente* para o modelo fluido markoviano C_I como sendo

$$C_I = \sum_{i=1}^{N}\left[\frac{\Delta\bar{p}_i T_i\hat{\lambda}_i - b + \sqrt{[\Delta\bar{p}_i T_i\hat{\lambda}_i - b]^2 + 4b\Delta\bar{p}_i T_i(1-T_i)\hat{\lambda}_i}}{2\Delta\bar{p}_i T_i}\right] \tag{2.21}$$

com $\Delta = -\log(\varepsilon)$ calculado a partir do limiar de perdas ε e $T_i = 1 - \lambda_i/\hat{\lambda}_i$. A expressão (2.21) superestima o valor da capacidade sempre que existam ganhos estatísticos de multiplexação [27]. Em consequência, a segunda métrica adotada C_{II} é baseada em uma

aproximação estacionária gaussiana

$$C_{II} = \sum_{i=1}^{N} \lambda_i + \sqrt{(2\Delta - \log(2\pi)) \sum_{i=1}^{N} \lambda_i (\hat{\lambda}_i - \lambda_i)} \tag{2.22}$$

A banda equivalente C calculada é o mínimo entre C_I e C_{II}

$$C = \min\{C_I, C_{II}\} \tag{2.23}$$

O valor obtido pelo método ainda é conservador, de modo que a banda equivalente permanece superestimada. Nos casos em que a aproximação gaussiana utilizada para obter a equação (2.22) não seja válida, isto é, quando os tráfegos agregados $\mathcal{A}_i$ diferirem muito em seus parâmetros individuais, o método pode tornar-se bastante impreciso [27]. O resultado obtido é sempre o valor *agregado* de banda equivalente. Para qualquer alteração de um dos tráfegos constituintes, é necessário o recálculo das equações (2.21) e (2.22).

2.2.3 Modelo de Le Boudec

Le Boudec propõe em [9] e [10] um modelo de atraso determinístico de pior caso, baseado em tempos de preenchimento e esvaziamento de *buffer*, para uma estimativa de banda efetiva.

Para uma fila com tamanho de *buffer* infinito servida à taxa c é possível assegurar um limite máximo d para o atraso de células caso

$$c = \max_t \frac{\mathcal{A}(t)}{t+d}; \quad t > 0 \tag{2.24}$$

onde $\mathcal{A}(t)$ é o número de elementos (células ou pacotes) de tráfego gerados pela fonte no intervalo $(0, t)$. Como o modelo não considera perdas, o *buffer* deve possuir um tamanho b tal que:

$$b = \sum_{i=1}^{N} c_i d_i \tag{2.25}$$

onde c_i e d_i são respectivamente a banda efetiva e o atraso máximo suportado para cada uma das N fontes presentes no enlace.

Uma fonte determinística variável com taxa de geração de elementos $\mathcal{A}_i(t)$, modelada por um sistema de dois estados *on-off* é definida pelos parâmetros: taxa de pico $\hat{\lambda}_i$, taxa média λ_i e duração máxima do surto $\hat{p}_i$. O atraso de pior caso é verificado quando um surto de tamanho máximo ocorre ininterruptamente. Os tempos T_{on} e T_{off} são determinados a partir dos parâmetros da fonte pelas relações

$$\hat{\lambda}_i = \frac{\hat{p}_i}{T_{\text{on}}} \tag{2.26}$$

$$\lambda_i = \frac{T_{\text{on}}}{T_{\text{on}} + T_{\text{off}}} \hat{\lambda}_i \tag{2.27}$$

A Figura 2.2 ilustra o preenchimento e o esvaziamento do *buffer* operando a uma taxa c, quando da ocorrência de um surto.

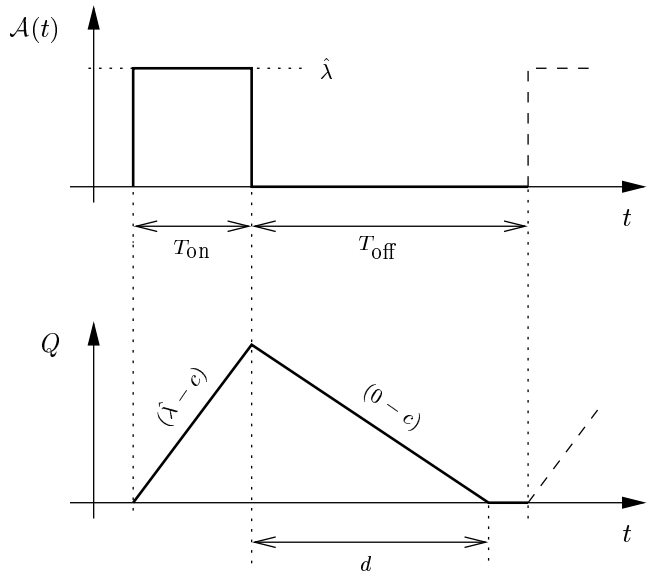

Figura 2.2: Fonte on-off (em cima) e tamanho de fila no buffer (embaixo)

O *buffer* é preenchido segundo a função $(\hat{\lambda} - c)t$ e esvaziado conforme ct. O número de elementos remanescentes ao final do surto no *buffer*, suposto inicialmente vazio é, portanto, $(\hat{\lambda} - c)T_{\text{on}}$. Esta quantidade de elementos deve ser esvaziada no tempo máximo d, a uma taxa c. Assim, a banda efetiva para uma fonte i é dada por:

$$c_i = \frac{\hat{\lambda}_i T_{\text{on}}}{T_{\text{on}} + d} \tag{2.28}$$

onde o parâmetro d representa o máximo atraso permitido no *buffer* (MaxCTD em sistemas ATM) ou mesmo a máxima variação permitida para este atraso (CDV, analogamente). A fim de garantir o completo esvaziamento do *buffer* após a ocorrência de cada surto a banda efetiva c_i deve ser maior que a taxa média λ_i. Assim, a situação $d > T_{\text{off}}$ na qual o atraso máximo se sobreporia a um novo surto é evitada com a restrição $c > \lambda_i$, pois [40]

$$(\hat{\lambda}_i - c_i)T_{\text{on}} < (\hat{\lambda}_i - \lambda_i)T_{\text{on}} = \frac{\hat{\lambda}_i T_{\text{on}} T_{\text{off}}}{T_{\text{on}} + T_{\text{off}}} < c_i T_{\text{off}} \tag{2.29}$$

Este modelo de atraso determinístico é conservador, uma vez que decorre de uma análise de pior caso que prevê colisão dos momentos de transmissão na taxa de pico das fontes multiplexadas.

2.3 Modelos de dois parâmetros

Aqui são apresentados modelos que se baseiam na taxa de pico e na taxa média, sem considerar o tamanho das rajadas (surto) ou outro parâmetro qualquer.

2.3.1 Modelo de Ritter

Um modelo empírico de dois parâmetros derivado por Ritter et al. [42] é proposto nos relatórios do projeto COST 242 [43]. É mais uma variante de modelo de perdas.

Dada uma fonte geradora de tráfego com taxa de pico $\hat{\lambda}$ e taxa média λ, e especificada uma probabilidade de perda devido a congestionamento ε num enlace de capacidade C, a banda efetiva c devido a uma fonte é calculada pelas equações

$$c = \begin{cases} f\lambda(1 + 3g(1 - \xi^{-1})) \; ; & 3g \leq \min\{3, \xi\} \\ f\lambda(1 + 3g^2(1 - \xi^{-1})) \; ; & 3 < 3g \leq \xi \\ f\hat{\lambda} \; ; & \text{caso contrário} \end{cases} \tag{2.30}$$

onde

$$\xi = \frac{\hat{\lambda}}{\lambda} \; ; \quad f = 1 + \frac{2\tilde{\Delta}}{100} \; ; \quad g = 2\tilde{\Delta}\frac{\hat{\lambda}}{C} \; ; \quad \tilde{\Delta} = -\log_{10}(\varepsilon) \tag{2.31}$$

Este modelo não apresenta dependência explícita do tamanho do *buffer*, embora tenha um escopo de validade limitado aos sistemas de *buffer* de grande capacidade. Os resultados são confiáveis para enlaces cuja taxa de transmissão é muito maior que as taxas de pico individuais das fontes ($C/\hat{\lambda} > 15$) mas sem que ultrapasse o valor ($C/\hat{\lambda} < 1000$). O parâmetro de pico deve ser tal que $\xi < 20$. Ainda assim, uma precisão razoável só ocorre quando o valor final c é menor que $\hat{\lambda}/2$. O fato da capacidade de enlace C ser uma das variáveis de entrada do modelo torna o uso inconveniente para dimensionamento, uma vez que métodos recursivos seriam necessários.

2.3.2 Modelo de Neame-Zukerman

Neame e Zukerman propuseram um modelo de dois parâmetros em um trabalho recente [38], visando à máxima simplificação do cálculo de banda alocada.

Seja um enlace que transporta N fontes de tráfego independentes do tipo *on-off* com tempos T_{on} e T_{off} arbitrários, implicando em uma taxa de pico $\hat{\lambda}$ e uma taxa média λ para cada fonte. O tráfego agregado oferecido ao enlace pode ser estimado por uma média $m = N\lambda$ e uma variância $\sigma^2 = N\lambda(\hat{\lambda} - \lambda)$. Para um número elevado de fontes N, o teorema do limite central permite afirmar que o comportamento do sistema converge para um processo gaussiano, o que justifica o uso da regra de dimensionamento tal que $C = m + \theta\sigma$. A banda total requerida vale [38]

$$C = \min\left\{ N\lambda + \theta\sqrt{N\lambda(\hat{\lambda} - \lambda)} \; ; \; N\hat{\lambda} \right\} \tag{2.32}$$

onde θ é um parâmetro arbitrário. O sucesso de tal método depende da escolha adequada de θ. Para distribuições gaussianas esta constante pode ser obtida analiticamente conforme [38] e [41]; já para processos não-gaussianos pode ser necessário o uso de simulações para tráfegos em sistemas reais [38]. O valor da probabilidade de perda máxima admitida ε, bem como o tamanho de *buffer*, devem ser levados em consideração no cálculo de θ. A Tabela 2.1 apresenta alguns valores típicos obtidos em [38] para $\varepsilon = 10^{-4}$ e número de fontes no intervalo $10 < N < 100$.

Este método se baseia em certos pressupostos, tais como o número de fontes elevado, com distribuições probabilísticas homogêneas. Também é necessário uma constante θ a ser determinada. Isso faz com que seu uso seja preferencial para políticas de CAC, onde a rapidez de cálculo pode ser atingida pelo armazenamento de *lookup tables* para valores

Tabela 2.1: Estimativas de θ do modelo Neame-Zukerman

Tamanho de *buffer*	Valor de θ
10	3.4
100	2.5
1000	1.8

previamente calculados da banda efetiva em função de N. É, todavia, um método menos eficaz para ser usado como algoritmo de dimensionamento de redes. Segundo os autores, os valores produzidos pela equação (2.32) são estimativas conservadoras de banda.

2.4 Modelos de um parâmetro

A aproximação poissoniana requer apenas um parâmetro, a taxa média. Os sistemas baseados em filas M/M/1 ou M/D/1 são aproximações freqüentemente utilizadas, devido à facilidade de tratamento analítico. Diversos resultados são obtidos nesta seção e uma contribuição na análise destes modelos é apresentada em seguida.

2.4.1 Dimensionamento por restrição de perdas

Baseando-se num modelo de restrição devido a perdas, alguns métodos de análise e dimensionamento de banda efetiva em sistemas com chegadas poissonianas conduzem a resultados fechados e de tratamento simplificado.

Chegadas poissonianas – filas M/M/1

Seja uma fila de comprimento infinito com processo de chegada poissoniano de taxa média λ, atendida por um único servidor de taxa μ com tempo de serviço exponencial negativo. A utilização média vale $\rho = \lambda/\mu$. Da resolução das equações de equilíbrio markovianas quando $\rho < 1$, calcula-se a probabilidade p_n de que existam n elementos no *buffer*, dada por [26]

$$p_n = (1-\rho)\ \rho^n; \quad n = 0, 1, \ldots \tag{2.33}$$

A probabilidade de que o número de elementos no sistema alcance ou ultrapasse um certo limiar b é dado por

$$Prob\{Q \geq b\} = 1 - \sum_{n=0}^{b-1} p_n = \rho^b = \left(\frac{\lambda}{\mu}\right)^b \tag{2.34}$$

A determinação da banda efetiva pode ser obtida através da resolução da inequação (2.1) com o valor da probabilidade de congestionamento do *buffer* dada pela equação (2.34). Para tanto, reescreve-se a relação (2.1) com o valor de máxima perda aceitável ε como uma função exponencial do limiar b no *buffer*

$$Prob\{Q \geq b\} \leq \varepsilon = e^{-\delta b} = e^{\Delta} \tag{2.35}$$

e portanto $\delta = -\log(\varepsilon)/b$ e $\Delta = -\log(\varepsilon)$. Combinando-se as relações (2.35) e (2.34) obtém-se

$$\left(\frac{\lambda}{\mu}\right)^b \leq e^{-\delta b} \quad \Longrightarrow \quad \mu \geq \lambda e^{\delta} \tag{2.36}$$

e adotando-se o valor da banda efetiva c como o *menor* valor de μ que satisfaça a relação (2.36) resulta em

$$c = \lambda e^{\delta} \tag{2.37}$$

A Figura 2.3 ilustra o comportamento do valor da banda normalizada c/λ em função do limiar b variando de 10 a 20, parametrizado para diferentes limites de perdas ε. A assíntota para todas as curvas na figura é $c = \lambda$, pois

$$\lim_{b \to +\infty} \lambda e^{\Delta/b} = \lambda \tag{2.38}$$

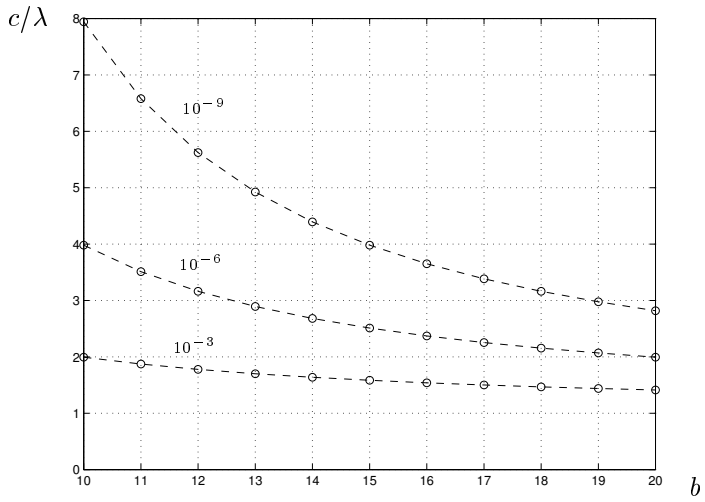

Figura 2.3: Valor da banda efetiva M/M/1 em função do tamanho de buffer para diferentes ε

Ao considerar-se no desenvolvimento acima uma fila M/M/1 (infinita) incorre-se em um erro com relação ao caso de uma fila finita M/M/1/b com tamanho máximo de fila b.

A probabilidade de estado p_n numa fila M/M/1/b para que o sistema contenha n elementos é dada por [26]

$$p_n = \begin{cases} \dfrac{(1-\rho)\rho^n}{1-\rho^b}; & \rho \neq 1 \\ \dfrac{1}{b}; & \rho = 1 \end{cases} \tag{2.39}$$

O congestionamento ocorre quando uma chegada encontra o sistema com a sua capacidade máxima b preenchida, igual ao tamanho do *buffer* mais o elemento sendo servido

$$Prob\{Q = b\} = \frac{(1-\rho)\rho^b}{1-\rho^b} \tag{2.40}$$

Uma comparação entre o uso de filas infinitas com limiar b e filas finitas com tamanho b pode ser feita para os casos M/M/1 e M/M/1/b, a fim de verificar a qualidade da

aproximação. Neste processo, a banda efetiva é dimensionada usando-se a aproximação M/M/1 da equação (2.37), utilizando ε e limiar b especificados. Estes valores assim obtidos de banda são usados como a taxa de transmissão c de uma fila M/M/1/b a fim de verificar qual a discrepância da perda *observada* face ao valor ε originalmente especificado. Para tanto, a equação (2.40) é utilizada fazendo $\rho = \lambda/c$, isto é

$$Prob\{Q = b\} = \frac{(1 - e^{-\delta})e^{-b\delta}}{1 - e^{-b\delta}} \tag{2.41}$$

Na Tabela 2.2 é apresentado o valor c/λ calculado (banda normalizada) para uma fila infinita equivalente, e a relação entre a perda em uma fila finita M/M/1/b (que utilize esta mesma banda) e o valor especificado ε, para diversos tamanhos de buffer b.

Tabela 2.2: Diferenças na perda da fila M/M/1/b

Limite ε especificado		10^{-3}	10^{-4}	10^{-5}	10^{-6}	10^{-7}	10^{-8}
Banda normalizada (fila infinita)	$b = 10$	1.995	2.512	3.162	3.981	5.012	6.310
	$b = 20$	1.412	1.584	1.778	1.995	2.239	2.512
	$b = 30$	1.259	1.395	1.468	1.585	1.711	1.848
Perda/ε (discrepância)	$b = 10$	0.499	0.602	0.684	0.749	0.808	0.842
	$b = 20$	0.499	0.602	0.684	0.749	0.808	0.842
	$b = 30$	0.499	0.602	0.684	0.749	0.808	0.842

Note que em todas as circunstâncias o valor da perda obtido para uma fila finita é sempre *menor* que o especificado (Perda/$\varepsilon < 1$), independentemente do tamanho b adotado. Portanto, a aproximação M/M/1 é aceitável como uma boa substituição à M/M/1/b nesta faixa analisada de valores de ε, produzindo resultados um pouco mais conservadores.

Chegadas poissonianas – fila M/D/1

Em certas situações a fila M/D/1 oferece um modelo mais adequado para a estimativa de banda efetiva. Isto é válido no caso de sistemas ATM, onde o tamanho fixo de célula implica em um tempo também fixo de atendimento no servidor, sendo portanto um processo de variância zero. O tempo de serviço determinístico implica num melhor desempenho quanto ao tempo médio de permanência na fila quando comparado com um servidor M/M/1 de mesma taxa média [26].

Para uma dada fila de comprimento infinito, com processo de chegada poissoniano de taxa λ atendida por um servidor com tempo de serviço constante $1/\mu$, a taxa de utilização é dada por $\rho = \lambda/\mu$. Utilizando um modelamento generalizado para filas M/G/1, a função geradora de probabilidades $\Pi(z)$ é dada por [26]

$$\Pi(z) = \frac{(1 - \rho)(1 - z)K(z)}{K(z) - z}; \qquad K(z) = \sum_{i=0}^{\infty} l_i z^i \tag{2.42}$$

onde l_n é a probabilidade de que ocorram n chegadas durante o tempo de atendimento $S = t$. No caso particular da fila M/D/1, a função geradora dada pela equação (2.42) toma a seguinte forma

$$\Pi(z) = \frac{(1 - \rho)(1 - z)}{1 - ze^{\rho(1-z)}} \tag{2.43}$$

A determinação das probabilidades de estado do sistema p_n pode ser feita por meio da expansão da equação (2.43) em série de Taylor em torno do ponto $z = 0$, conforme sugerido em [45], resultando em

$$\begin{aligned} p_0 &= (1-\rho) \\ p_1 &= (1-\rho)(e^{\rho}-1) \\ &\vdots \\ p_n &= (1-\rho)\sum_{j=1}^{n}(-1)^{n-j}e^{j\rho}\left[\frac{(j\rho)^{n-j}}{(n-j)!}+\frac{(j\rho)^{n-j-1}}{(n-j-1)!}\right] \qquad (n \geq 2) \end{aligned} \tag{2.44}$$

onde a segunda parcela da soma entre colchetes para a probabilidade p_n deve ser ignorada sempre que $j = n$. Desta forma, a probabilidade de que o tamanho da fila exceda o limiar b é dada por

$$Prob\{Q \geq b\} = 1 - \sum_{n=0}^{b-1} p_n \tag{2.45}$$

e a banda efetiva para este sistema é determinada de maneira análoga à da fila M/M/1.

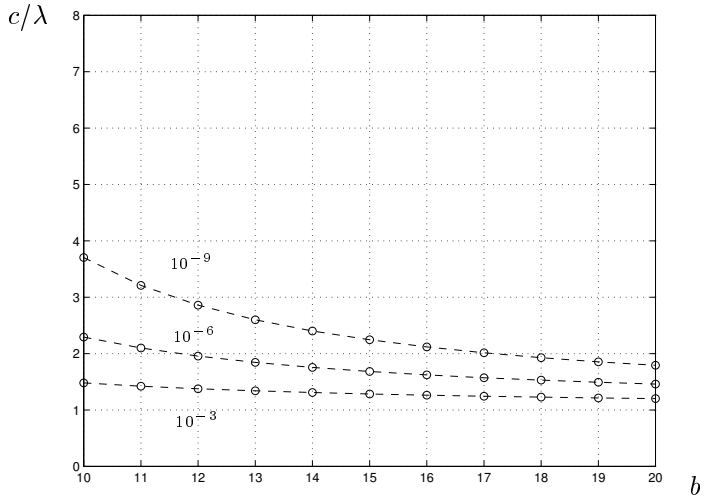

Figura 2.4: Valor da banda efetiva M/D/1 em função do tamanho de buffer para diferentes ε

A complexidade algébrica resultante da combinação das equações (2.44) e (2.45) impede a solução analítica em μ, sendo necessário recorrer a métodos numéricos (inversão por bissecção). A Figura 2.4 ilustra o comportamento de c/λ, parametrizado para diferentes valores do limite de perdas ε. Em comparação com a fila M/M/1, a banda efetiva calculada pela M/D/1 é menor, sendo este efeito mais pronunciado para valores pequenos ε e b.

Ganho de multiplexação em filas M/M/1

Considere N fontes distintas, com processos de chegada poissonianos independentes de taxa de chegada λ_i, onde $i = 1, \ldots, N$. Suponha que a QoS requerida devido a perdas para cada uma delas seja a mesma ε.

Na Figura 2.5 duas situações distintas são consideradas e uma relação de equivalência entre ambas é deduzida.

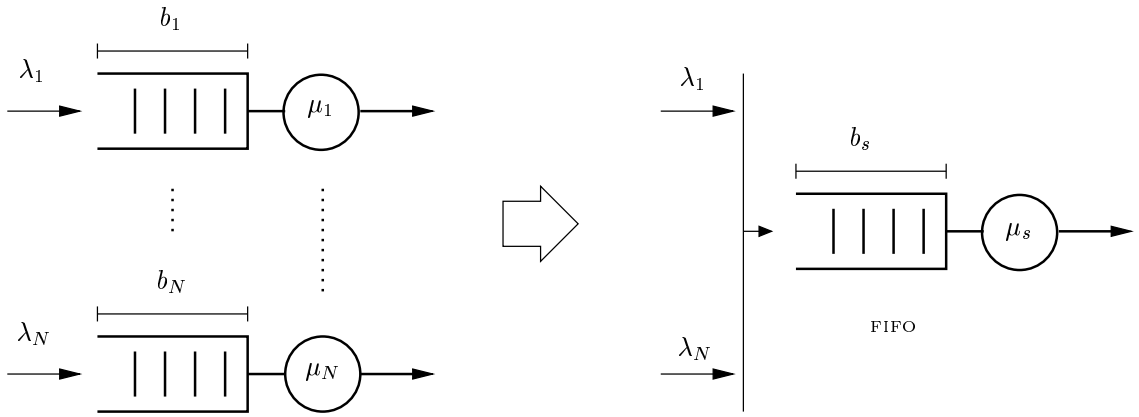

Figura 2.5: *Buffers* distintos versus *buffer* único para fontes independentes

Na primeira situação, cada uma das fontes é apresentada a uma fila exclusiva, independente das demais, com limiar b_i e taxa de serviço μ_i. A relação (2.36) produz

$$\mu_i \geq \lambda_i e^{\delta_i} \tag{2.46}$$

com $\delta_i = -\log(\varepsilon)/b_i$. Portanto

$$c_i = \lambda_i e^{\delta_i} \tag{2.47}$$

é o valor individual para cada banda efetiva, com $i = 1, \ldots, N$.

Na segunda situação, as N fontes independentes são oferecidas a uma única fila com disciplina de atendimento FIFO, cujo limiar é b_s e a taxa de serviço, μ_s.

Aplicando a relação (2.36) ao conjunto de fontes resulta em

$$\left(\frac{1}{\mu_s}\sum_{i=1}^{N}\lambda_i\right)^{b_s} \leq e^{-\delta_s b_s} \tag{2.48}$$

onde

$$\delta_s = \frac{-\log(\varepsilon)}{b_s} = \frac{\Delta}{b_s} \tag{2.49}$$

Isolando μ_s resulta

$$\mu_s \geq \sum_{i=1}^{N}\lambda_i e^{\delta_s} \tag{2.50}$$

A banda efetiva agregada do conjunto de fontes é igual a

$$C = e^{\delta_s}\sum_{i=1}^{N}\lambda_i \tag{2.51}$$

A comparação das equações (2.47) e (2.51) pode ser feita de duas formas distintas.

- ***Caso*** $b_s = b_1 + \cdots + b_N$

Quando o tamanho do *buffer* comum b_s é igual a soma dos tamanhos dos *buffers* individuais, da equação (2.49) tem-se

$$\delta_s = \frac{\Delta}{\sum_{i=1}^{N} b_i} \tag{2.52}$$

A susbstituição na equação (2.51) resulta em

$$C = \sum_{i=1}^{N} \lambda_i \exp\left(\left[\sum_{i=1}^{N} b_i\right]^{-1} \Delta\right) \tag{2.53}$$

e o somatório das bandas efetivas individuais, dados pela equação (2.47), implica

$$\sum_{i=1}^{N} c_i = \sum_{i=1}^{N} \lambda_i \exp\left(\frac{\Delta}{b_i}\right) \geq C \tag{2.54}$$

O *buffer* compartilhado b_s implica num *ganho de banda* na determinação da banda efetiva (pois $C \leq c_1 + c_2 + \ldots + c_N$), conforme se observa na equação (2.54). Quando a mesma quantidade de *buffer* é empregada no sistema, a banda efetiva resultante é *menor* que a soma das bandas efetivas individuais.

- ***Caso*** $b_s = b_1 = \cdots = b_N$

Quando o tamanho de *buffer* é o mesmo para todas as filas, da equação (2.49) vem

$$\delta_s = \frac{\Delta}{b} \tag{2.55}$$

e a equação (2.51) pode ser reescrita como

$$C = \sum_{i=1}^{N} \lambda_i \exp\left(\frac{\Delta}{b}\right) \tag{2.56}$$

e o somatório das bandas efetivas dadas pela equação (2.47) resulta em

$$\sum_{i=1}^{N} c_i = \sum_{i=1}^{N} \lambda_i \exp\left(\frac{\Delta}{b}\right) = C \tag{2.57}$$

A propriedade de aditividade das bandas efetivas ($C = c_1 + c_2 + \ldots + c_N$) se preserva neste caso. O ganho de multiplexação agora assume a forma de um *ganho de buffer*, isto é, economia da quantidade total de *buffer* comum utilizado, que é N vezes menor do que no caso anterior.

2.4.2 Dimensionamento conjunto

Tomando-se *independentemente* as restrições dadas pelas relações (2.1) e (2.2), a banda efetiva é determinada pela restrição mais exigente

$$c = \max\{c_D; c_L\} \tag{2.58}$$

onde c_D é a banda devido ao atraso e c_L a banda devido à perda.

Em sistemas de um único tráfego oferecido a uma fila M/M/1, a equação (2.37) determina o valor de c_L. A banda efetiva devido ao atraso é obtida a partir da equação (2.2) e portanto

$$c_L = \lambda e^{\delta} \tag{2.59}$$
$$c_D = b/d \tag{2.60}$$

com d sendo o valor do máximo atraso admissível.

No caso de sistemas compostos por múltiplos tráfegos poissonianos *iguais e independentes*, oferecidos a uma fila comum M/M/1 (disciplina FIFO), com um *buffer* de limiar b (invariante com o número de fontes) conforme ilustrado na Figura 2.5, c_L deriva da equação (2.51). Nesta situação a aditividade se mantém e a banda agregada vale

$$C = \sum_{i=1}^{N} c_i = Nc \tag{2.61}$$

o que implica que tanto o atraso devido ao armazenamento no *buffer* quanto a banda agregada devido a perdas são N vezes menores. Assim

$$c_L = \frac{1}{N} \sum_{i=1}^{N} \lambda_i e^{\delta} \tag{2.62}$$
$$c_D = \frac{b}{Nd} \tag{2.63}$$

Observe que o ganho de multiplexação se manifesta na diminuição da banda c_D, inversamente proporcional ao número das fontes presentes no sistema.

Neste tipo de dimensionamento, o tamanho de *buffer* é um dado fixo, restando a banda efetiva para ser determinada. Com um *buffer* de limiar b e uma probabilidade ε de congestionamento, a taxa agregada C para a soma de N fontes iguais e independentes é

$$C = \sum_{i=1}^{N} c_i = N\lambda e^{\delta} \tag{2.64}$$

pois as propriedades de independência e aditividade se mantêm. Do ponto de vista do atraso d no *buffer* as N fontes multiplexadas experimentam um melhor desempenho. Neste caso atende-se o congestionamento e o atraso se beneficia do ganho de multiplexação.

A Figura 2.6 mostra a utilização da fila $\rho = N\lambda/C$, o máximo atraso observado $\tilde{d}$ no *buffer*, a banda efetiva normalizada c/λ e o tamanho de fila b (dado), todos em função do aumento do número N de fluxos de tráfego multiplexados [39]. O limite da probabilidade

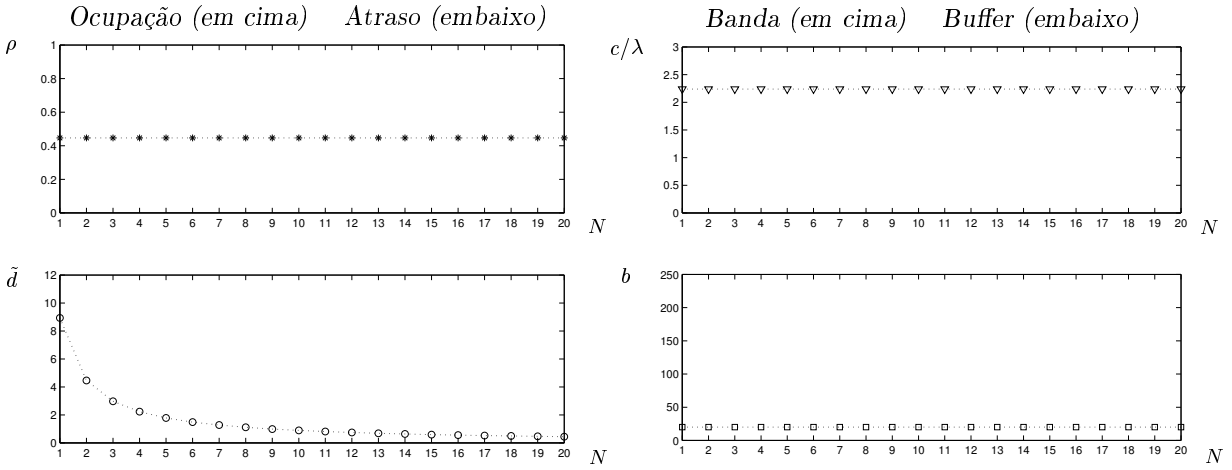

Figura 2.6: Variação do desempenho com N – Dimensionamento conjunto

de congestionamento do *buffer* vale $\varepsilon = 10^{-7}$, a taxa de chegada de cada tráfego individual vale λ e o máximo atraso admissível, $d = 10$ms. O tamanho do *buffer* dado é $b = 20$.

A soma das bandas efetivas calculada devido a perdas para qualquer número de fontes idênticas varia *linearmente* com N e vale $C \approx 2.24N\lambda$, pois não há ganho na utilização da fila e $\rho \approx 0.45$ é constante. Por sua vez, o máximo atraso verificado $\tilde{d} \leq d$ diminui com o aumento de N.

2.4.3 Dimensionamento combinado

É possível *combinar* as duas restrições dadas pelas relações (2.1) e (2.2) para o dimensionamento de sistemas de filas M/M/1 de modo a aproveitar o ganho de multiplexação de maneira mais conveniente. Tanto a banda efetiva c quanto o tamanho da fila b são valores a determinar.

O objetivo é dimensionar a banda pelo atraso máximo e utilizar o ganho de multiplexação para a diminuição da banda efetiva conforme o aumento do número N. Para cada uma das fontes idênticas com taxa de chegada λ e probabilidade máxima de congestionamento ε, a variável b pode ser eliminada das equações (2.62) e (2.63) resultando em

$$\left(\frac{N\lambda}{C}\right)^{Cd} = \varepsilon \tag{2.65}$$

onde d dado é o valor máximo admissível do atraso no *buffer*. A resolução da equação (2.65) em função do número de fontes N produz

$$C = \frac{\Delta}{\mathcal{W}\left(\frac{\Delta}{N\lambda d}\right) d} \tag{2.66}$$

onde $\Delta = -\log(\varepsilon)$ e $\mathcal{W}(y)$ é a função *W de Lambert* [18], definida como sendo a solução de $y = xe^x$. O tamanho do *buffer* b é obtido pela substituição da equação (2.66) em (2.2)

$$b = \left\lceil \frac{\Delta}{\mathcal{W}\left(\frac{\Delta}{N\lambda d}\right)} \right\rceil \tag{2.67}$$

com $\lceil \cdot \rceil$ indicando o máximo inteiro.

A Figura 2.7 mostra a utilização da fila ρ, o máximo atraso d no *buffer*, a banda efetiva normalizada c/λ e o tamanho de fila b, todos em função de N, num exemplo de [39]. A probabilidade de congestionamento do *buffer* vale $\varepsilon = 10^{-7}$, com a taxa de chegada de cada tráfego λ e máximo atraso admissível $d = 10$ms.

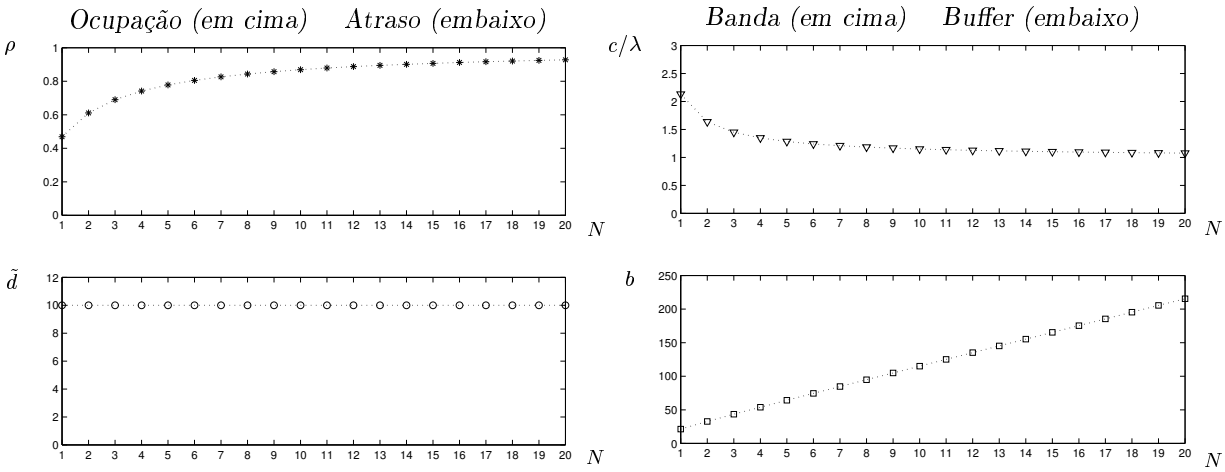

Figura 2.7: Variação do desempenho com N – Dimensionamento combinado

A utilização da fila $\rho(N)$ varia com o número de fontes, aumentando sensivelmente quando comparado ao caso da Figura 2.6, pois a banda efetiva agregada cresce mais lentamente que o caso do dimensionamento conjunto

$$C = \frac{1}{\rho(N)} N\lambda \tag{2.68}$$

O máximo atraso verificado $\tilde{d}$ permanece constante e idêntico a d qualquer que seja o número de fontes.

2.4.4 Momento logarítmico com parâmetro único

O modelo de momento logarítmico para dimensionamento por perda pode ser utilizado para fontes de um único parâmetro, do tipo poissoniano, com taxa média λ. Estas fontes têm a distribuição de probabilidades

$$Prob\{\mathcal{A}(t) = n\} = \frac{(\lambda t)^n}{n!} e^{-\lambda t} \tag{2.69}$$

e utilizando a equação (2.6) vem

$$E\{\exp[\mathcal{A}(t)\delta]\} = \sum_{n=0}^{+\infty} e^{n\delta} \frac{(\lambda t)^n}{n!} e^{-\lambda t} = \exp[\lambda t(e^{\delta} - 1)] \tag{2.70}$$

e a respectiva banda efetiva vale

$$c = \lambda \frac{(e^{\delta} - 1)}{\delta} \tag{2.71}$$

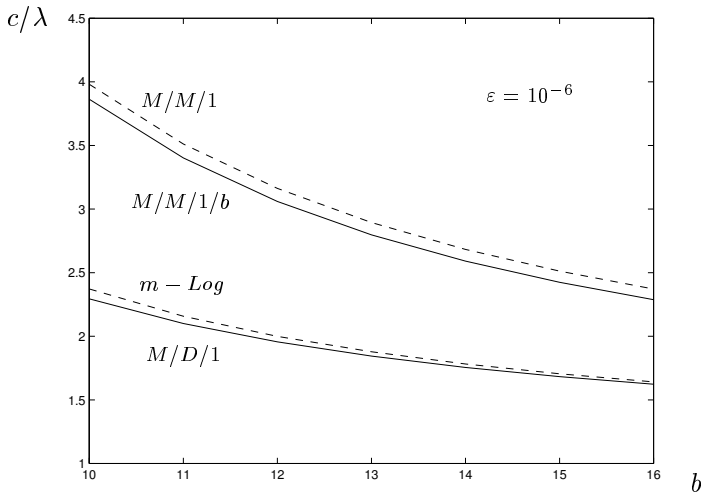

Figura 2.8: Comparação de modelos de um parâmetro

O caso particular de uma fonte determinística de taxa constante λ é calculada de maneira similar, com o número de elementos gerados no intervalo t sendo $\mathcal{A}(t) = \lambda t$. A banda efetiva é igual à própria taxa λ.

Os modelos de perdas para chegadas poissonianas, tanto para tempos de serviço variável (M/M/1 e M/M/1/b) quanto para tempos de serviço constante (M/D/1 e Momento Logarítmico) são comparados na Figura 2.8 em função de b, para $\varepsilon = 10^{-6}$. Observa-se que para modelos com tempo de serviço aleatório o M/M/1 é uma aproximação mais conservadora que o M/M/1/b, enquanto que nos modelos com tempo de serviço determinístico o Momento Logarítmico é um pouco mais conservador que o M/D/1.

2.5 Estimativas de buffer

Muitas vezes o tamanho de *buffer* é um parâmetro a ser especificado durante o processo de dimensionamento do nó de transmissão. O aspecto econômico é sempre importante, pois maiores tamanhos de *buffer* implicam em maiores custos. Critérios adicionais são sugeridos no desenvolvimento a seguir e podem fornecer outras estimativas adequadas.

2.5.1 Sistemas de um parâmetro

No caso de sistemas de um parâmetro, poissoniano, o valor adotado é a situação de melhor compromisso entre possíveis combinações entre a banda efetiva alocada c e tamanho de fila b, de modo a respeitar também a condição de atraso máximo admissível d especificada. Utiliza-se um processo de dimensionamento conjunto, e qualquer escolha do par (c, b) que satisfaça as relações (2.60) é aceitável. Um dos critérios de projeto é adotar o tamanho de *buffer* correspondente aos menores valores de banda efetiva. Seja o exemplo ilustrado na Figura 2.9. Para perdas de $\varepsilon = 10^{-3}$ e atraso máximo $d = 160$ms, um sistema que receba pacotes a uma taxa de $\lambda = 40$pps terá um tamanho desejável de fila para a vizinhança $11 \leq b \leq 12$. Note que outros valores de *buffer* também seriam possíveis, desde que acompanhados de uma aumento da banda. O ponto assinalado na figura corresponde ao mínimo do dimensionamento conjunto e fornece a faixa de escolhas adequadas para b.

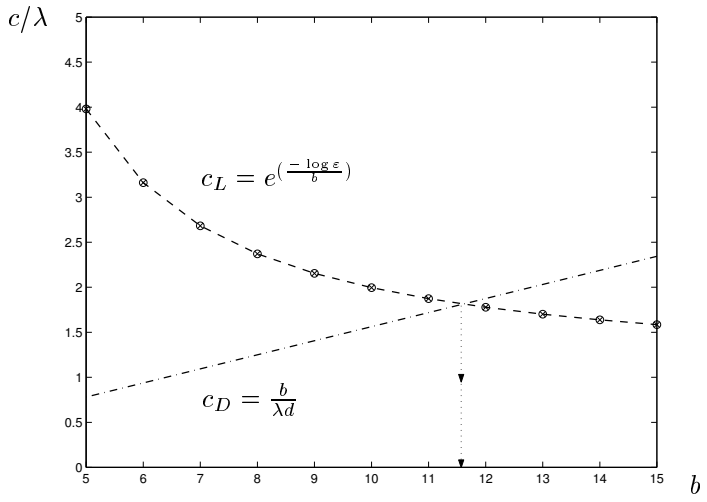

Figura 2.9: Escolha de buffer para modelo de um parâmetro – Dimensionamento conjunto

Sistemas poissonianos que utilizem o dimensionamento combinado têm o valor de b calculado diretamente pelo procedimento, por intermédio da equação (2.67), não sendo necessário arbitrar valores.

2.5.2 Sistemas de múltiplos parâmetros

Sistemas definidos por vários parâmetros (como por exemplo PCR, SCR e MBS) possuem maior liberdade de escolha dos valores de b. Da mesma forma que no caso poissoniano, é possível utilizar a restrição de atraso máximo admissível da relação de perdas (2.2) para obter um limitante superior para os valores possíveis de b.

Considerando apenas as perdas é possível inferir uma estimativa razoável para o tamanho de *buffer* baseado no tamanho máximo de surto MBS. Seja um sistema descrito segundo o modelo de fonte fluido markoviano, tipo *on-off*, com os parâmetros tempo médio de permanência no estado ativo T_{on}, tempo médio de permanência no estado inativo T_{off} e taxa de transmissão no estado ativo $\hat{\lambda}$. Estes parâmetros podem ser especificados em função do PCR, SCR e MBS como

$$\hat{\lambda} = \text{PCR}; \qquad T_{\text{on}} = \frac{\text{MBS}}{\text{PCR}} \qquad T_{\text{off}} = \text{MBS}\left(\frac{1}{\text{SCR}} - \frac{1}{\text{PCR}}\right) \tag{2.72}$$

No exemplo mostrado na Figura 2.10, um sistema de fluido markoviano é especificado para $\varepsilon = 10^{-3}$. A banda efetiva é calculada pelo método do momento logarítmico, equações (2.18), (2.19) e (2.20). A relação de pico $\xi = \text{PCR}/\text{SCR}$ igual a 4 é utilizada, e o cálculo da banda efetiva é normalizado pela taxa mádia SCR. Para uma faixa de variação do parâmetro de surto $10 \leq \text{MBS} \leq 20$ uma família de curvas é plotada para a banda c/SCR em função do tamanho de *buffer*. Observa-se que o lugar geométrico dos pontos das curvas em que $b = \text{MBS}$, marcados por um círculo no gráfico, formam uma reta horizontal, fixando um valor idêntico de banda efetiva, qualquer que seja b. Isto torna conveniente a escolha de um tamanho de *buffer* b igual ao valor MBS esperado.

Há pelo menos dois argumentos para tal escolha:

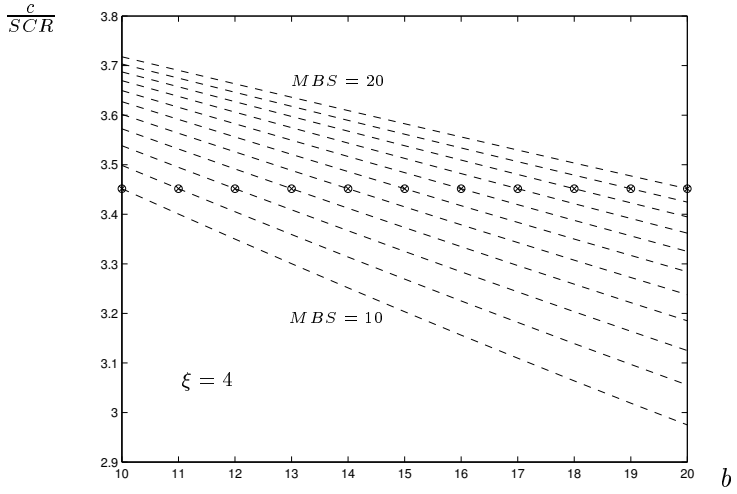

Figura 2.10: Banda efetiva para modelo de três parâmetros

i) O espaço de *buffer* deve ser grande o suficiente para absorver *pelo menos* uma rajada de tamanho máximo, isto é, MBS, a fim de minimizar a probabilidade de congestionamento;

ii) A conversão de modelos de dois parâmetros para três parâmetros fica facilitada. Portanto, o cálculo da banda passa a depender exclusivamente da relação ξ, e a escolha feita de *buffer* permite que sistemas de três parâmetros possam ser tratados como sistemas de dois parâmetros (PCR e SCR) cujo tamanho de *buffer* valha $b = \text{MBS}$. Reciprocamente, sistemas com dois parâmetros e *buffer* a determinar podem ser modelados como sistemas de três parâmetros, onde $\text{MBS} = b$.

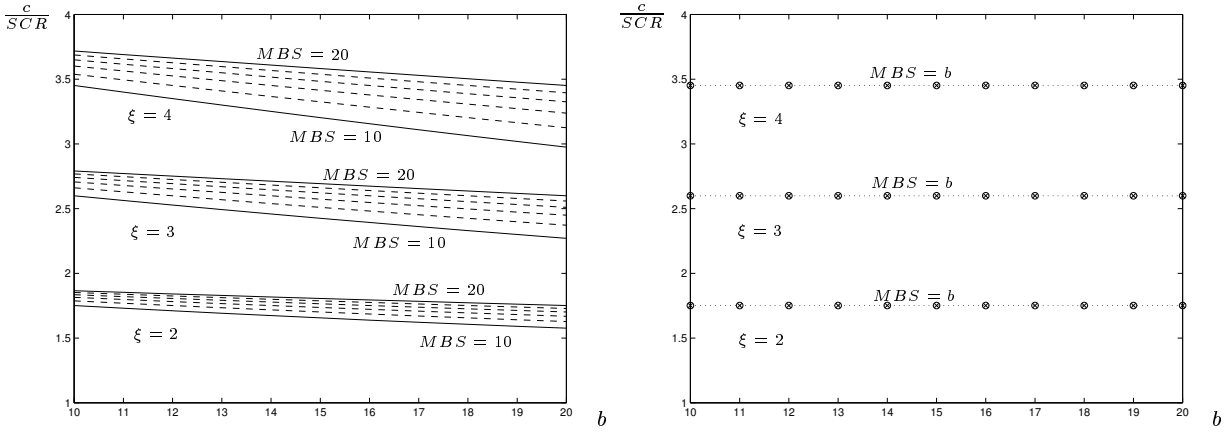

Figura 2.11: Sistema de três parâmetros com MBS livre (esquerda) e com $\text{MBS} = b$ (direita)

A Figura 2.11 mostra a equivalência entre dois sistemas. O gráfico da esquerda é calculado com três parâmetros, com $10 \leq \text{MBS} \leq 20$. Três grupos de curvas são plotados para diferentes relações entre PCR e SCR. O valor da banda efetiva é normalizado como c/SCR, em função da variação de b. No gráfico da direita, adota-se sempre $\text{MBS} = b$ para o cálculo da banda efetiva normalizada c/SCR, em função de b. Também neste caso três curvas são plotadas para relações entre PCR e SCR diferentes. Em ambos os casos, $\varepsilon = 10^{-3}$.

Utilizando o tamanho de fila igual ao tamanho de surto máximo, é possível calcular a Tabela 2.3 onde os valores de banda efetiva normalizada para SCR são dados em função do ε especificado e da relação ξ. Nota-se que para qualquer situação SCR $< c <$ PCR.

Tabela 2.3: Banda Efetiva normalizada para sistemas com $b =$ MBS

Perda ε	$\xi = 2$	$\xi = 3$	$\xi = 4$	$\xi = 5$	$\xi = 6$	$\xi = 7$	$\xi = 8$	$\xi = 9$	$\xi = 10$
10^{-3}	1.75	2.60	3.45	4.30	5.16	6.01	6.87	7.72	8.58
10^{-4}	1.81	2.69	3.58	4.47	5.36	6.26	7.15	8.04	8.93
10^{-5}	1.84	2.75	3.66	4.58	5.49	6.40	7.31	8.23	9.14
10^{-6}	1.87	2.79	3.72	4.64	5.57	6.50	7.43	8.35	9.28
10^{-7}	1.88	2.82	3.76	4.69	5.63	6.57	7.51	8.45	9.38
10^{-8}	1.90	2.84	3.79	4.73	5.68	6.62	7.57	8.51	9.46
10^{-9}	1.91	2.86	3.81	4.76	5.71	6.66	7.62	8.57	9.52

2.6 Simulações

Simulações permitem comprovar alguns dos resultados do cálculo de banda efetiva quando aplicado aos modelos mais simplificados. Também permite, no caso dos modelos mais complexos, testar disciplinas de agendamento mais elaboradas que a FIFO. Todas as simulações de eventos discretos realizadas foram computadas para um número mínimo de eventos da ordem de 10^7, utilizando o software *ARENA*. Vide apêndice para os modelos de simulação.

2.6.1 Filas M/M/1 e M/D/1

Seja as filas M/M/1 e M/D/1 às quais é oferecido o tráfego de pacotes com uma taxa de chegada de $\lambda = 40$ pacotes por segundo (pps). O tamanho da fila b em ambas situações vale 12, com uma perda máxima especificada $\varepsilon = 10^{-3}$. A taxa média de atendimento é exponencial negativa no primeiro caso e constante no segundo, com valor $\mu = c$ igual à banda efetiva requerida em cada caso, obtida respectivamente das Figuras 2.3 e 2.4.

Tabela 2.4: Comparação das bandas efetivas para as filas M/M/1 e M/D/1

Parâmetros	M/M/1	M/D/1
c (pps)	71	55
ρ	0.56	0.73
$E\{Q\}$	0.71	0.96
$E\{W\}$ (ms)	17.8	23.1
L_b	$0.42 \cdot 10^{-3}$	$0.28 \cdot 10^{-3}$

A Tabela 2.4 apresenta os resultados da simulação para taxa de ocupação média do servidor ρ, comprimento médio da fila $E\{Q\}$, tempo médio de espera na fila $E\{W\}$ e probabilidade de congestionamento $L_b = Prob\{Q \geq b\}$. Em ambos os casos o critério de perdas é satisfeito, observando que para o caso da taxa de atendimento determinística a banda efetiva também é menor, permindo validar o modelo de simulação.

2.6.2 Ganho de multiplexação

Para N fontes idênticas oferecidas a um mesmo *buffer* de tamanho $b = 8$ com disciplina de atendimento FIFO, verifica-se que o ganho de multiplexação manifesta-se por meio da diminuição do tempo de permanência no *buffer*, supondo o mesmo limiar de perda por congestionamento $\varepsilon = 10^{-3}$ para todas as fontes. O atraso máximo admissível especificado é de $d = 100$ms.

Tabela 2.5: Simulação para dimensionamento conjunto em fila M/M/1

Parâmetros	$N = 1$	$N = 2$	$N = 3$	$N = 4$	$N = 5$
C (pps)	95	190	284	379	474
c (pps)	95	95	95	95	95
b	8	8	8	8	8
ρ	0.42	0.42	0.42	0.42	0.42
$E\{Q\}$	0.30	0.30	0.30	0.30	0.30
$E\{W\}$ (ms)	7.60	3.80	2.53	1.90	1.52
$L_b/10^{-3}$	0.59	0.58	0.58	0.55	0.60

A Tabela 2.5 apresenta os resultados para uma fila M/M/1, onde cada uma da fontes independentes tem uma taxa de chegada $\lambda = 40$pps. A banda efetiva de cada fonte é calculada pela equação (2.37) e vale $c = 95$pps. Isto implica um atraso *máximo* esperado igual a $\tilde{d} = (84.2/N)$ milissegundos, o que satisfaz o limite de atraso para qualquer número de fontes. A banda agregada C é obtida por meio da equação (2.64) para cada situação com N fontes.

No dimensionamento conjunto o ganho de multiplexação é utilizado de modo a reduzir os requisitos totais de banda efetiva C com o aumento de N. A Tabela 2.6 mostra os resultados numa fila M/M/1 com disciplina FIFO.

Tabela 2.6: Simulação para dimensionamento combinado em fila M/M/1

Parâmetros	$N = 1$	$N = 2$	$N = 3$	$N = 4$	$N = 5$
C (pps)	88	134	177	219	261
c (pps)	88	67	59	55	52
b	9	13	18	22	26
ρ	0.46	0.60	0.68	0.73	0.77
$E\{Q\}$	0.38	0.87	1.41	1.95	2.52
$E\{W\}$ (ms)	9.41	10.9	11.7	12.2	12.6
$L_b/10^{-3}$	0.46	0.48	0.30	0.24	0.25

Os valores da banda efetiva agregada C são calculados por meio da equação (2.66) e os valores de *buffer* correspondentes por $b = Cd$, adotando o inteiro mais próximo. O valor de atraso máximo admissível é $d = 100$ms. O valor do tráfego individual por fonte vale $\lambda = 40$pps. O limiar de perda é adotado como $\varepsilon = 10^{-3}$. A banda efetiva individual c varia conforme o número de fontes e é dada por $c = C/N$.

2.6.3 Disciplinas de agendamento

Diferentes disciplinas de agendamento além da FIFO são utilizadas com frequência em sistemas reais. Uma possibilidade é o uso de prioridades absolutas, com as classes divididas em múltiplos níveis, cada uma com privilégio de atendimento distinto, conforme estudado por Mata em [36]. A outra, é por meio de disciplinas cíclicas.

Neste item, são consideradas as seguintes possibilidades:

- **Round-Robin ou Weighted Round-Robin:** Para cada uma das N fontes ou k classes independentes é alocada uma fila própria de tamanho fixo. As filas são servidas ciclicamente por um único servidor que atende por vez um (RR) ou um número pré-definido (WRR) de elementos de cada fila. Filas vazias são ignoradas e não consomem tempo de servidor.

- **Cíclica exaustiva:** É similar ao caso da *Round-Robin*, diferindo no fato de que são servidos *todos* os elementos que se encontrarem numa fila, até que não reste mais nenhum, antes que o servidor se desloque para a fila seguinte.

- **Weighted Round-Robin com buffer compartilhado:** Trata-se de uma variante do WRR (ou RR) em que uma quantidade fixa de *buffer comum* é compartilhada por todas as fontes. O comprimento virtual das filas individuais é alterado dinamicamente conforme o número de elementos presentes. A restrição é que a somatória dos tamanhos de todas estas filas deve ser menor ou igual à quantidade de *buffer* disponível, além da qual os elementos excedentes são descartados. Esse sistema também é denominado *WRR dinâmico*.

Estas três disciplinas são comparadas na Tabela 2.7 para um número N diferente de fontes, com tráfego individual igual a $\lambda = 40$pps. O tamanho do *buffer* para cada fila é idêntico e vale $b = 12$, para as políticas RR com fila estática e cíclica exaustiva. No caso do RR com fila dinâmica, o *buffer* comum tem tamanho $b_s = 12$. O valor de banda efetiva para cada uma das fontes ou classes é calculado pela equação (2.37) e vale $c = 71$pps. A banda agregada C é dada por Nc.

Tabela 2.7: Agendamentos Round-Robin (2 tipos) e Cíclico exaustivo

		RR estático		RR dinâmico		Exaustivo	
Parâmetros	$N = 1$	$N = 2$	$N = 3$	$N = 2$	$N = 3$	$N = 2$	$N = 3$
C (pps)	71	142	213	142	213	142	213
ρ	0.56	0.56	0.56	0.56	0.56	0.56	0.56
$E\{X\}$	0.71	0.36	0.24	0.36	0.24	0.36	0.24
$E\{W\}$ (ms)	18	9.0	6.0	9.0	6.0	9.0	6.0
$L_b/10^{-3}$	0.4	0.02	0.003	0.6	0.6	0.07	0.007

A exemplo da disciplina FIFO, os três agendamentos exibem um ganho de multiplexação com respeito ao tempo médio de permanência na fila $E\{W\}$, inversamente proporcional ao número N. Nos casos *RR-estático* e *Cíclico exaustivo* observa-se um ganho adicional de multiplexação sob a forma de um desempenho melhor que o especificado quanto à probabilidade de congestionamento L_b. Isto deve-se à redistribuição automática de recursos

ociosos de banda que estas duas disciplinas permitem. Filas vazias não consomem banda no servidor, de modo que a fatia alocada pode ser substancialmente maior que c para as filas não-vazias em certos instantes de tempo. Para uma discussão mais detalhada do fenômeno, vide [31]. Isto não ocorre para o caso *RR-dinâmico*, onde a quantidade total de espaço de armazenamento realocável b_s não aumenta conforme o número de fontes.

A vantagem destes agendamentos sobre o FIFO é que as classes de serviço de maior prioridade, como por exemplo CBR e VBR-rt, podem ser isoladas umas das outras, o que permite que cada uma tenha sua banda efetiva calculada de maneira independente das demais. Além disso, para a disciplina *WRR-dinâmica*, a aditividade se mantém. Nos demais casos, o ganho adicional de multiplexação age no sentido de *melhorar* o desempenho quanto ao congestionamento, garantindo que será superior ao especificado. O cálculo quantitativo deste desempenho melhorado todavia depende do conhecimento exato do comprimento médio das filas – e portanto a correspondente parcela de tempo inativa de cada uma – sendo difícil de se obter para sistemas com fontes heterogêneas. Os tráfegos do tipo UBR são tratados como uma *fonte interferente*, isto é, não tem banda própria alocada num enlace e utilizam o tempo ocioso do servidor para serem transmitidos com uma QoS do tipo *melhor esforço*. Presume-se que este serviço menos prioritário não deva degradar apreciavelmente o serviço das demais classes. A Tabela 2.8 apresenta esta influência para diversas intensidades u do tráfego UBR. As classes prioritárias são três, com taxas de chegada $\lambda = 40$pps para cada uma delas. Os *buffers* individuais (e o comum, no caso RR-dinâmico) têm tamanho $b = 12$ e a banda efetiva agregada vale $C = 213$pps.

Tabela 2.8: Simulação para comportamento sob tráfego interferente UBR

	RR estático			RR dinâmico			Exaustivo		
Parâmetros	$u = 0$	$u = \lambda$	$u \to \infty$	$u = 0$	$u = \lambda$	$u \to \infty$	$u = 0$	$u = \lambda$	$u \to \infty$
ρ	0.56	0.75	1.00	0.56	0.75	1.00	0.56	0.75	1.00
$E\{Q\}$	0.24	0.28	0.33	0.24	0.28	0.33	0.24	0.28	0.33
$E\{W\}$ (ms)	6.0	7.0	8.0	6.0	7.0	8.0	6.0	7.0	8.0
$L_b/10^{-3}$	0.003	0.004	0.004	0.7	0.8	0.8	0.007	0.012	0.012

Verifica-se que os três agendamentos têm comportamento adequado para proteção das fontes prioritárias, mesmo com um tráfego UBR tendendo ao infinito. Esta proteção não seria possível com uma disciplina FIFO. Choudhury et al. [15] argumentam com simulações numéricas que os cálculos de banda efetiva em sistemas FIFO podem conduzir a resultados super ou subestimados para um número muito elevado de fontes agregadas. Fontes cujo fator de pico (razão variância-média ζ) se afaste do comportamento poissoniano ($\zeta = 1$), tanto para mais ($\zeta \gg 1$) quanto para menos ($\zeta \ll 1$) também podem interferir nos cálculos de banda efetiva.

2.6.4 Fontes fluido markoviano

Para a simulação de modelos de três parâmetros o modelo de fonte fluido markoviano é utilizado. Para um sistema com tempo de atendimento constante, a banda efetiva é calculada pelas equações (2.18), (2.19) e (2.20). O valor adotado para a taxa média SCR é de 40pps para todas as situações, e o tamanho de buffer b é tomado como igual ao máximo tamanho de surto MBS = 12 pacotes. A taxa de pico PCR é tomada em relação

a taxa média como ($\xi \cdot \mathrm{SCR}$). Para o cálculo dos tempos T_{on} e T_{off} utilizam-se as equações (2.72). A Tabela 2.9 mostra os resultados da simulação para diferentes relações ξ, para uma perda máxima aceitável de $\varepsilon = 10^{-3}$. Observa-se que em todas as situações a perda permanece dentro do especificado. Verifica-se também que para as maiores relações de pico ξ a utilização da fila ρ diminui sensivelmente, bem como o tempo médio de permanência no *buffer*.

Tabela 2.9: Simulação para fontes fluido markoviano

Parâmetros	$\xi = 2$	$\xi = 4$	$\xi = 6$	$\xi = 8$	$\xi = 10$
c (pps)	70	138	206	275	343
ρ	0.31	0.21	0.16	0.13	0.11
$E\{Q\}$	0.62	0.44	0.33	0.26	0.22
$E\{W\}$ (ms)	28.4	18.5	10.1	7.54	6.07
$L_b/10^{-3}$	0.20	0.45	0.35	0.32	0.42

2.7 Conclusão

Neste capítulo foram apresentados alguns métodos de cálculo de banda efetiva para sistemas em que se conheça um determinado número de parâmetros descritores de tráfego (três, dois ou um).

O conceito de banda efetiva torna-se mais simples e eficaz quando apresenta as propriedades de aditividade e independência, notadamente para o dimensionamento de sistemas que contenham várias fontes com características distintas de tráfego. Na maioria das vezes a quantidade disponível de dados confiáveis sobre a característica dos tráfegos é pequena, o que exige modelos que dependam de um menor número de parâmetros. Desta forma, para os diversos métodos de cálculo analisados, o método genérico do momento logarítmico (Kesidis et al.) é o mais adaptável por se adequar tanto à função que modela a fonte de tráfego quanto ao respectivo número de parâmetros descritores. Outros métodos de três ou dois parâmetros não apresentam vantagens significativas que recomendem o seu uso sobre o método genérico.

Cálculos realizados pelo modelo de um parâmetro são simples de executar usando a aproximação da fila M/M/1 e conduzem a bons resultados para processos puramente poissonianos ou para sistemas em que se conheça apenas a taxa média (SCR). Na comparação das variantes de um parâmetro, as aproximações recomendáveis são a M/M/1 para sistemas com tempo de atendimento variável (aleatório) e a momento logarítmico para sistemas com tempo de atendimento constante (determinístico).

O ganho de multiplexação devido à agregação de fontes estatisticamente independentes pode manifestar-se de duas maneiras. Uma é através do ganho de banda e a outra é através do ganho no tamanho de *buffer* necessário. Esta segunda apresenta como conseqüência indireta a diminuição do tempo de permanência no *buffer*. Especificamente no caso do dimensionamento conjunto (perdas ou atraso), o ganho se apresenta sob a forma de redução do tempo de permanência na fila, sendo difícil de ser explorado.

O uso do método de dimensionamento combinado (perdas e atraso) usa o ganho de multiplexação na forma de redução da banda efetiva agregada com relação à simples soma das bandas individuais, para um grande número de fontes. Em contrapartida, um aumento significativo ocorre no tamanho do *buffer* necessário, o que limita a aplicação deste dimensionamento aos casos em que o custo de banda é maior do que o custo de memória para *buffer*. Neste caso, a disponibilidade limitada de banda (ex. sistemas rádio) é um dos poucos casos que podem tornar o dimensionamento combinado atrativo.

O tamanho de *buffer* pode ser também uma variável a ser dimensionada na síntese de uma rede. A estimativa para sistemas de um parâmetro, poissonianos, pode ser feita para o dimensionamento conjunto para a vizinhança ponto de mínimo da banda efetiva c em função de b. Para o dimensionamento combinado, este tamanho é obtido como parte do processo de cálculo. Sistemas de três parâmetros têm uma escolha conveniente de *buffer* como $b = \text{MBS}$. Este mesmo critério permite que sistemas de dois parâmetros sejam tratados como modelos de três parâmetros, fazendo neste caso $\text{MBS} = b$ caso o *buffer* seja conhecido. Portanto, o mesmo equacionamento de momento logarítmico para fontes tipo fluido markoviano baseado em PCR, SCR e MBS pode ser utilizado para os sistemas de dois parâmetros.

Agendamentos do tipo WRR-estático, WRR-dinâmico e Cíclico exaustivo têm desempenho semelhante quando comparados entre si para situações de tráfego mediano e taxas de perda melhores que 10^{-3}, conforme verificado nas simulações realizadas. A proteção contra tráfegos interferentes é igualmente eficiente nos três agendamentos, o que não seria possível de implementar com FIFO apenas. É interessante observar que o agendamento WRR-dinâmico utiliza o ganho de escala de multiplexação de modo a economizar uma quantidade significativa de *buffer*, quando comparado ao WRR-estático ou ao cíclico exaustivo.

As estimativas dos valores de banda efetiva são bastante acuradas quando o tráfego tem um comportamento próximo ao poissoniano. À medida em que estes tráfegos se afastam desta hipótese, os valores obtidos tendem a ser *superestimados* quando a variância da distribuição do processo de chegada é muito maior que a respectiva média, ou *subestimados* na situação oposta.

Capítulo 3

Bloqueio em Enlaces

O bloqueio como conseqüência do acesso comutado aos recursos de rede é discutido neste capítulo e diferentes estratégias de reserva de recursos são comparadas. Algumas características da interação entre classes de serviço também são consideradas. Finalmente, comparações numéricas entre estratégias são apresentadas.

3.1 Introdução

Existem duas formas de se tratar os circuitos virtuais dos tráfegos de informação que atravessam uma rede. A maneira mais simples é estabelecer uma *conexão permanente*, estática e dedicada, que não compete por recursos com outros tráfegos simultâneos. Os recursos são alocados por um gerente de rede e permanecem invariantes durante toda a operação do sistema. Conexões do tipo PVC pertencem a esta categoria. A outra maneira é por meio do estabelecimento dinâmico das conexões a cada chamada, e posterior desconexão ao término do serviço. Estes sistemas são denominados *sistemas comutados* e conexões do tipo SVC são exemplos desta categoria. A principal vantagem deste modo de operação é liberar mais recursos para outras chamadas que de outra forma permaneceriam ociosos após o término do fluxo de informação. A contrapartida é que agora se introduz uma *probabilidade de bloqueio de acesso* à rede nos sistemas SVC, devido à maneira estocástica com que as requisições de conexão são geradas.

Em sistemas nos quais convivem diversas classes de serviço com requisitos diferentes, as correspondentes probabilidades de bloqueio também apresentam valores diferentes, além de interagirem mutuamente. Uma estratégia simples e eficiente em sub-sistemas (nós ou enlaces) de uma rede multisserviços é a *reserva de recursos*. A reserva faz uso de recursos dedicados dentro da rede e é, em geral, adotada para modificar características do sistema, sejam elas bloqueio, desempenho ou mesmo custo de implementação. Estas *políticas de reserva* podem ser implementadas segundo diferentes estratégias levando em conta fatores diversos, tais como quantidade total de recursos (circuitos ou banda) disponível, quantidade de recursos previamente alocados para uma determinada classe de serviço ou mesmo a quantidade instantânea de recursos tomados por alguma classe.

3.2 Sistemas multiclasses

Define-se um *sistema multiclasses* como aquele no qual uma quantidade C de recursos (circuitos ou unidades de banda) pode ser utilizada simultaneamente por uma população de clientes agrupados em k diferentes *classes de serviço*. Cada uma das classes de serviço é definida pelo conjunto de parâmetros $D_i = [\ a_i, c_i, \zeta_i, \mathcal{N}_i\ ]$, para $i = 1, \ldots, k$, onde

- a_i : tráfego oferecido pela i-ésima classe de serviço em erlangs, definido como λ_i/μ_i, com
  - λ_i : taxa média de ocorrência da distribuição de probabilidade do processo de chegada de requisições de solicitação (chamadas) ao sistema, suposta estacionária, da i-ésima classe de serviço.
  - $1/\mu_i$: média de duração da distribuição de probabilidade do tempo de retenção da i-ésima classe. Consideram-se as distribuições que possuem uma transformada de Laplace racional.
- c_i : quantidade de unidades de recursos requisitadas simultaneamente pelos clientes da i-ésima classe, podendo assumir apenas valores inteiros positivos.
- ζ_i : fator de pico ou *peakedness factor*, definido como sendo a relação entre a variância e a média da distribuição do processo de chegada. Processos poissonianos têm ζ igual a 1.
- $\mathcal{N}_i$: tamanho da população da i-ésima classe, que pode ser tanto finita quanto infinita.

No conjunto universo $\mathcal{U}^k$ dos infinitos estados de sistemas multisserviços de k classes distintas, composto por números naturais, define-se $\Omega \subset \mathcal{U}^k$ como sendo o sub-espaço de estados finito que contém todos os estados *admissíveis* para um dado sistema e que é determinado pela capacidade total C e pela particular política de acesso. A descrição do espaço Ω é definida pelo vetor de estado

$$n = (n_1, n_2, \cdots, n_k) \tag{3.1}$$

onde n_i é o número de clientes da i-ésima classe de serviço utilizando o sistema, tal que

$$0 \leq n_i \leq \mathcal{N}_i \tag{3.2}$$

com $i = 1, \ldots, k$.

Para o total de C recursos disponíveis, qualquer cliente de uma classe i requer simultaneamente c_i unidades para ser atendido. Considera-se então o vetor dos *requisitos de recursos* das k classes como sendo

$$c = (c_1, c_2, \cdots, c_k) \tag{3.3}$$

e cujos valores admissíveis de um dado c_i são inteiros tais que

$$c_i \in \ \{0, 1, 2, \ldots, C\} \tag{3.4}$$

3.3 Limites e reservas

A definição de políticas de acesso impõe limites adicionais ao espaço de estados Ω admissível, além da quantidade máxima de recursos C existentes. Estas restrições adicionais decorrem das *políticas de reserva* de recursos adotadas.

3.3.1 Reserva de recursos

As políticas de reserva são subdivididas nos seguintes grupos:

- A política de *compartilhamento total*, descrita como

$$\Omega = \left\{ 0 \leq \sum_{i=1}^{k} n_i c_i \leq C \right\} \tag{3.5}$$

  cujos clientes podem ter acesso a qualquer um dos C recursos, respeitados os limites físicos de disponibilidade dados pelas restrições das equações (3.2) e (3.4).

- Uma política de acesso do tipo *reserva parcial* toma a forma

$$\Omega = \left\{ 0 \leq \sum_{i=1}^{k} n_i c_i \leq C_0 + \sum_{i=1}^{k} C_i = C \ ; \ 0 \leq n_i c_i \leq C_0 + C_i; \ i = 1, \ldots, k \right\} \tag{3.6}$$

  na qual os C recursos são repartidos entre as i-classes de serviço. Os C_i recursos são destinados exclusivamente ao uso de cada uma delas e os C_0 recursos restantes são compartilhados livremente por todas as classes. Outras adições interessantes à política de acesso parcial podem ser executadas de modo a permitir que os recursos comuns sejam subdivididos em fatias acessíveis a alguns grupos de classes (mas não a outros), ou que algumas classes não tenham acesso algum aos recursos comuns, etc.

- No caso particular em que $C_0 = 0$ tem-se a política de *reserva plena*

$$\Omega = \{ 0 \leq n_i c_i \leq C_i; \quad i = 1, \ldots, k \} \tag{3.7}$$

  na qual cada classe vale-se de um conjunto distinto de recursos sem interferência entre elas.

Os limites das *fronteiras de reserva* R_i de cada uma das i classes (isto é, onde as componentes em i do vetor de estado devem atender à restrição $n_i \leq R_i$) são dados por

$$R_i = \frac{1}{c_i} \cdot \left[C - \sum_{\substack{j=1 \\ j \neq i}}^{k} C_j \right] \tag{3.8}$$

A Figura 3.1 ilustra um exemplo das políticas de compartilhamento total (esquerda) e parcial (direita), para o caso bidimensional ($k = 2$). Cada estado possível é indicado por um pequeno círculo, de coordenadas $(n_1 \ , \ n_2)$. As classes têm requisitos de recursos $c_1 = 2$

e $c_2 = 1$, e o número de recursos disponíveis é $C = 18$; para o sistema com reservas, $C_1 = 6$ e $C_2 = 4$, com limites $R_1 = 7$ e $R_2 = 12$.

O lugar geométrico dos estados permitidos pertencentes a Ω está contido em um politopo no espaço k-dimensional. Nota-se que, com o uso da reserva, as fronteiras do espaço de estados Ω se alteram, reduzindo-o sensivelmente.

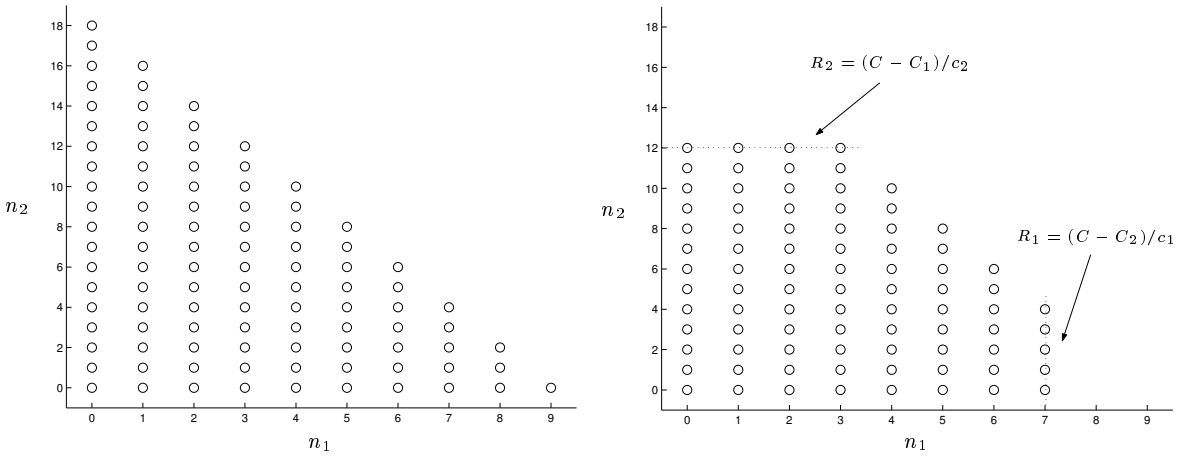

Figura 3.1: Espaço de estados - compartilhamento total e parcial

O *bloqueio* experimentado por uma chegada de um cliente associado a uma determinada classe i pode ser definido pela impossibilidade da transição do vetor de estado n para um outro vetor de estado adjacente n_i^+. O *impedimento* de uma partida é definido pela impossibilidade de transição do estado n_i para um estado adjacente n_i^-. Estas transições possíveis podem ser representadas pela notação

$$\delta_i^+(n) = \begin{cases} 1\ ; & n_i^+ \in \Omega \\ 0\ ; & \textit{caso contrário} \end{cases} \tag{3.9}$$

$$\delta_i^-(n) = \begin{cases} 1\ ; & n_i^- \in \Omega \\ 0\ ; & \textit{caso contrário} \end{cases} \tag{3.10}$$

onde os *vetores de estados adjacentes* n_i^+ e n_i^- definem-se como:

$$\begin{aligned} n_i^+ &= (n_1, \cdots, n_{i-1}, n_i + 1, n_{i+1}, \cdots, n_k) \\ n_i^- &= (n_1, \cdots, n_{i-1}, n_i - 1, n_{i+1}, \cdots, n_k) \end{aligned}$$

Um grupo de políticas de particular interesse é apresentado por Aein e Kosovych em [3], sendo denominadas de políticas *coordenada-convexas*. Uma política Ω é coordenada-convexa se satisfizer

$$\Omega : \ \{ n \in \Omega,\ \forall n_i > 0 \ \Rightarrow \ n_i^- \in \Omega \ \} \tag{3.11}$$

Portanto nas políticas coordenada-convexas as partidas (fim de serviço) nunca são impedidas, isto é, $\delta_i^-(n) = 1$ para todo $n \in \Omega$ tal que $n_i > 0$, $i = 1, \ldots, k$. Além disso, os estados bloqueantes pertencem a alguma fronteira de Ω.

3.3.2 Políticas de reserva e simetria

Para analisar a questão da bidirecionalidade de trasição, considerem-se os dois métodos distintos de política de reserva de recursos:

A *reserva simples* de recursos (estática) do tipo *ao menos* σ_i não leva em conta a quantidade de recursos livres após a admissão de uma requisição da classe i, pois garante-se apenas que uma quantidade máxima de σ_i recursos não será tomada por qualquer outra classe que não a i-ésima. Este tipo de reserva não lida com grupos de recursos atribuídos fisicamente *a priori* para qualquer classe reservada (também conhecido como *marcação*), e somente a *quantidade* requisitada σ_i dos mesmos é garantida.

A *reserva por limiar* θ_i, impõe uma maior estruturação ao conjunto Ω. Uma requisição de serviço de uma i-ésima classe ingressante no sistema só é admitida se após a aceitação restarem *no mínimo* θ_i recursos desocupados. O número θ_i é chamando de *limiar de admissão* para a i-ésima classe. Uma dada classe não sofre restrição adicional de acesso ao sistema se seu valor de limiar de admissão for $\theta_i = 0$.

Esta diferenciação do tipo de reserva, embora sutil, é de vital importância para a aplicabilidade de alguns métodos de cálculo empregados, uma vez que a simetria entre as transições de estado poderá ou não ser preservada. Por simetria entende-se a existência da possibilidade de transição bidirecional entre dois estados n quaisquer tal que

$$n \in \Omega : \begin{cases} se \quad \delta_i^+(n) = 1 \quad \Rightarrow \quad \delta_i^-(n_i^+) = 1 \\ \\ se \quad \delta_i^-(n) = 1 \quad \Rightarrow \quad \delta_i^+(n_i^-) = 1 \end{cases} \tag{3.12}$$

Reservas do tipo *ao menos* σ_i preservam a simetria para todos os estados, enquanto as reservas do tipo *por limiar* θ_i destroem esta simetria para alguns dos pares de estados $n = [n_1, n_2]$.

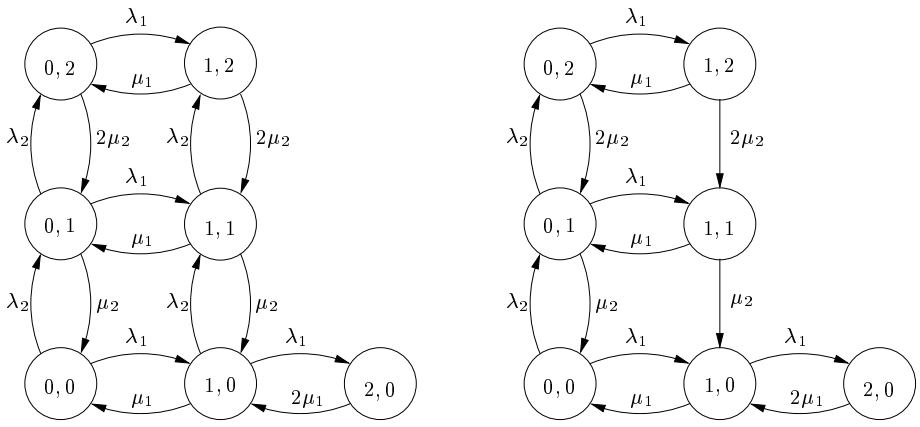

Figura 3.2: Cadeias de Markov para reserva tipo σ (esquerda) e tipo θ (direita)

A Figura 3.2 exemplifica um sistema em que $c_1 = 2$ e $c_2 = 1$, para um total de recursos $C = 4$, onde se empregam duas estratégias diferentes de reserva. O lado esquerdo representa a situação $\sigma_1 = 2$ e $\sigma_2 = 0$, enquanto o lado direito representa $\theta_1 = 0$ e $\theta_2 = 2$ onde

verificam-se quebras de simetria: por exemplo, para a transição $\delta_2^-(1,1) = 1$ não existe correspondente no sentido oposto, pois $\delta_2^+(1,0) = 0$.

3.4 Cálculo do bloqueio

O cálculo das probabilidades de bloqueio pode ser executado por diferentes métodos, com complexidade computacional variável. Também a aplicabilidade ou não de um método depende das políticas de reserva adotadas naquele particular sistema.

3.4.1 Método das equações de equilíbrio

Sistemas com processos de chegada poissonianos e com tempo de retenção exponencial negativo podem ser modelados pelas equações de equilíbrio markoviano associadas ao sistema, e sua resolução permite inferir qual o bloqueio associado a cada uma das k classes de serviço. Isso é feito a partir da determinação das probabilidades P_n de permanência em cada um dos estados do sistema. O bloqueio é então entendido como a soma das probabilidades de todos os estados bloqueantes para uma certa classe i de serviço.

Um conjunto de estados $\Omega_i^B \subset \Omega$ é dito *bloqueante* para uma i-ésima classe se para uma política de acesso arbitrária Ω tem-se

$$\Omega_i^B = \{n \in \Omega; \quad n_i^+ \notin \Omega\} \tag{3.13}$$

Portanto, a probabilidade de bloqueio B_i para a classe i vale

$$B_i = \sum_{n \in \Omega_i^B} P_n \tag{3.14}$$

Considere agora um sistema com k classes de serviço, com as respectivas taxas de chegada λ_i poissonianas e tempo médio de retenção μ_i exponencial negativo sujeito a uma dada política de acesso Ω. Adicionalmente supõe-se que não haja fila de espera para as requisições entrantes e que sejam liberadas imediatamente caso haja bloqueio. As equações de equilíbrio podem ser escritas na seguinte forma [32]

$$\left[\sum_{i=1}^{k} \lambda_i \delta_i^+(n) + \sum_{i=1}^{k} n_i \mu_i \delta_i^-(n)\right] P_n = \sum_{i=1}^{k} \lambda_i \delta_i^-(n) P_{n_i^-} + \sum_{i=1}^{k} (n_i + 1)\mu_i \delta_i^+(n) P_{n_i^+} \tag{3.15}$$

Caso exista simetria de transição para todos os estados, conforme definido pela condição (3.12) então as condições locais de equilíbrio da equação (3.15) para todo $n \in \Omega$ podem ser simplificadas e reescritas como

$$\lambda_i \delta_i^-(n) P_{n_i^-} = n_i \mu_i \delta_i^-(n) P_n; \quad i = 1, \ldots, k \tag{3.16}$$

O sistema de equações (3.15) tem tantas equações quantos forem os n estados presentes em Ω. Por se tratar de um sistema com equações linearmente dependentes, a resolução nas incógnitas P_n pode ser feita pela determinação do espaço nulo da matriz das equações e

subseqüente normalização para que $\sum_n P_n = 1$. Embora o método conduza a resultados exatos e opere sobre espaços de estados mesmo sem simetria de transições o problema pode rapidamente tornar-se numericamente intratável mesmo para uma quantidade modesta de classes k em espaços Ω contendo muitos estados possíveis. Estas dificuldades estimulam o desenvolvimento de aproximações para as probabilides de bloqueios, em geral visando o dimensionamento. Por exemplo, em [22] é proposto um cálculo para o limite superior de estimativa do bloqueio, e que pode ser usado em redes cujas informações exatas sobre o tráfego sejam complexas ou mesmo incompletas.

3.4.2 Método do teorema do produto

A resolução das equações (3.15) pode ser facilitada pelo uso do *Teorema do Produto* [25], [32], que possibilita a obtenção das probabilidades P_n sem a resolução explícita do sistema de equações. Para uma política de acesso arbitrária Ω, coordenada-convexa, com transições simétricas, contendo k classes com os respectivos tráfegos $a_i = \lambda_i/\mu_i$, a distribuição das probabilidades de estado P_n admite a representação na forma

$$P_n = \prod_{i=1}^{k} \frac{a_i^{n_i}}{n_i!} \cdot G^{-1}(\Omega) \tag{3.17}$$

$$G(\Omega) = \sum_{n \in \Omega} \left(\prod_{i=1}^{k} \frac{a_i^{n_i}}{n_i!} \right) \tag{3.18}$$

para todo $n \in \Omega$. O teorema pode ser verificado pela substituição da equação (3.17) na equação (3.16), sendo portanto uma solução que satisfaz também a equação (3.15) para o espaço de estados.

Uma dificuldade para a aplicação do teorema do produto reside em determinar de forma eficiente o conjunto de estados Ω. O conjunto dos estados bloqueantes Ω_i^B permite o cálculo direto das probabilidades de bloqueio B_i das classes por meio de um corolário do teorema dado por

$$B_i = \frac{G(\Omega_i^B)}{G(\Omega)} \tag{3.19}$$

com Ω_i^B definido pelas condições (3.13) e a função $G(.)$ dada pela equação (3.18). A complexidade numérica das soluções em forma de produto ainda é elevada, da ordem de $O(C^k)$ conforme [43], o que limita o uso a sistemas pequenos. Choudhury et al. propõem em [14] um algoritmo alternativo de cálculo baseado em inversão numérica de funções geradoras da constante de normalização, de modo a reduzir a complexidade computacional.

A maior restrição ao uso da forma-produto contudo encontra-se no fato da mesma ser solução apenas dos sistemas markovianos cujas políticas de reserva mantenham a simetria das transições entre os estados [48],[43]. O fato do sistema não ser coordenada-convexo não possui relevância aparente conforme [32]. Porém, políticas de reserva que coloquem regras adicionais às definidas pela equação (3.6) podem destruir a simetria, especialmente aquelas baseadas em marcação explícita de recursos reservados ou aquelas do tipo *por limiar* θ_i. Isto reduz sensivelmente os casos genuinamente tratáveis sob a óptica do teorema do produto basicamente àquelas políticas do tipo *ao menos* σ_i.

3.4.3 Método de Kaufman

Kaufman propôs um método de cálculo das probabilidades de bloqueio B_i baseado em um cálculo iterativo [32] que reduz bastante o esforço computacional, sendo a complexidade do método da ordem de $O(C \cdot k)$, conforme [43]. A fim de reduzir a quantidade de cálculos envolvidos, uma agregação das probabilidades de estado que ocupem a mesma quantidade de recursos C é usada. O espaço de estados original Ω com k dimensões é remapeado para outro espaço unidimensional pertencente a $\mathcal{U}^1$, onde as novas probabilidades de estados P_j valem

$$P_j = \sum_{\Omega^j} P_n$$

onde

$$\Omega^j = \{n; \quad n \cdot c = j\}; \quad j = 0, 1, \cdots, C$$

e representam a probabilidade de encontrar o sistema num estado n em que j unidades dos C recursos estão tomadas. Isso permite a localização dos *estados agregados bloqueantes* para uma i-ésima classe de serviço, sabendo-se quantos recursos c_i são necessários para atendê-la, como todos aqueles estados em que $j \geq C - (c_i - 1)$.

Algoritmo

Seja $a_i = \lambda_i/\mu_i$ o tráfego oferecido para cada uma das k classes de serviço, num sistema com um total de C recursos. As probabilidades unidimensionais não-normalizadas $\tilde{P}_j$ valem

$$\tilde{P}_j = \begin{cases} 1 & ; j = 0; \\ 0 & ; j < 0; \\ \dfrac{1}{j} \cdot \displaystyle\sum_{i=1}^{k} a_i c_i \tilde{P}_{j-c_i} & ; 0 < j \leq C; \end{cases} \tag{3.20}$$

após a normalização, tem-se:

$$P_j = \tilde{P}_j \left(\sum_{m=0}^{C} \tilde{P}_m \right)^{-1} \tag{3.21}$$

e os bloqueios B_i por classe de serviço,

$$B_i = \sum_{j=C-c_i+1}^{C} P_j \tag{3.22}$$

Este mapeamento não discrimina *como* se compõe a ocupação de recursos por parte de cada uma das i classes, focando apenas na *quantidade* de recursos ocupados. O método aplica-se apenas para sistemas sem reservas por classe. A simetria de transições é condição *sine qua non* para a validade do método. Algumas alterações foram propostas por Roberts [44] e Johnson [30] para incluir reservas, embora neste caso o método se torne aproximado segundo [43], com resultados melhores para quantidades baixas de recursos reservados. Outra generalização do método foi proposta por Delbrouck em [19], onde uma equação recursiva é adicionada de modo a tratar tráfegos cujo processo de chegada seja diferente do poissoniano (i.e. fator de pico $\zeta \neq 1$). O autor argumenta que tais tráfegos quando combinados com chegadas poissonianas podem alterar significativamente as probabilidades de bloqueio verificadas.

3.4.4 Método baseado em Erlang

A formulação clássica devida a Erlang para a probabilidade de bloqueio de um serviço com chegada poissoniana de taxa λ e tempo médio de retenção exponencial negativo $1/\mu$, oferecido a um grupo contendo C circuitos é dada por

$$B = Erl(a, C) = \frac{a^C/C!}{\sum_{m=0}^{C} a^m/m!} \tag{3.23}$$

sendo $a = \lambda/\mu$. Esta fórmula, também conhecida como *Erlang-B*, pode ser empregada para o cálculo do bloqueio em sistemas multisserviços que utilizem a política de reserva total. Neste caso, cada uma das i classes opera com C_i circuitos sem recursos comuns e portanto

$$C = \sum_{i=1}^{k} C_i \tag{3.24}$$

A quantidade de recursos c_i requerida por uma dada classe i determina o número de blocos de recurso disponíveis q_i, sendo dada por

$$q_i = \lfloor C_i/c_i \rfloor \tag{3.25}$$

onde $\lfloor \cdot \rfloor$ denota o maior inteiro menor ou igual a C_i/c_i e serve de base para cada uma das funções Erlang-B associadas aos bloqueios. Assim cada bloqueio torna-se

$$B_i = Erl(a_i, q_i) \tag{3.26}$$

A complexidade de cálculo da função Erlang-B é baixa, tipicamente da ordem de $O(C)$ para cada classe de serviço. Os algoritmos de cálculo já foram exaustivamente discutidos em diversos trabalhos, como por exemplo [23].

3.4.5 Método de LeGall

LeGall e Bernussou [24] propuseram um método aproximado para o cálculo de um sistema composto de uma única classe de serviço oferecida a dois grupos de circuitos com transbordo mútuo e reserva de um número máximo de circuitos (recursos) ocupados. Embora este seja um método restrito à existência de uma única classe, pode ser usado como parte na implementação de políticas de reserva mais complexas.

A fórmula derivada em [24] permite a aplicação de reservas para o tráfego mutuamente transbordante. Sejam os tráfegos oferecidos a_P e a_S pertencentes a uma mesma classe encaminhados respectivamente aos grupos de recursos C_P e C_S. Estes tráfegos não sofrem restrição. Os correspondentes transbordos resultantes do bloqueio em primeira escolha, representados como a_P^t e a_S^t, são reciprocamente re-oferecidos ao outro grupo de circuitos. Estes tráfegos "intrusos" estão sujeitos à política de reservas θ_P^t e θ_S^t, sendo portanto obrigados a deixar um total de recursos livres de pelo menos θ_S^t (no caso de a_P^t) ou θ_P^t (para o tráfego a_S^t). A Figura 3.3 apresenta o processo esquematicamente.

A probabilidade conjunta de que todos os $C_P + C_S$ recursos estejam indisponíveis é

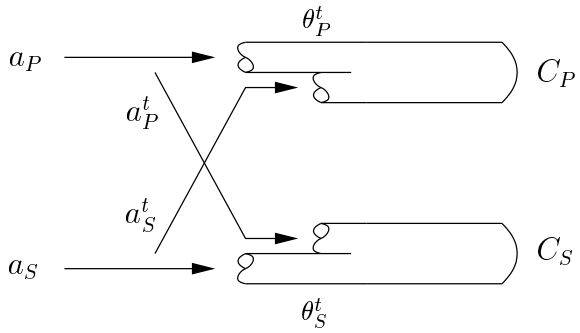

Figura 3.3: Esquema de transbordo mútuo

$$B_{P,S} = \left[\frac{a_P \; C_S!}{a_T \; a_S^{C_S}} \sum_{j=C_S-\theta_S^t}^{C_S} \frac{a_S^j/j!}{Erl(a_T, C_P + j)} + \frac{a_S \; C_P!}{a_T \; a_P^{C_P}} \sum_{j=C_P-\theta_P^t}^{C_P} \frac{a_P^j/j!}{Erl(a_T, C_S + j)} \right]^{-1} \tag{3.27}$$

onde $a_T = a_P + a_S$ e $Erl(.)$ é definida pela equação (3.23). Esta equação (3.27) recebe a seguinte notação:

$$LGall(a_P, a_S, C_P, C_S, \theta_P^t, \theta_S^t)$$

Os bloqueios totais dos tráfegos a_P e a_S são dados por

$$B_P = \frac{C_P!}{a_P^{C_P}} \; B_{P,S} \sum_{j=C_P-\theta_P^t}^{C_P} a_P^j/j! \tag{3.28}$$

$$B_S = \frac{C_S!}{a_S^{C_S}} \; B_{P,S} \sum_{j=C_S-\theta_S^t}^{C_S} a_S^j/j! \tag{3.29}$$

Tomando como ponto de partida esta formulação bastante genérica, simplificações são feitas de modo a tornar o uso das equações mais prático ao caso de interesse deste trabalho, onde é considerado o transbordo multiclasses em apenas um sentido. Uma classe (protegida) é considerada como o primeiro tráfego a_P e todas as demais são agrupadas no segundo tráfego a_S.

Para a situação $\theta_S^t = C_S$, ou seja, não é permitido transbordo do primeiro tráfego sobre o segundo, este possui todos os C_S circuitos para si e disputa os C_P recursos com o primeiro, seja livremente ($\theta_P^t = 0$), seja de maneira restrita ($0 < \theta_P^t \leq C_P$). Além disso, quando $C_S = 0$, apenas os C_P circuitos são disputados por ambos os tráfegos, conforme mostra a Figura 3.4. A proteção do tráfego a_P depende do valor escolhido para θ_P^t, podendo ser exclusiva ($\theta_P^t = C_P$), parcial ($0 < \theta_P^t < C_P$) ou ainda compartilhada ($\theta_P^t = 0$).

A equação (3.27) torna-se então

$$B_{P,*} = Erl(a_T, C_P) \left[\frac{a_P}{a_T} + \left(\frac{a_S}{a_T} \; \frac{Erl(a_T, C_P)}{a_P^{C_P}/C_P!} \sum_{j=C_P-\theta_P^t}^{C_P} \frac{a_P^j/j!}{Erl(a_T, j)} \right) \right]^{-1} \tag{3.30}$$

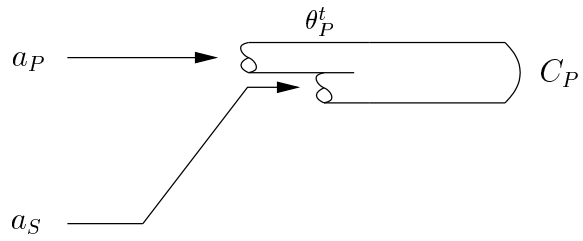

Figura 3.4: Esquema de proteção ao transbordo em um único sentido

Para o tráfego principal a_P limitado e o tráfego intruso na faixa $0 \leq a_S < +\infty$, os valores-limite assintóticos valem

$$T_0(a_P, C_P) \stackrel{\triangle}{=} \lim_{a_S \to 0} B_{P,*} = Erl(a_P, C_P) \tag{3.31}$$

$$T_\infty(a_P, C_P, \theta_P^t) \stackrel{\triangle}{=} \lim_{a_S \to +\infty} B_{P,*} = \frac{a_P^{C_P}/C_P!}{\sum_{j=C_P-\theta_P^t}^{C_P} a_P^j/j!} \tag{3.32}$$

Note que a equação (3.32) é similar à fórmula Erlang-B e mesmo para o tráfego intruso a_S tendendo ao infinito a mesma encontra-se limitada pois

$$T_\infty(\theta_P^t = C_P) = Erl(a_P, C_P) \tag{3.33}$$

$$T_\infty(\theta_P^t = 0) = 1 \tag{3.34}$$

Assim é possível dimensionar, com a escolha adequada de uma reserva θ_P^t, um valor para $T_\infty(\cdot)$ garantindo que uma i-ésima classe de serviço tenha um bloqueio no intervalo

$$T_0(\cdot) \leq B_i \leq T_\infty(\cdot) \leq 1 \tag{3.35}$$

independentemente do valor dos demais tráfegos intrusos.

3.5 Análise da distribuição dos estados

A correlação entre o tráfego oferecido e o número médio de recursos ocupados por classe pode ser inferida pela distribuição dos estados mais prováveis de um sistema multisserviços.

Seja um espaço $\Omega \in \mathcal{U}^k$ coordenada-convexo que possua simetria entre transições de estado, a probabilidade não-normalizada $\tilde{P}_n$ de um estado n_i qualquer, tomada a partir da equação (3.17) vale

$$\tilde{P}_n = \prod_{i=1}^{k} \frac{a_i^{n_i}}{n_i!} \tag{3.36}$$

cujo valor independe da política de acesso empregada.

Seja a função $\tilde{p}(\nu)$ generalizada a partir da equação (3.36) nas variáveis de estado contínuas tal que $0 \leq \nu_i < \infty$, para $i = 1, \ldots, k$

$$\tilde{p}(\nu) = \prod_{i=1}^{k} \frac{a_i^{\nu_i}}{\Gamma(\nu_i + 1)} \tag{3.37}$$

com $\Gamma(\cdot)$ sendo a função Gama definida para valores positivos pela integral

$$\Gamma(x) = \int_0^\infty e^{-t} t^{x-1} dt \tag{3.38}$$

Por construção, $\tilde{p}(n) = \tilde{P}_n$. A função $\tilde{p}(\nu)$ tem seu máximo para valores positivos das componentes ν_i, dado que cada uma delas é estritamente positiva e crescente para qualquer $a_i > 1$. No ponto de máximo da função $\tilde{p}(\nu)$ as derivadas parciais de cada uma das k classes deve se anular

$$\max_{\nu} \tilde{p}(\nu) \Longrightarrow \left[\frac{\partial \tilde{p}(\nu)}{\partial \nu_1}, \cdots, \frac{\partial \tilde{p}(\nu)}{\partial \nu_k} \right] = 0 \tag{3.39}$$

onde cada i-ésima componente é dada por

$$\frac{\partial \tilde{p}(\nu)}{\partial \nu_i} = \left[\frac{\partial}{\partial \nu_i} \left(\frac{a_i^{\nu_i}}{\Gamma(\nu_i + 1)} \right) \right] \prod_{\substack{j=1 \\ j \neq i}}^{k} \frac{a_j^{\nu_j}}{\Gamma(\nu_j + 1)} \tag{3.40}$$

A derivada no segundo membro da equação (3.40) deve se anular e portanto

$$\frac{a_i^{\nu_i}}{\Gamma(\nu_i + 1)} \left(\ln(a_i) - \frac{\partial}{\partial \nu_i} \ln(\Gamma(\nu_i + 1)) \right) = 0 \tag{3.41}$$

A equação (3.39) só é nula nos pontos determinados pelas raízes das k equações transcedentais contidas em (3.41)

$$\frac{\partial}{\partial \nu_i} \ln(\Gamma(\nu_i + 1)) = \ln(a_i) \tag{3.42}$$

O lado esquerdo da equação (3.42) é monotônico para qualquer $\nu_i \geq 1$ implicando uma solução única e dependente do valor de a_i.

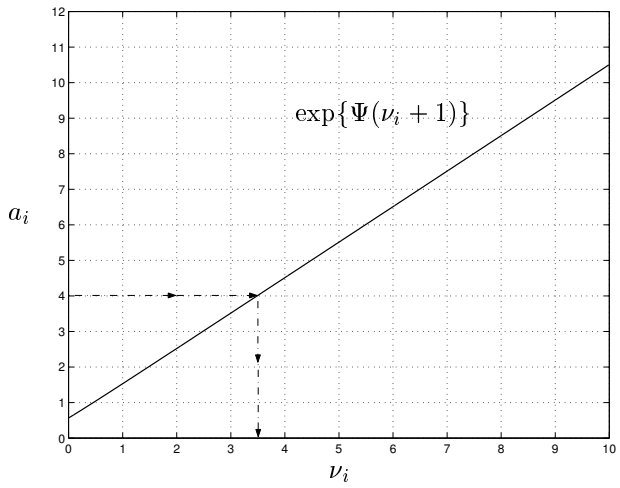

Figura 3.5: Relação entre o tráfego e a posição do ponto de máximo

A resolução numérica da equação (3.42) permite construir o gráfico da Figura 3.5, onde $\Psi(\nu_i + 1)$ representa o primeiro membro da equação (3.42), ilustrando o fato de que a i-ésima coordenada do ponto de máximo é tal que $|a_i - 1| < |\nu_i| < |a_i|$.

A função normalizada pelo pico $p^*(\cdot)$ é definida como

$$p^*(\nu) = \frac{\tilde{p}(\nu)}{\max_\nu \tilde{p}(\nu)} \tag{3.43}$$

e a *região de maior probabilidade* dos estados ν por sua vez é dada por

$$\Lambda(x) = \{\Lambda \subset \mathcal{U}^k;\ p^*(\nu) \leq x\};\quad 0 \leq x \leq 1 \tag{3.44}$$

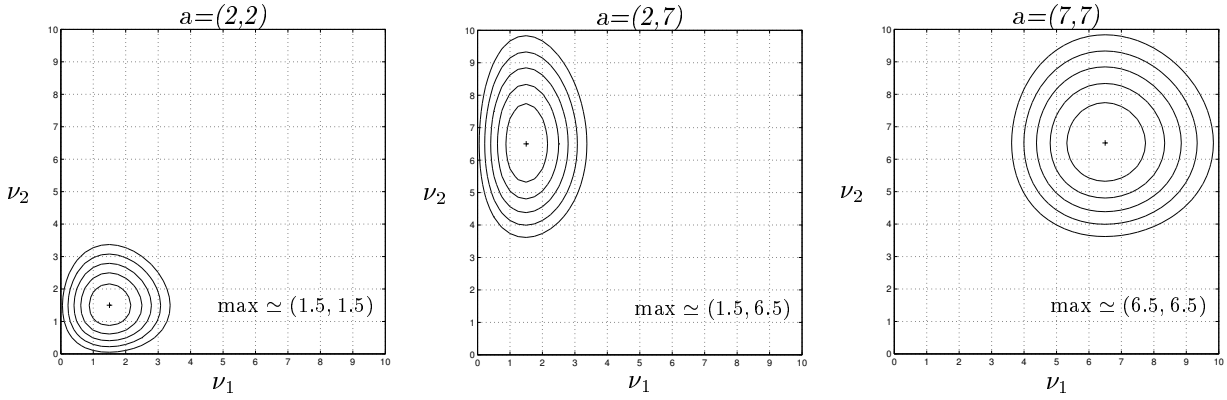

Figura 3.6: Deslocamento de posição da região de maior probabilidade Λ com o tráfego

A equação (3.44) permite a estimativa da dispersão de p^* em torno do máximo. A Figura 3.6 ilustra $\Lambda(x)$ como curvas de nível para o caso bidimensional ($k = 2$) em três situações distintas de tráfego, para os valores $x \in \{0.9;\ 0.8;\ 0.7;\ 0.6;\ 0.5\}$. Observa-se como o ponto de máximo segue a coordenada definida pelo valor a do tráfego na proporção indicada pelo gráfico da Figura 3.5. As curvas de $\Lambda(x)$ formam um padrão concêntrico, não simétrico, ao redor do máximo. A região Λ se dispersa, cobrindo uma área maior quando a coordenada do máximo se afasta do eixo do respectivo ν_i.

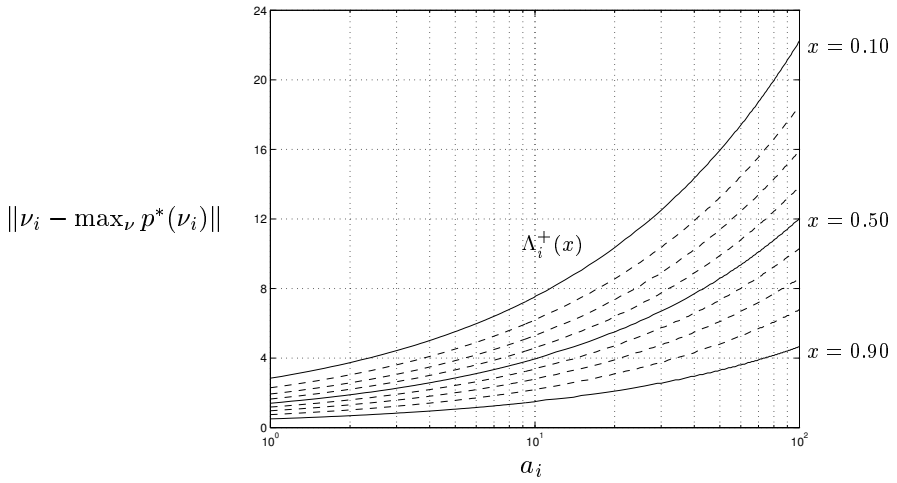

Figura 3.7: Alargamento da dispersão da região Λ_i^+ com o tráfego a_i

A Figura 3.7 mostra a distância radial de um ponto qualquer da curva $\Lambda_i^+(x)$. Esta é definida como $\Lambda(x)$ em uma i-ésima direção positiva dos eixos coordenados, medida a partir

do ponto de máxima probabilidade, para diversos valores do parâmetro x. Verifica-se que para uma curva de nível x a dispersão cresce com o aumento do tráfego a_i.

3.6 Efeitos das reservas

A reserva de recursos altera a probabilidade de bloqueio das classes em um ambiente multisserviços sendo empregada pela sua simplicidade e eficiência. Contudo, diversos outros parâmetros são afetados neste procedimento.

3.6.1 Alteração do bloqueio

Um sistema multisserviços com k diferentes classes, caracterizadas pelo descritor de serviços D_i, pode ser representado pelo *descritor abreviado de serviços* $\hat{D}_i$ tal que

$$\hat{D}_i[a_i, c_i] = D_i[a_i, c_i, 1, +\infty] \tag{3.45}$$

onde $\zeta_i = 1$ e $\mathcal{N}_i = +\infty$ para todas as classes, e que é aplicável para o caso de tráfego poissoniano. Cada uma das i-classes é caracterizada pelo $\hat{D}_i$ numa matriz $\hat{D}$ de dimensões $(k \times 2)$ onde cada linha contém a dupla correspondente à i-ésima classe. Considerando $C_i = 0$ para todo i (compartilhamento total) as probabilidades de bloqueio das classes podem ser calculadas a partir das fórmulas de recorrência de Kaufman, equações (3.20), (3.21), (3.22) e representadas pelo vetor B de k elementos dado por $B = Kauf(\hat{D}, C)$.

A enumeração das classes de serviço de modo que

$$c_1 \geq c_2 \geq \cdots \geq c_k \tag{3.46}$$

implica que as respectivas probabilidades de bloqueio obedecem à relação

$$B_1 \geq B_2 \geq \cdots \geq B_k \tag{3.47}$$

independentemente dos tráfegos. Esta relação foi inicialmente observada por Aein e Kosovych em [3]. Isto pode ser verificado por exemplo na equação (3.22), onde o somatório varia entre $(C - c_i + 1)$ e (C), o que adiciona um número maior de termos $P_j > 0$ para maiores valores de c_i.

A relação das magnitudes dos B_i pode ser alterada pela reserva de recursos para a i-ésima classe na qual se planeja reduzir o bloqueio. Em contrapartida, as demais classes sofrem um aumento em seus respectivos bloqueios já que o universo de recursos disponíveis foi reduzido de C para $C - C_i$.

Considere o exemplo de um *sistema de referência* (chamado aqui de `Ref`) com duas classes e cujo descritor tem os parâmetros

$$\hat{D} = \begin{bmatrix} 3 & 4 \\ 16 & 1 \end{bmatrix}; \qquad C = 44;$$

Para este sistema, `Ref-I` designa a política de compartilhamento total, enquanto em `Ref-II` uma política de reservas estática do tipo *ao menos* σ_i com $C_1 = 20$ e $C_2 = 0$ é adotada. A Tabela 3.1 mostra os resultados.

Tabela 3.1: Comparação de sistemas com e sem reservas

classe i	tráfego a_i	recursos c_i	Bloqueio B_i	
			`Ref-I`	`Ref-II`
1	3 erl	4	4.08 %	3.75 %
2	16 erl	1	0.76 %	1.96 %

As magnitudes dos bloqueios, num sistema com processos de chegada poissonianos utilizando compartilhamento total de recursos, são geralmente inferiores às dos sistemas com reserva plena para o *conjunto* das classes conforme estudo realizado por Aein em [2]. A reserva pode ser feita nos casos em que se deseje influenciar apenas o bloqueio de algumas das classes mais críticas, ainda que em detrimento das demais.

3.6.2 Sensibilidade ao tráfego

Quando um sistema multisserviços é dimensionado, o parâmetro principal a ser considerado é, em geral, o bloqueio experimentado pelas classes de serviço. Todavia, um outro conjunto de parâmetros intimamente associado é a *variação* da taxa nominal de bloqueio conforme o sistema se afasta do tráfego oferecido nominal – isto é, o tráfego para o qual foi originalmente dimensionado. A variação deste tráfego em uma certa classe i não só afeta a própria como também as demais, num grau maior ou menor dependendo de como opera o sistema. Dito de outra maneira, uma certa classe que exceda ao seu tráfego nominal não só aumentará o bloqueio experimentado pela mesma como degradará perceptivelmente as demais.

Seja a matriz J das derivadas parciais dos bloqueios em relação aos tráfegos oferecidos, calculadas para todas as classes

$$J = \begin{bmatrix} \partial B_1/\partial a_1 & \partial B_1/\partial a_2 & \cdots & \partial B_1/\partial a_k \\ \partial B_2/\partial a_1 & \partial B_2/\partial a_2 & \cdots & \partial B_2/\partial a_k \\ \vdots & \vdots & & \vdots \\ \partial B_k/\partial a_1 & \partial B_k/\partial a_2 & \cdots & \partial B_k/\partial a_k \end{bmatrix} \iff J(x,y) = \left[\frac{\partial B_x}{\partial a_y}\right] \tag{3.48}$$

sendo a matriz J simétrica, isto é

$$\frac{\partial B_x}{\partial a_y} = \frac{\partial B_y}{\partial a_x} \tag{3.49}$$

A propriedade da equação (3.49) é denominada *relação de reciprocidade* e foi demonstrada por Virtamo em [48] para sistemas poissonianos, coordenada-convexos e cujas transições entre estados adjacentes sejam permitidas em ambos os sentidos. Nestes sistemas a interferência mútua entre classes possui o mesmo grau de sensibilidade e caso uma certa política de acesso diminua a influência do tráfego de uma classe x sobre o bloqueio de uma outra classe y, a recíproca é verdadeira e tem a mesma magnitude. Em `Ref-I` a matriz J^I calculada no ponto de tráfego nominal ($a_1 = 3$ e $a_2 = 16$) vale

$$J^I_{(3,16)} = \begin{bmatrix} 0.0396 & 0.0082 \\ 0.0082 & 0.0017 \end{bmatrix}$$

Em `Ref-II` a matriz J^{II} correspondente ao mesmo tráfego assume os valores

$$J^{II}_{(3,16)} = \begin{bmatrix} 0.0389 & 0.0062 \\ 0.0062 & 0.0079 \end{bmatrix}$$

Nota-se aqui a redução da sensibilidade mútua quando a política de reservas é adotada. Caso uma política de reserva plena seja utilizada, estes mesmos termos vão a zero. A matriz J é apenas uma descrição *local* da sensibilidade e não se presta para extrapolações para muito além do ponto em que foi calculada.

A Figura 3.8 mostra o valor dos bloqueios B_1 e B_2 para a variação dos tráfegos oferecidos a_1 e a_2 entre -100% e $+100\%$ do valor nominal, para o sistema `Ref-I`.

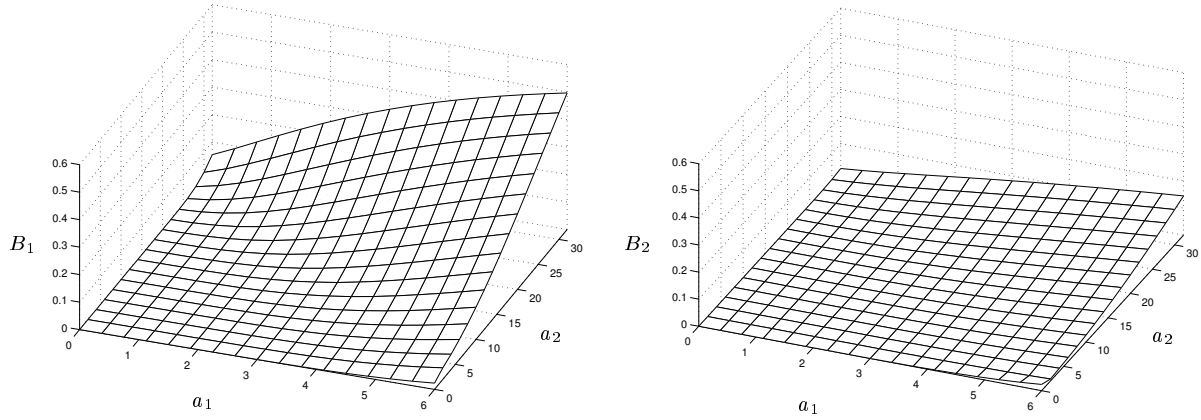

Figura 3.8: Bloqueio das classes 1 e 2 - Compartilhamento total (`Ref-I`)

A Figura 3.9 representa B_1 e B_2 com a mesma variação de -100% a $+100\%$ do tráfego nominal para o sistema `Ref-II`. A reserva de recursos favorece o bloqueio B_1 em certas regiões de tráfego (por exemplo $a_1 \sim 2$ e $a_2 \sim 32$) enquanto ao mesmo tempo penaliza o bloqueio B_2.

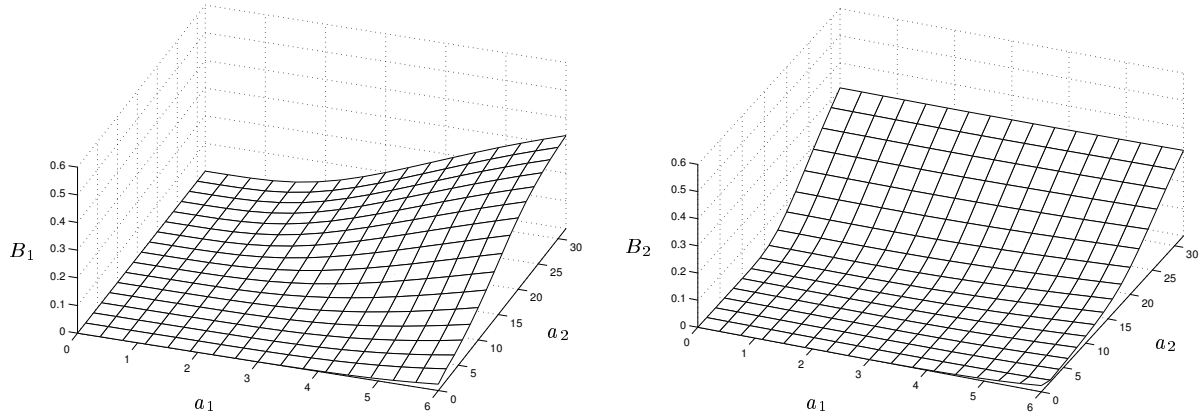

Figura 3.9: Bloqueio das classes 1 e 2 - Reserva parcial (`Ref-II`)

Observa-se que sempre quando uma classe é beneficiada pela reserva de recursos isso se faz à custa das demais.

3.6.3 Fenômeno oscilatório

A variação dos bloqueios com o aumento do tráfego, embora seja tendencialmente crescente, não é necessariamente monotônica em sistemas contendo classes de serviço com diferentes requisitos de recursos, notadamente nas classes de menor c_i. Por exemplo, o bloqueio B_2 na Figura 3.8 possui uma região em torno das coordenadas ($a_1 = 6$, $a_2 = 1$) em que este comportamento pode ser observado, conforme indica a matriz J correspondente

$$J_{(6,1)} = \begin{bmatrix} 0.0265 & 0.0079 \\ 0.0079 & -0.0065 \end{bmatrix}$$

pois o valor negativo do termo $(2,2)$ aponta uma diminuição do bloqueio com o aumento do tráfego próprio. A Figura 3.10 ilustra este fenômeno para o cenário `Ref-I`.

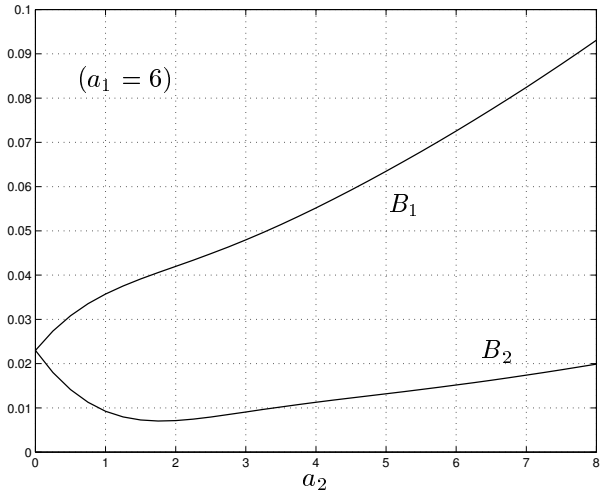

Figura 3.10: Fenômeno oscilatório para duas classes de serviço com variação de a_2

Este fenômeno é local e se observa quando a diferença de banda c_i entre as classes e a diferença entre os tráfegos é significativa, tendo sido constatado já em [30], [48] e [43].

Uma variante do fenômeno oscilatório pode ser observada quando, para um dado sistema com tráfego oferecido constante, altera-se progressivamente a quantidade total C de recursos disponíveis. A Figura 3.11 ilustra este comportamento para um sistema sem reservas com $c_1 = 4$, $c_2 = 1$ sujeito a um tráfego nominal $a = (6, 1)$ e onde $30 \leq C \leq 44$.

O bloqueio B_2 apresenta uma oscilação sobreposta de período 4 unidades, que é exatamente o valor de c_1. Observa-se também que para valores de C que não são múltiplos inteiros de c_1 o valor de B_2 assume valores comparativamente menores. Usando o exemplo dado, este fato pode ser entendido utilizando a argumentação que se segue. Para valores de C simultaneamente múltiplos de c_1 e c_2 (por exemplo $32, 36, 40, \ldots$), ambos os serviços podem fazer uso de todos os C recursos. Seja por exemplo $C = 36$. Ao acresecentar-se um recurso a mais (chegando a 37), apenas a segunda classe é favorecida diretamente pela adição do recurso $C + 1$. A primeira classe continua possuindo apenas os C recursos para seu uso, sem poder tomar o recurso adicional, que não é múltiplo inteiro de c_1. Ainda assim, esta classe é beneficiada indiretamente, pois o tráfego concorrente percebido nos seus C recursos visíveis é ligeiramente menor. Isso se repete *mutatis mutandis* para a adição dos dois próximos recursos totais (isto é, para os casos $C+2$ e $C+3$). Quando o quarto recurso

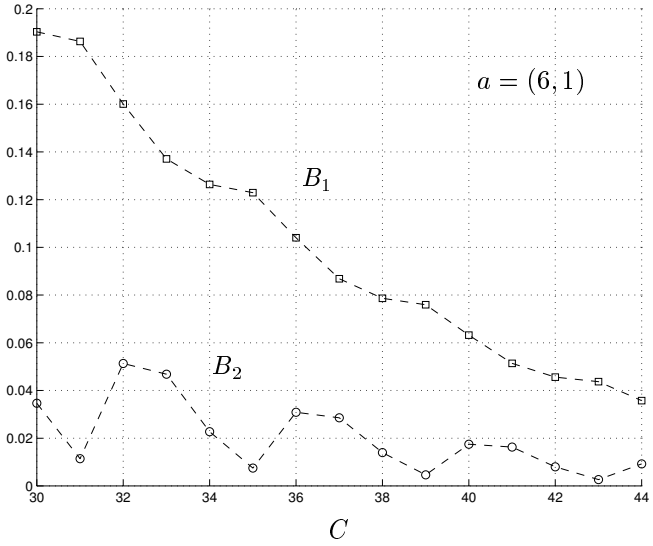

Figura 3.11: Fenômeno oscilatório para duas classes de serviço com variação de C

é acrescido (totalizando 40 no exemplo), os agora $C+4$ recursos, múltiplo simultâneo de c_1 e c_2, beneficiam diretamente ambas as classes. A segunda passa a sofrer uma interferência maior do tráfego da primeira, já que estes quatro recursos adicionais são disputados agora em condição de igualdade. O efeito líquido é um pequeno aumento do bloqueio da segunda classe. Este processo se repete segundo esse ciclo para cada quatro recursos adicionados, estabelecendo o período oscilatório observado na Figura 3.11.

3.7 Usos das reservas

Políticas de reserva visando a alteração dos bloqueios em sistemas multiclasses podem ser adotadas tendo em mente dois propósitos principais. Um deles é a economia de recursos utilizados objetivando a redução do custo de implementação de um determinado sistema. O outro é assegurar um grau mínimo de proteção de uma determinada classe de serviço contra uma possível sobrecarga de tráfego nas demais, isto é, uma garantia mínima de grau de serviço.

3.7.1 Economia de recursos

No dimensionamento de um sistema é comum referir-se a um certo *bloqueio objetivo* $B^\star$ e manter os bloqueios nominais calculados abaixo deste determinado valor, isto é

$$\max\{B_1, B_2, ..., B_k\} \leq B^\star$$

Num ambiente multisserviços com compartilhamento total os bloqueios experimentados pelas classes variam na relação direta da quantidade de recursos requeridos pela classe. Define-se o *bloqueio natural* B^N de um sistema, contendo C recursos e representado pelo descritor de serviços $\hat{D}$, como o vetor de bloqueios B_i de cada classe de serviço numa política sem reservas por

$$B^N = Kauf(\hat{D}, C) \tag{3.50}$$

Supondo as classes ordenadas tal que $c_1 > c_2 > \cdots > c_k$, os bloqueios naturais terão suas magnitudes ordenadas da mesma forma $B_1^N > B_2^N > \cdots > B_k^N$. Define-se o *bloqueio resultante* da adoção de uma certa política de reservas como B^R. A possibilidade de economia de recursos baseia-se no pressuposto de que, dado um valor $B^\star$, por meio da reserva é possível a redução do valor do bloqueio B_1^R abaixo do seu valor natural B_1^N sem contudo que o aumento correspondente observado em $B_2^R, \ldots, B_k^R$ ultrapasse o valor especificado de $B^\star$. O novo B_1^R por sua vez deve ser dimensionado de modo a também permanecer abaixo de $B^\star$.

3.7.2 Estratégias de reserva

Diferentes estratégias de reserva podem ser adotadas visando a redução do número de recursos necessários. Para um descritor dado $\hat{D}$, a descrição de algumas destas possíveis estratégias $\mathcal{S}$ é definida pelos procedimentos a seguir.

• Compartilhamento total $\mathcal{S}^N$ (método de referência)

O acesso sem restrição a todos os C recursos pelas i classes, calculado valendo-se do método de Kaufman. Define-se esta operação matemática recursiva como uma função representada da forma

$$B = Kauf(\hat{D}, C)$$

onde $\hat{D}$ é a matriz dos descritores de serviços.

O algoritmo segue os seguintes passos:

- Dados um descritor de serviço $\hat{D}$ para k classes ordenadas e um certo bloqueio objetivo $B^\star$;

(1) Calcula-se como estimativa inicial $C = \sum_{i=1}^{k} c_i$;

(2) Calcula-se o bloqueio natural $B = Kauf(\hat{D}, C)$;

(3) Se $\max\{B_i\} \leq B^\star$ o processo termina e o resultado é igual a C;

(4) Caso contrário, $C \leftarrow C + 1$ e repete-se desde o passo (2).

Este método de cálculo também é utilizado como parte das estratégias de reserva descritas adiante, para a determinação do bloqueio natural.

• Reserva parcial $\mathcal{S}^{Sr}$ (método estático *ao menos* σ, para r classes)

Nesta estratégia é adotada uma reserva inicial $\sigma_i = C_i$ para a i-ésima classe de serviço por meio de um critério heurístico. O tipo de política é tal que permite o uso do teorema do produto para o cálculo do bloqueio. Define-se esta operação matemática como

$$B = Tprod(\hat{D}, \sigma, C)$$

onde $\sigma = (\sigma_1, \cdots, \sigma_k)$ é o vetor de reservas de recursos.

O dimensionamento é descrito por:

- Dados o descritor de serviço $\hat{D}$ para k classes ordenadas, uma reserva inicial $C_i = \sigma_i$ e um bloqueio objetivo $B^\star$;

(1) Calcula-se pela estratégia anterior $\mathcal{S}^N$ o bloqueio natural que satisfaça a condição dada por $B^\star$ e adota-se o resultado como o valor inicial de C;

(2) Define-se a variável auxiliar $g = 0$;

(3) Calcula-se o bloqueio com reserva $B = Tprod(D_s, \sigma, C - g)$;

(4) Caso $\max\{B_i\} \leq B^\star$ faz-se $g \leftarrow g + 1$ e repete-se o processo desde o passo (3);

(5) Recalcula-se o bloqueio como $B = Tprod(D_s, \sigma, C - g)$;

(6) Caso $\max\{B_i\} > B^\star$ faz-se $g \leftarrow g - 1$ e repete-se o processo desde o passo (5);

(7) Se não, a redução máxima de recursos foi atingida e C deverá ser adotado como $C \leftarrow C - g$.

Quanto à decisão sobre o valor inicial adotado da reserva σ_i, considere-se por hora que diversos valores múltiplos dos respectivos c_i podem ser verificados por tentativa e erro, sendo que aquele que apresentar melhor resultado (i.e. maior valor possível de g) será adotado, embora esta estratégia implique o uso do algoritmo acima descrito diversas vezes. Após a resolução de um número suficiente de problemas semelhantes, heurísticas podem ser inferidas sobre qual o número adequado de reservas a ser adotado.

Por uma questão de simplicidade e eficácia, este algoritmo pode ser mais comumente usado para reserva da primeira classe de serviço (maior c_i) apenas, permanecendo as demais sem recursos dedicados, ou seja $\sigma = (\sigma_1, 0, \cdots, 0)$, caso em que a estratégia é denominada como $\mathcal{S}^{S1}$.

• Reserva parcial $\mathcal{S}^{L1}$ (transbordo controlado por limite θ, para uma classe)

Aqui adota-se como base um método baseado na aproximação de LeGall para a primeira classe em conjunto com dimensionamento auxiliar baseado em $Kauf(.)$ para as demais. Para efeito de nomenclatura, define-se a seguinte estruturação para os C recursos:

$$\begin{aligned} C_1 :&\quad \textit{recursos reservados para a classe } 1; \\ C_{2..k} :&\quad \textit{recursos compartilhados pelas classes } 2, \ldots, k; \\ C_V :&\quad \textit{recursos para transbordo de todas as classes}; \end{aligned}$$

A soma de C_1 com C_V é representada como C_{1+V}. As requisições de serviço são apresentadas primeiramente às suas respectivas reservas e, caso sejam bloqueadas em primeira escolha, sofrerão transbordo para o grupo comum de recursos C_V. Esta estratégia de transbordo não é passível de tratamento pelo método do teorema do produto da equação

(3.17), pois devido a quebra de simetrias internas no espaço de estado o sistema markoviano equivalente não admite uma solução na forma produto.

Da primeira fórmula de LeGall (3.27), para os grupos de recursos primários P e secundários S

$$B_{P,S} = LGall(a_P, a_S, C_P, C_S, \theta_P^t, \theta_S^t)$$

impõe-se $C_S = 0$ e $\theta_S^t = 0$, reduzindo a fórmula para a equação (3.30) cujos limites assintóticos T_0 e T_∞ são dados respectivamente pelas equações (3.31) e (3.32). Além disso, $\theta_P^t = \theta_1$. Nestas condições, a estratégia de dimensionamento assume a forma:

- Dados um descritor de serviço $\hat{D}$ para k classes ordenadas, e um bloqueio objetivo $B^\star$ de pior caso, além de uma faixa auxiliar ϵ tal que $0 < \epsilon < B^\star$

(1) Define-se um *bloqueio de melhor caso* $\tilde{B^\star}$ tal que $\tilde{B^\star} = B^\star - \epsilon$;

(2) Atribui-se o valor inicial $C_{1+V} = c_1$;

(3) Calcula-se $\tilde{B_1} = T_0(a_1, C_{1+V})$;

(4) Se $\tilde{B_1} > \tilde{B^\star}$ faz-se $C_{1+V} \leftarrow C_{1+V} + 1$ e repete-se o passo (3);

(5) Toma-se $\theta_1 = C_{1+V}$;

(6) Calcula-se agora $B_1 = T_\infty(a_1, C_{1+V}, \theta_1)$;

(7) Caso $B_1 \leq B^\star$ faz-se $\theta_1 \leftarrow \theta_1 - 1$ e repete-se o passo (6);

(8) Caso contrário, o valor do limiar adotado é $\theta_1 \leftarrow \theta_1 + 1$;

(9) Excluída a primeira classe do descritor de serviços tal que agora $\hat{D}^\dagger = \hat{D}(2 \cdots k)$, calcula-se $B_{2..k}$ pelo método de dimensionamento sem reservas ($\mathcal{S}^N$) para $\hat{D}^\dagger$ e $B^\star$, obtendo como resultado $C_{2..k}$;

(10) Finalmente, $C = C_{1+V} + C_{2..k}$.

Para comparação com o método de referência, g pode ser tomado indiretamente, como sendo a diferença entre o valor de C calculado com os bloqueios naturais segundo o método $\mathcal{S}^N$ com o valor obtido por este algoritmo.

Esta forma de dimensionamento tem a vantagem de garantir que todas as classes terão um bloqueio menor que $B^\star$, além de garantir que uma eventual sobrecarga no tráfego de a_1 não degrade o serviço do conjunto das demais classes a_i com $i = 2, \ldots, k$ e vice-versa. Outras estratégias semelhantes poderiam ser imaginadas para a inclusão seqüencial das outras $(k-1)$ classes de serviço num mecanismo de reservas com transbordo em cascata ou seja, uma estratégia generalizada $\mathcal{S}^{Lr}$.

• Reserva plena $\mathcal{S}^T$

A estratégia de reserva plena isola todas as classes de serviço das influências mútuas de tráfego. Dito de outra forma, nesta política de reserva a matriz J é sempre uma matriz

diagonal para quaisquer valores de a_i, $i = 1, \ldots, k$.

O algoritmo de dimensionamento pode ser descrito como:

- Dados um descritor de serviço $\hat{D}$ para k classes ordenadas e um bloqueio objetivo $B^\star$;

(1) Inicializa-se a variável auxiliar $i = 1$;

(2) Atribui-se o valor $q = 1$;

(3) Calcula-se o bloqueio da i-ésima classe como sendo $B_i = Erl(a_i, q)$;

(4) Caso $B_i > B^\star$ faz-se $q \leftarrow q + 1$ e repete-se o processo desde o passo (3);

(5) Caso contrário, o bloqueio já foi atingido e então corrige-se o valor da quantidade de recursos $C_i = q \cdot c_i$;

(6) A reserva desta classe vale C_i;

(7) Caso $i \leq k$, faz-se $i \leftarrow i + 1$ e repete-se o processo desde o passo (2);

(8) Finalmente, $C = (C_1 + C_2 + \cdots + C_k)$.

3.7.3 Comparações numéricas

Sejam três exemplos onde as estratégias de reserva acima são aplicadas para a primeira classe (maior c_i). Em cada um deles, a *ocupação de recursos* para cada classe, definida por $w_i = a_i c_i$ é tal que a interferência do grupo das classes $(2 \cdots k)$ na primeira classe pode ter pesos relativos distintos, tal que

$$\begin{array}{lcl} Caso\ H & : & (w_2 + \cdots + w_k) > w_1; \\ Caso\ M & : & (w_2 + \cdots + w_k) \approx w_1; \\ Caso\ L & : & (w_2 + \cdots + w_k) < w_1; \end{array} \tag{3.51}$$

para o conjunto de todas as classes.

Caso H – alta interferência

Neste caso, $(w_2 + \cdots + w_k) > w_1$ e o descritor é dado por

$$\hat{D}^H = \begin{bmatrix} 1 & 4 \\ 32 & 1 \end{bmatrix} \tag{3.52}$$

e os resultados dos procedimentos são mostrados na Tabela 3.2. Observa-se que a estratégia $\mathcal{S}^{S1}$ é a mais bem sucedida para a redução do número de circuitos para uma ampla faixa de valores de $B^\star$, com reservas de $C_1 = 8$. Para um bloqueio objetivo muito elevado (25%), a estratégia $\mathcal{S}^{L1}$ mostrou-se melhor, embora seja este um valor atípico para especificação na síntese de redes.

Tabela 3.2: Número de circuitos requeridos por estratégia – Tráfego $\hat{D}^H$

$B^\star$	$\mathcal{S}^N$		$\mathcal{S}^{S1}$		$\mathcal{S}^{L1}$		$\mathcal{S}^T$	
	C	$(C_1;C_2)$	C	$(C_1;C_2)$	C	$(C_1;C_2)$	C	$(C_1;C_2)$
1%	57	(0;0)	**56**	(8;0)	64	(16;0)	64	(20;44)
2%	54	(0;0)	**53**	(8;0)	58	(12;0)	58	(16;42)
5%	50	(0;0)	**48**	(8;0)	54	(12;0)	54	(16;38)
12%	46	(0;0)	**41**	(8;0)	45	(8;0)	45	(12;33)
25%	41	(0;0)	35	(8;0)	**33**	(4;0)	35	(8;27)

Caso M – média interferência

Aqui, $(w_2 + \cdots + w_k) \approx w_1$ e o descritor é dado por

$$\hat{D}^M = \begin{bmatrix} 4 & 4 \\ 16 & 1 \end{bmatrix} \tag{3.53}$$

com os resultados de dimensionamento mostrados na Tabela 3.3. Neste caso, a estratégia $\mathcal{S}^{S1}$ apresentou resultados razoáveis apenas nas situações de $B^\star$ com valor elevado (maior que 10%), mesmo para um nível de reservas $C_1 = 20$. Abaixo disso, o dimensionamento sem reservas $\mathcal{S}^N$ foi sempre melhor ou no mínimo igual a qualquer outra estratégia com reservas.

Tabela 3.3: Número de circuitos requeridos por estratégia – Tráfego $\hat{D}^M$

$B^\star$	$\mathcal{S}^N$		$\mathcal{S}^{S1}$		$\mathcal{S}^{L1}$		$\mathcal{S}^T$	
	C	$(C_1;C_2)$	C	$(C_1;C_2)$	C	$(C_1;C_2)$	C	$(C_1;C_2)$
1%	**57**	(0;0)	**57**	(20;0)	69	(20;0)	65	(40;25)
2%	**54**	(0;0)	**54**	(20;0)	64	(20;0)	60	(36;24)
5%	**48**	(0;0)	**48**	(20;0)	56	(20;0)	53	(32;21)
12%	42	(0;0)	**41**	(20;0)	47	(12;0)	42	(24;18)
25%	36	(0;0)	**35**	(20;0)	38	(12;0)	35	(20;15)

Caso L – baixa interferência

Neste caso, $(w_2 + \cdots + w_k) < w_1$ e o descritor é dado por

$$\hat{D}^L = \begin{bmatrix} 8 & 4 \\ 1 & 1 \end{bmatrix} \tag{3.54}$$

com os resultados mostrados na Tabela 3.4. Aqui sempre os melhores resultados foram obtidos com $\mathcal{S}^N$, para qualquer faixa de $B^\star$. O número de reservas neste caso parece ser irrelevante para o desempenho das outras estratégias, por mais elevado que seja.

Observa-se que estratégias alternativas nem sempre são eficazes para a economia de recursos quando comparadas ao sistema sem reservas $\mathcal{S}^N$. Esta eficácia varia com a faixa do bloqueio objetivo desejado e com o perfil de tráfego, sendo mais significativa para sistemas com perfil de tráfego do tipo $\hat{D}^H$.

Tabela 3.4: Número de circuitos requeridos por estratégia – Tráfego $\hat{D}^L$

$B^\star$	$\mathcal{S}^N$		$\mathcal{S}^{S1}$		$\mathcal{S}^{L1}$		$\mathcal{S}^T$	
	C	$(C_1;C_2)$	C	$(C_1;C_2)$	C	$(C_1;C_2)$	C	$(C_1;C_2)$
1%	**62**	(0;0)	62	(56;0)	69	(32;0)	65	(60;5)
2%	**58**	(0;0)	58	(52;0)	64	(28;0)	60	(56;4)
5%	**52**	(0;0)	52	(48;0)	56	(24;0)	56	(52;4)
12%	**44**	(0;0)	44	(40;0)	47	(20;0)	47	(44;3)
25%	**34**	(0;0)	34	(32;0)	38	(12;0)	34	(32;2)

3.7.4 Proteção de serviços

A proteção de uma ou mais classes contra a variação de tráfego das demais e conseqüente degradação assume duas formas principais. A primeira é através de um método de descarte mediante uma probabilidade de aceitação estocástica das requisições a partir de um determinado limiar, de modo a manter o tráfego oferecido dentro de limites estabelecidos. Estas são chamadas *políticas de aceitação probabilística*. Os tráfegos são verificados continuamente e, ao tráfego ofensor acima de um determinado valor, uma *restrição* ao ingresso de suas chamadas é imposta. A outra forma é através do uso da reserva de recursos, por meio da escolha de uma política de acesso adequada (reserva parcial ou total). Considere o sistema de referência, cenários `Ref-I` e `Ref-II`, respectivamente sem resevas e com reservas para a primeira classe com $C_1 = 20$. O tráfego nominal é $a_1 = 3$ e $a_2 = 16$. O bloqueio objetivo $B^\star$ é especificado *em condições normais* para melhor que 5%. A Figura 3.12 mostra as curvas de nível traçadas para diversos valores de bloqueio B_1 em função dos tráfegos a_1 e a_2. Nota-se que no cenário `Ref-II` para variações de a_2 o bloqueio cresce mais lentamente que no cenário `Ref-I`, além de atingir valores menores para uma mesma variação do tráfego. Uma variação no tráfego a_2 de 100% em `Ref-II` resulta num bloqueio para a classe 1 menor que 10%.

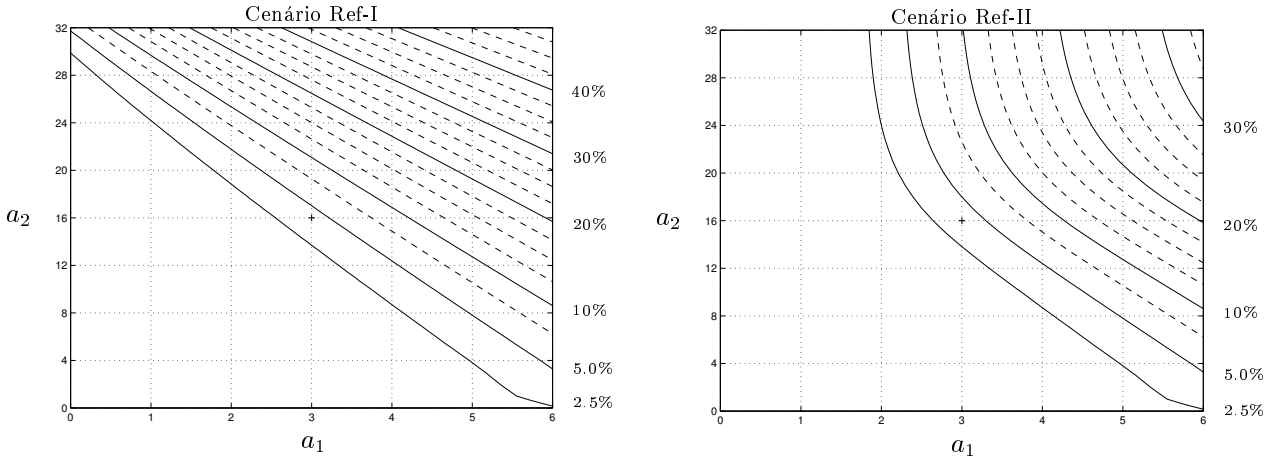

Figura 3.12: Curvas de bloqueio da classe 1

O valor da assíntota da curva de nível na direção do eixo das ordenadas em $a_1 = 3$ pode

ser obtido para $a_2 \to +\infty$ como sendo

$$\lim_{a_2 \to +\infty} B_1 = Erl(a_1, C_1) \approx 11\%$$

Este comportamento permite estimar o valor para a degradação máxima absoluta experimentada por um serviço face à sobrecarga dos demais, sendo possível dimensionar-se o sistema de modo a acomodar esta situação anormal de tráfego.

A Figura 3.13 mostra as curvas de nível para diversos valores do bloqueio B_2 em função de a_1 e a_2. Nota-se que no cenário `Ref-II` o valor de B_2 cresce rapidamente com o aumento do tráfego a_2, degradando este serviço mais severamente que no caso `Ref-I`.

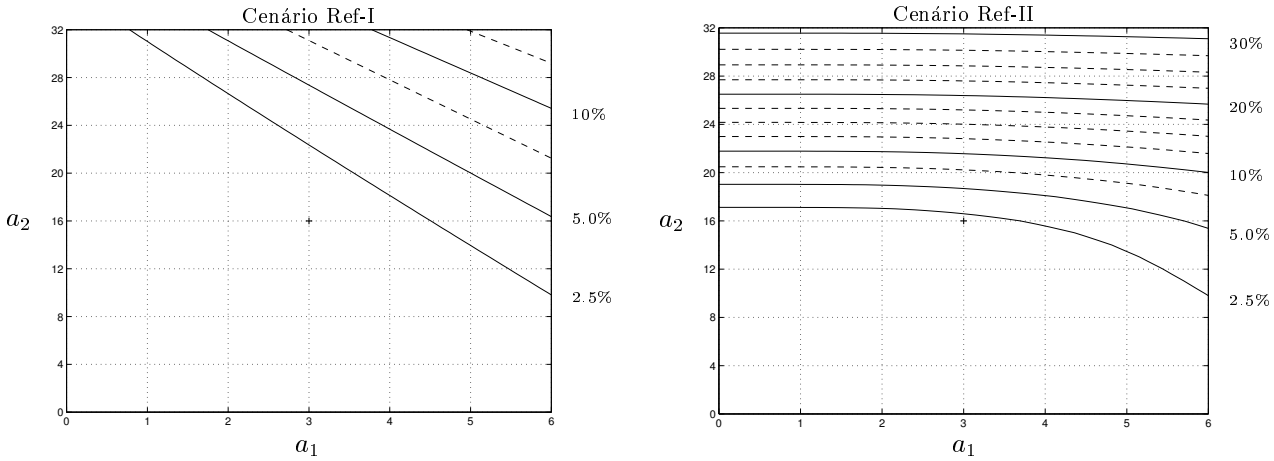

Figura 3.13: Curvas de bloqueio da classe 2

O *menor* bloqueio esperado pela classe de serviço 2 pode ser determinado por

$$\lim_{a_1 \to 0} B_2 = Erl(a_2, C - C_1)$$

e caso este valor seja maior que $B^\star$, o sistema com compartilhamento parcial de recursos torna-se infactível para a reserva C_1 escolhida, sendo limitado portanto pelo bloqueio da classe 2. Se possível, menores valores de C_1 devem ser adotados.

Exemplo

Para avaliar a eficácia da proteção seja o exemplo com três classes de serviço, dadas pelo descritor $\hat{D}(a, c)$

$$\hat{D} = \begin{bmatrix} 3 & 4 \\ 6 & 2 \\ 12 & 1 \end{bmatrix} \tag{3.55}$$

O dimensionamento sem reservas $\mathcal{S}^N$ com bloqueio máximo especificado $B^\star = 2\%$ para todas as classes resulta $C = 58$ e bloqueios nominais de $B_1 = 1.75\%$, $B_2 = 0.71\%$ e $B_3 = 0.32\%$. A classe a ser protegida é a classe 1. Para analisar o desempenho do enlace na situação de sobrecarga, considere o tráfego de cada classe definido como

$$\begin{bmatrix} \tilde{a}_1 \\ \tilde{a}_2 \\ \tilde{a}_3 \end{bmatrix} = \begin{bmatrix} 1 + \alpha L_1 & 0 & 0 \\ 0 & 1 + \alpha L_2 & 0 \\ 0 & 0 & 1 + \alpha L_3 \end{bmatrix} \cdot \begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix} \tag{3.56}$$

onde a_i são os valores nominais de tráfego dados em $\hat{D}$ e o fator de sobrecarga é dado pelo parâmetro α. O vetor de sobrecarga L possui componentes $[L_1\ L_2\ \cdots\ L_k]$ tais que

$$L_i = \begin{cases} 1\ ; & \textit{se há sobrecarga em } a_i \\ 0\ ; & \textit{caso contrário} \end{cases} \tag{3.57}$$

Para a proteção da classe 1, é adotada uma política de reserva parcial simples, do tipo *ao menos* com $\sigma_1 = 20$ e com $\sigma_2 = \sigma_3 = 0$. Para estes parâmetros, o valor assintótico do bloqueio para a classe 1 é $B_1|_\infty = 19.9\%$. A Figura 3.14 mostra à esquerda as probabilidades de bloqueio das classes sem reservas em função da sobrecarga das classes 2 e 3. O vetor de sobrecarga correspondente vale $L = [0\ 1\ 1]$. Note o crescimento de B_1 apesar de a_1 permanecer no seu valor nominal. Um incremento $\alpha = 0.5$ resulta em $B_1 = 12.0\%$, $B_2 = 5.5\%$, e $B_3 = 2.6\%$.

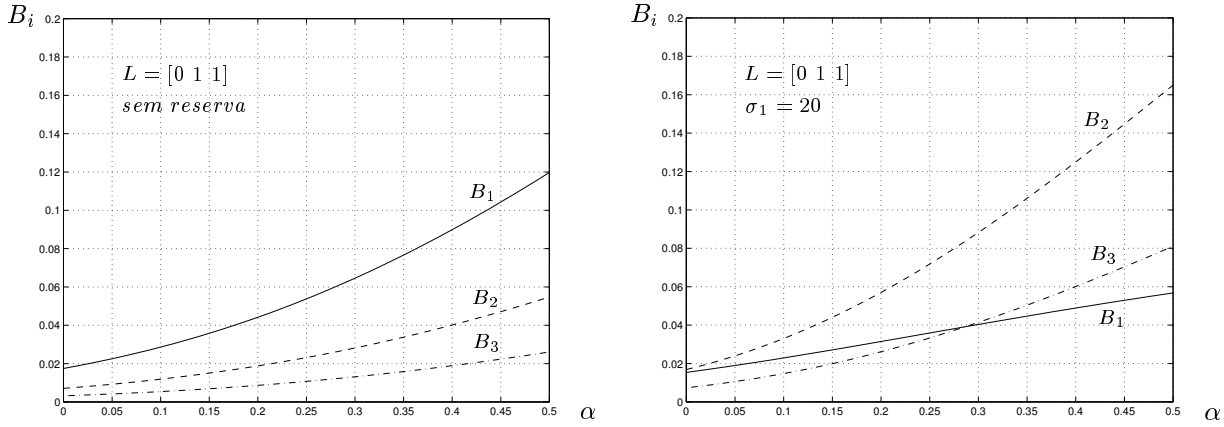

Figura 3.14: Probabilidades de bloqueio em função da sobrecarga de tráfego – Sem reserva (esquerda) e com reserva (direita)

Da mesma forma, o lado direito da Figura 3.14 mostra as probabilidades de bloqueio em função da sobrecarga de tráfego $L = [0\ 1\ 1]$ para o mesmo enlace de capacidade total $C = 58$. O bloqueio nominal ($\alpha = 0$) vale $B_1 = 1.6\%$, $B_2 = 1.8\%$, e $B_3 = 0.75\%$.

Verifica-se apenas uma pequena alteração nos valores do bloqueio nominal da condição de compartilhamento total, ainda dentro do $B^\star$ especificado. Com a sobrecarga das classes 2 e 3, o bloqueio da classe protegida B_1 possui uma variação modesta, quando comparada com o lado esquerdo da Figura 3.14. Para o mesmo $\alpha = 0.5$ tem-se $B_1 = 5.5\%$, $B_2 = 16.5\%$, e $B_3 = 8.0\%$. A estratégia de reserva da classe prioritária também apresenta um comportamento adequado tanto para o caso em que todos os tráfegos experimentem sobrecarga simultâneamente quanto para o caso em que apenas a classe 1 esteja sobrecarregada. A Figura 3.15 ilustra estes casos. No lado esquerdo com o vetor de sobrecarga $L = [1\ 1\ 1]$ nota-se que todas as curvas aumentam de forma similar. O lado direito com o vetor de sobrecarga $L = [1\ 0\ 0]$ mostra que quando apenas a classe protegida está sobrecarregada, as demais não são seriamente penalizadas.

Para a determinação de um valor adequado de reservas, métodos heurísticos baseados em premissas simples podem ser utilizados, como o algoritmo apresentado em [7].

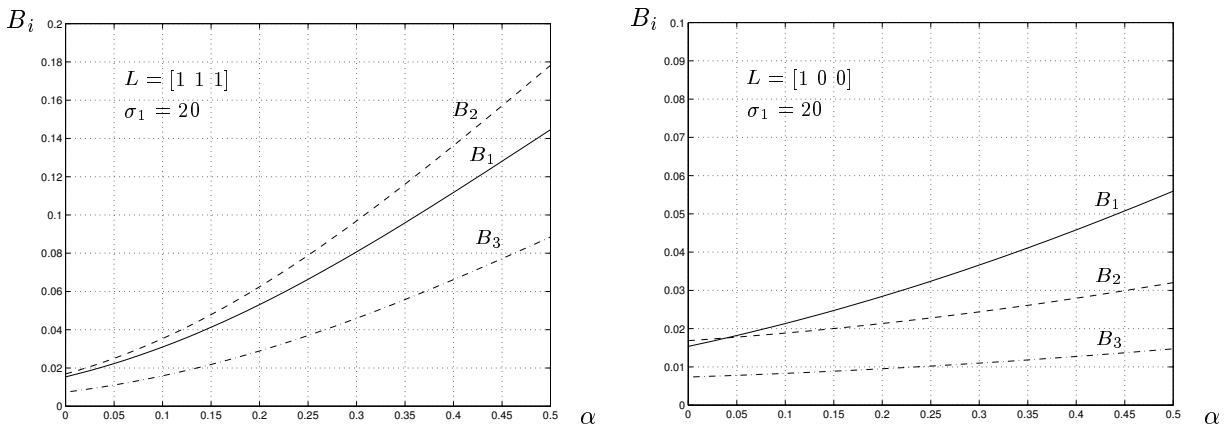

Figura 3.15: Probabilidades de bloqueio em função de sobrecargas diversas de tráfego no sistema com reserva

3.8 Conclusão

Neste capítulo foram discutidos os sistemas multiclasses, para políticas de acesso com e sem reservas, com exemplos ilustrativos.

Dentre as diversas políticas de reserva destacam-se aquelas chamadas coordenada-convexas por serem o tipo mais comum em sistemas reais, além de serem importantes para a validade de alguns métodos de cálculo. Também as políticas de acesso cujos espaços de estado preservam a simetria interna de transições são mais facilmente tratadas pelos métodos analíticos que aquelas que possuem assimetrias de transição.

O bloqueio natural (sem reservas) em sistemas multiclasses sempre apresenta maiores probabilidades de bloqueio para as classes mais exigentes em termos de requisitos de recursos. A variação destes bloqueios com o aumento do tráfego a_i ou do número total C de recursos nem sempre é monotônica, apresentando o fenômeno oscilatório. Este comportamento é inevitável para sistemas com compartilhamento de recursos (total ou parcial) onde as classes possuam requisitos c_i distintos, mas não apresenta grandes conseqüências para o desempenho do sistema. Políticas de reserva total não apresentam tal fenômeno.

Entre os algoritmos de cálculo de bloqueio estudados destaca-se o método iterativo de Kaufman, sendo o método preferencial (análise e síntese) pela eficiência numérica apresentada. Já o cálculo pelo teorema do produto deve ser aplicado apenas para a análise de sistemas com pequeno número de classes que utilizem reservas de recursos, pois ainda que se trate de um método polinomial com relação a C, exibe o comportamento exponencial com relação ao número de classes k. O método baseado na fórmula de LeGall somente deve ser usado quando a estratégia de proteção por limiares de transbordo é adotada na síntese de uma rede.

A utilização de reservas para economia de recursos C no procedimento de síntese deve ser empregada em um sistema onde $c_1 > c_2 > \cdots > c_k$ para a classe mais exigente de recursos. Há bons resultados apenas quando existir uma condição apropriada dos tráfegos das classes tal que satisfaça $a_1 c_1 \gg (a_2 c_2 + \ldots + a_k c_k)$. Nestas condições, o método de reserva $\mathcal{S}^S$ é o mais eficaz dentre os analisados. Caso a condição não seja satisfeita, o dimensionamento sem reservas para bloqueio natural $\mathcal{S}^N$ produz melhores resultados.

A proteção contra sobrecarga de outras classes por meio das políticas de reserva apro-

priadas é mais eficaz para a melhora do desempenho global do sistema. Como método para sistemas multiclasses com requistos c_i distintos, é utilizável com bons resultados somente para as classes mais críticas (i.e. maiores c_i). Bons desempenhos para sobrecargas da ordem de 50% foram obtidos nos exemplos analisados. Quanto às políticas de reserva, o método de reserva simples (estática) $\mathcal{S}^S$ apresenta bons resultados de desempenho. Todavia, o cálculo exato é difícil uma vez que utiliza o teorema do produto. A reserva com limites de transbordo $\mathcal{S}^L$ pode ser usada alternativamente desde que se estabeleçam apenas valores assintóticos para proteção e o cálculo exato do bloqueio para diferentes situações de sobrecarga não seja necessário.

É importante ressaltar a ausência de métodos alternativos com baixa complexidade numérica para cálculo de bloqueios em sistemas com reservas. Algumas propostas para cálculos aproximados existem, mas um algoritmo genérico que cubra várias modalidades de reserva permanece como objetivo de pesquisa no futuro.

Capítulo 4

Bloqueio em Redes

A representação de redes por matrizes é discutida neste capítulo, bem como alguns métodos de encaminhamento associados. O problema de ponto fixo é tratado sob a óptica dos sistemas multiclasses e uma generalização do método é proposta, em conjunto com o cálculo recursivo de Kaufman para bloqueios em enlaces. Por fim, algumas redes simples são analisadas para validação da metodologia.

4.1 Introdução

As requisições de conexões (chamadas) têm de atravessar redes cuja topologia em geral é mais complexa do que uma simples conexão de um enlace. Os diversos saltos ou *hops* que uma chamada atravessa contribuem com uma probabilidade de bloqueio, um tempo de atraso, uma taxa de perda, etc. de modo que a qualidade de serviço (QoS) fim-a-fim é uma combinação de todas estas contribuições individuais. Algumas delas, como por exemplo o atraso máximo (MaxCTD no caso ATM) são aditivas, enquanto outras, como a probabilidade de bloqueio, exigem modelos mais elaborados para seu cálculo preciso, tal como a aproximação de *ponto fixo*. Mesmo o processo de se encontrar um caminho factível através da rede, denominado aqui como *encaminhamento* exige algoritmos bastante sofisticados e eficientes, uma vez que este processo opera na maioria das vezes em tempo real e freqüentemente com informações incompletas sobre o estado da rede. Simplificações do procedimento podem ser adotadas, tais como restringir as opções a uma lista de caminhos previamente estabelecida. A análise do desempenho e o gerenciamento da rede neste caso são facilitados, ainda que à custa da flexibilidade do sistema.

4.1.1 Topologia e tráfego

Uma *rede multisserviços* que seja composta por N nós e S enlaces físicos multiclasses pode ser representada por um grafo (N, S) onde N são os respectivos vértices e S as correspondentes arestas. Cada s-ésimo enlace possui uma capacidade C^s e interconecta dois nós (x, y) conforme uma dada topologia que é assumida *invariante* no tempo. Desta forma a representação topológica da rede toma a forma de uma matriz de adjacência

$$\mathcal{R}^{N\times N}(x,y) = \begin{cases} C^s \ ; & x, y \ adjacentes; \\ 0 \ ; & caso \ contrário; \end{cases} \tag{4.1}$$

onde $s = 1, 2, \ldots, S$ e lista todos os enlaces da rede. A matriz de adjacência é simétrica ($\mathcal{R}(x, y) = \mathcal{R}(y, x)$) e possui a diagonal principal nula ($\mathcal{R}(x, x) = 0$). Os enlaces de interconexão são supostos bidirecionais e com a mesma capacidade C^s em ambos os sentidos.

O tráfego multisserviços composto por k classes, oferecido em cada nó x da rede tendo como destino outro nó y também pertencente à mesma é representado pelas *matrizes de tráfego* origem-destino A_i, com componentes

$$A_i = a_i(x, y) \; ; \quad \begin{cases} i = 1, 2, \ldots, k \\ x, y \in \{1, 2, \ldots, N\} \end{cases} \tag{4.2}$$

As k matrizes A_i existentes, uma para cada classe de serviço, têm as mesmas dimensões e são independentes, podendo conter elementos nulos caso aquela classe não apresente interesse de tráfego entre um par de nós.

4.2 Encaminhamento

Um determinado elemento de tráfego denotado por um par (x, y) comporta com freqüência diversos caminhos possíveis para que uma potencial conexão possa transmitir os dados através da rede. A decisão de quais caminhos devem ser adotados é feita por um processo geral conhecido como *encaminhamento*, que é responsável por criar um conjunto finito de caminhos não cíclicos para todo elemento de tráfego da matriz $A_i(x, y)$. Este conjunto resultante é designado por $\mathcal{P}$ e contém por hipótese todos os caminhos *admissíveis* através da rede para todos os pares dados de origem-destino e que respeitem uma QoS dada. O número de elementos deste conjunto pode ser bastante elevado dependendo da topologia da rede, o que motiva com freqüência o emprego de seleções arbitrárias de caminhos admissíveis visando limitar o número total de opções que se apresentam, a critério do planejador do sistema.

Por outro lado, a qualidade de serviço a ser atendida numa única conexão (ou em um grupo de conexões) especifica uma série de *restrições* adicionais ao conjunto de caminhos possíveis. Essas restrições podem ser de duas espécies:

- **Restrições de enlace:** Impõem limites na utilização de enlaces individuais. Por exemplo, uma restrição de banda que é imposta a uma conexão requer uma largura de banda mínima de cada um dos enlaces a ser utilizado. Apenas aqueles enlaces capazes de atender aos requisitos da conexão podem tomar parte na composição de caminhos.

- **Restrições de caminho:** Impõem limites na utilização de caminhos completos. Por exemplo, o atraso fim-a-fim de uma conexão não pode exceder um certo máximo, que é totalizado para diversos caminhos alternativos. Somente aqueles caminhos que respeitem as restrições podem ser considerados como escolhas factíveis.

Restrições e regras adicionais de encaminhamento são acrescentadas ao processo genérico que passa a ser denominado encaminhamento *com restrições* ou *estruturado*. Este processo permite a priorização ou mesmo exclusão de cada uma das diferentes escolhas contidas no grupo $\mathcal{P}_{x,y}$ de caminhos *possíveis* para um certo elemento de tráfego (x, y), o que produz

um subconjunto de caminhos *factíveis* designado por $\mathcal{F}_{x,y}$. Todos estes subconjuntos para todos os pares (x, y) são agrupados no conjunto total de caminhos factíveis $\mathcal{F}$.

Restrições comumente especificadas são por exemplo: atraso fim-a-fim, largura de banda disponível, taxa de perda de pacotes/células, probabilidade de bloqueio, custo de uso do canal de transmissão. A Figura 4.1 mostra dois exemplos de caminhos $\mathcal{F}_{1,6}$ entre os nós $1 \leftrightarrow 6$ numa rede e que simultaneamente atendam às especificações dadas [13].

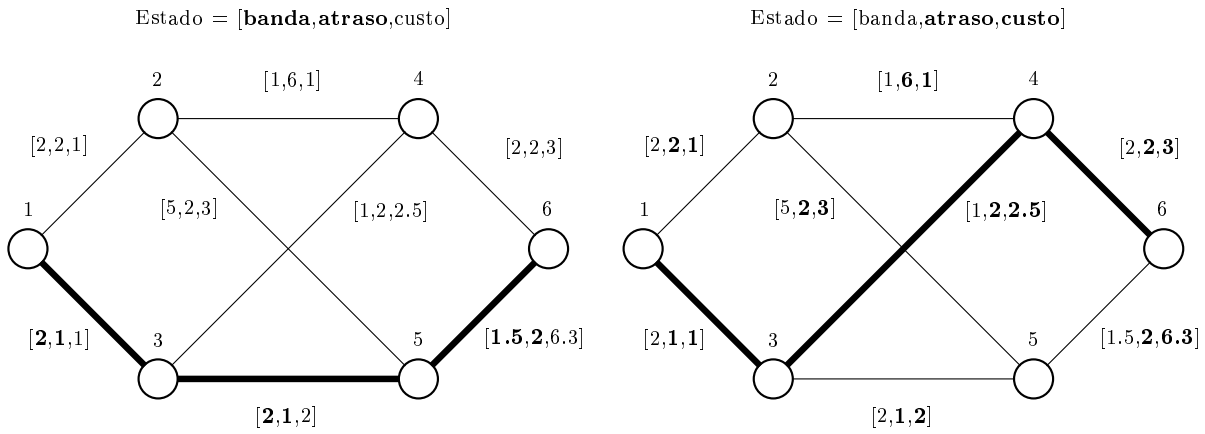

Figura 4.1: Caminhos determinados com requisitos dados: (esquerda) banda ≥ 1.5 e atraso ≤ 5; (direita) custo $= min$ e atraso ≤ 5

Embora as restrições apresentadas no exemplo resultem em apenas um elemento (caminho) em cada um dos dois conjuntos, o cenário mais freqüente é aquele em que diversas opções persistam dentro de cada um dos grupos $\mathcal{F}_{x,y}$.

4.2.1 Processos de encaminhamento

Existem dois grandes grupos de métodos de encaminhamento. Caso o conjunto total $\mathcal{F}$ não mude com o passar do tempo, o processo é denominado *encaminhamento estático*. O conjunto de caminhos é obtido por meio de algoritmos de cálculo ou heurísticas executados pelo planejador ou gerente da rede *a priori* e cuja atualização é apenas circunstancial ou mesmo sazonal. Caso $\mathcal{F}$ seja função de cada requisição individual de acesso à rede o processo é chamado de *encaminhamento dinâmico*. Neste caso, algoritmos automatizados que atuem em tempo real de operação da rede se fazem necessários.

As estratégias de encaminhamento empregadas em cada caso são agrupadas em três tipos gerais, dependendo de como a informação a respeito do estado da rede é mantida e como a procura de caminhos factíveis se processa:

- **Encaminhamento na origem:** Cada nó mantém o estado global completo, incluindo a topologia da rede e o estado de cada enlace. Os caminhos factíveis são computados localmente no nó de origem, e no caso de encaminhamento dinâmico, mensagens contendo variáveis de estado local são enviadas para todos os demais nós da rede para a atualização do estado global;

- **Encaminhamento distribuído:** Os caminhos factíveis são computados por meio de um algoritimo distribuído. A informação de estado de cada nó é utilizada para o cálculo e apenas mensagens de controle são trocadas pelos nós;

- **Encaminhamento hierárquico:** Os nós são agregados em grupos que por sua vez são agregados recursivamente em outros grupos de nível hierárquico superior, criando uma estrutura com múltiplos níveis. Cada nó mantém informações detalhadas sobre o estado global do próprio grupo e informações genéricas agregadas sobre os demais.

Para redes com reduzido número de nós, procedimentos heurísticos podem ser utilizados na determinação dos caminhos factíveis no caso do encaminhamento estático. Redes maiores ou encaminhamento dinâmico exigem algoritmos mais elaborados [13]. Algoritmos tradicionais, como os de Dijkstra [20] e Bellman-Ford [6] são empregados com freqüência nestes cálculos. Dentre os algoritmos recentes de encaminhamento na origem pode-se citar os de Ma-Steenkiste [34] e Guérin-Orda [28] para restrição de banda, Wang-Crowcroft [49] para restrição de banda mais atraso e Chen-Nahrstedt [12] para banda e custo. No caso do encaminhamento distribuído destacam-se os Shin-Chou [47] para atraso, de Salama [46] para atraso e custo mínimos, e Cidon [17] como genérico. Para o encaminhamento hierárquico emprega-se o PNNI do ATM Forum [4].

4.2.2 Listas enumeradas de encaminhamento

O conjunto $\mathcal{F}$ de todos os subconjuntos de caminhos factíveis da rede pode ter os seus elementos listados de uma forma alternativa para uso no cálculo de ponto fixo através de um mapeamento que enumere seqüencialmente todos os caminhos para todos os pares (x, y), produzindo um outro conjunto unidimensional de tamanho L igual ao número total de caminhos, com cujos elementos ℓ mantém uma correspondência biunívoca tal que se $\mathcal{F}_{x,y} \subset \mathcal{F}$, para todos os pares (x, y) existentes, então

$$\mathcal{F} \Leftrightarrow \{\ell\} \tag{4.3}$$

Cada um dos elementos de $\mathcal{F}$ recebe então um número de seqüência ℓ. A matriz de tráfego origem-destino $A(x, y)$ pode ser reescrita portanto como uma matriz de tráfego denominada como *enumerada por caminhos factíveis* A^{ℓ}, onde os diversos tráfegos oferecidos pelas classes existentes são enviados no interior da rede pelos L caminhos disponíveis enumerados.

Esta representação dos caminhos ℓ através da rede se faz por meio de uma *matriz de incidência* $I^{S \times L}$ definida como

$$I_{s,\ell} = \begin{cases} 1\ ; & s \in \ell; \\ 0\ ; & \textit{caso contrário}; \end{cases} \tag{4.4}$$

para todo s-ésimo enlace que pertença à rede.

Para a utilização de cada conjunto $\mathcal{F}_{x,y}$ de caminhos factíveis alternativos entre um par (x, y) da rede pelo menos duas maneiras distintas podem ser adotadas.

Balanceamento de carga

Um encaminhamento estático sem transbordos de qualquer espécie, mas cujas alternativas factíveis para um elemento de tráfego recebam percentuais para o escoamento de parcelas do tráfego total é denominado de *balanceamento de carga* ou *load sharing*. Para o conjunto

de caminhos $\mathcal{F}$ da rede o mapeamento que enumera cada um dos pares origem-destino (x, y) produz listas de caminhos ℓ para este par tais que

$$\mathcal{F}_{x,y} \Leftrightarrow \{\ell_{x,y}^{\ r}\} \tag{4.5}$$

O índice x, y é associado a todo o tráfego que tem origem em x e destino em y, não levando em conta as possíveis diversidades de caminho. Este índice é único para cada *interesse de tráfego*. O índice r por sua vez enumera os diferentes caminhos permitidos dentro de um certo conjunto $\{\ell_{x,y}^1, \ell_{x,y}^2, \ldots\}$. Cada um destes r-ésimos caminhos pode receber uma parcela de tráfego $\tau_{x,y}^{\ r}$ denominada de *partição de carga* para o encaminhamento $\mathcal{F}_{x,y}$. Decorre também desta definição que

$$\sum_r \tau_{x,y}^{\ r} = 1; \tag{4.6}$$

para qualquer um dos interesses de tráfego x, y existentes.

Transbordo

Uma rede que opere segundo um encaminhamento estático mas com caminhos alternativos para requisições que sejam bloqueadas em seu caminho de primeira escolha é chamado de *transbordo* ou *overflow*. À semelhança do balanceamento de carga, diversos caminhos são oferecidos para cada interesse de tráfego x, y, mas ao contrário daquele, apenas uma das alternativas poderá ser utilizada de cada vez. Seja um total de v alternativas possíveis, para as quais é feita uma ordenação segundo um critério de preferência pré-estabelecido. A representação dos caminhos a partir do subconjunto $\mathcal{F}_{x,y}$ resultante mapeia todos os pares origem-destino produzindo listas de caminhos ℓ tais que

$$\mathcal{F}_{x,y} \Leftrightarrow \{\ell_{x,y}^{\ v}\} \tag{4.7}$$

Da mesma forma que no balanceamento de carga, o índice x, y é associado a todo o tráfego que tem origem em x e destino em y, não levando em conta as possíveis diversidades de caminho. O índice v por sua vez enumera os diferentes caminhos alternativos para um certo conjunto $\{\ell_{x,y}^1, \ell_{x,y}^2, \ldots\}$. Cada um destes caminhos possui uma prioridade seqüencial $(v_1, v_2, \ldots)$ de modo que v_q só poderá ser ocupado se v_{q-1} estiver bloqueado. A requisição de conexão só será finalmente bloqueada se *todas* as alternativas para um dado par (x, y) estiverem tomadas

$$\ell_{x,y}^{\ 1} \rightarrow \ell_{x,y}^{\ 2} \rightarrow \cdots \rightarrow \ell_{x,y}^{\ v} \rightarrow \textit{bloqueio} \tag{4.8}$$

O processo de transbordo é seqüencial, ocupando o primeiro caminho livre encontrado e, esgotadas as possibilidades, não leva em consideração novamente caminhos já tentados. Embora certos procedimentos possam ser estabelecidos para reconsiderar alternativas já tentadas previamente, o usual é testar as opções de transbordo apenas uma vez.

4.3 Ponto fixo

Uma maneira de estender os resultados de bloqueio computáveis no enlace para uma rede complexa é através do método de cálculo conhecido como *ponto fixo*. Através deste processo obtém-se resultados com boa precisão quando os processos de requisição de conexões são poissonianos e operam em balanceamento de carga.

4.3.1 Formulação

Seja uma rede composta por N nós e S enlaces, com uma topologia definida pela matriz de adjacência $\mathcal{R}^{N\times N}$. Para o conjunto $\mathcal{P}$ dos caminhos possíveis não-cíclicos aplica-se um processo de encaminhamento estático para a obtenção do conjunto de todos os caminhos factíveis $\mathcal{F}$. Apenas o balanceamento de carga é admissível para cada um dos $\mathcal{F}_{x,y}$. Os caminhos são enumerados conforme o critério dado pela relação (4.3) em elementos ℓ. Constrói-se a matriz A^ℓ correspondente a partir da matriz de tráfego dada. A matriz de incidência da definição (4.4) é usada para representar o conjunto de caminhos $\{\ell\}$ através da rede.

Assume-se que a probabilidade de bloqueio de caminho $\mathcal{B}^\ell$ seja resultado da probabilidade de que *pelo menos um* dos enlaces $s \in \ell$ esteja bloqueado e, além disso, que estes eventos sejam todos independentes. Então a expressão

$$\mathcal{B}_i^\ell = 1 - \prod_{s\in\ell} [1 - B_i^s] \tag{4.9}$$

é válida para um certo caminho ℓ admissível na rede onde B_i^s é a probabilidade de bloqueio para a i-ésima classe de serviço presente no enlace s pertencente àquele caminho. Esta equação pode ser reescrita com auxílio dos elementos da matriz de incidência $I_{s,\ell}$ da forma

$$\mathcal{B}_i^\ell = 1 - \prod_{s=1}^{S} [1 - B_i^s]^{I_{s,\ell}} \tag{4.10}$$

O tráfego que é efetivamente escoado via caminho ℓ, denominado $\bar{A}^\ell$ é uma função do bloqueio $\mathcal{B}^\ell$ dado pela equação (4.10) e cujos elementos são definidos pela seguinte relação com os elementos da matriz de tráfego A^ℓ dada

$$\bar{A}_i^\ell = [1 - \mathcal{B}_i^\ell]\ A_i^\ell \tag{4.11}$$

onde cada i-ésima classe presente num caminho ℓ é considerada individualmente. Os diversos enlaces que fazem parte dos caminhos têm seus bloqueios calculados por algum dos métodos conhecidos para sistemas multiclasses.

Seja agora o tráfego fictício $a_i^{s,\ell}$ oferecido ao correspondente enlace s, devido a um caminho ℓ, para a i-ésima classe. O tráfego escoado associado $\bar{a}_i^{s,\ell}$ pode ser representado em função do bloqueio sob a forma

$$\bar{a}_i^{s,\ell} = [1 - B_i^s]\ a_i^{s,\ell} \tag{4.12}$$

onde ambos os membros são escalares. Considere-se um caminho ℓ qualquer que transporte um feixe composto pelas diferentes classes de serviço presentes na rede. Os elementos da matriz de tráfego total $\bar{A}_i^\ell$ escoado pelo caminho deverão ser, pela hipótese de continuidade de fluxo do tráfego através da rede, os mesmos que fluem em cada um dos enlaces que constituem o caminho. Assim cada escalar $\bar{a}_i^{s,\ell}$ vale

$$\bar{a}_i^{s,\ell} = \bar{A}_i^\ell; \quad s \in \ell \tag{4.13}$$

A relação de continuidade acima impõe que nenhum tráfego *escoado* desaparece ao longo de qualquer caminho ℓ. Tomando a equação (4.13) para todos os caminhos existentes que

passam pelo enlace s e, com auxílio dos elementos da matriz de incidência $I_{s,\ell}$ para incluir a condição $s \in \ell$, vem que

$$\sum_{\ell} \bar{a}_i^{s,\ell} = \sum_{\ell} I_{s,\ell}\ \bar{A}_i^{\ell} \tag{4.14}$$

e substituindo os tráfegos escoados pelos seus valores em função dos bloqueios nas equações (4.11) e (4.12) produz para a_i^s

$$\sum_{\ell} [1 - B_i^s]\ a_i^{s,\ell} = \sum_{\ell} I_{s,\ell}\ A_i^{\ell}\ [1 - \mathcal{B}_i^{\ell}] \quad \Longrightarrow$$

$$a_i^s = [1 - B_i^s]^{-1} \sum_{\ell} I_{s,\ell}\ A_i^{\ell}\ [1 - \mathcal{B}_i^{\ell}] \tag{4.15}$$

Com o auxílio da equação (4.10) elimina-se a dependência explícita do bloqueio de caminho $\mathcal{B}^{\ell}$ e a equação (4.15) toma a seguinte forma

$$a_i^s = [1 - B_i^s]^{-1}\ \sum_{\ell} I_{s,\ell}\ A_i^{\ell}\ \prod_{s=1}^{S} [1 - B_i^s]^{I_{s,\ell}} \tag{4.16}$$

Na equação anterior (4.16) os bloqueios B^s para os enlaces são função dos respectivos tráfegos a_i^s

$$B^s = \mathcal{K}\left\{a_1^s, \ldots, a_k^s; c_1, \ldots, c_k; C^s\right\} \tag{4.17}$$

onde $B^s = [B_1^s, B_2^s, \cdots, B_k^s]$ é o vetor dos bloqueios para todas as k classes no s-ésimo enlace e c_i são os requisitos de recursos das classes. A função $\mathcal{K}\{\cdot\}$ pode ser qualquer método de cálculo para bloqueio multisserviços que obedeça às regras locais estabelecidas (reservas, limiares, etc.) do enlace s de capacidade dada C^s.

O sistema não-linear formado pelas equações (4.16) e (4.17) é conhecido como *equações de ponto fixo*. Desde que se disponha de um método adequado de resolução, este sistema possibilita, para todos os s enlaces da rede, o cálculo dos valores B_i^s e a_i^s, conhecidos como *bloqueio de enlace* e *tráfego fictício* para carga reduzida, respectivamente. A partir da obtenção destes resultados e com a utilização das equações (4.10), (4.11) calcula-se o valor de $\mathcal{B}_i^{\ell}$, dito *bloqueio fim-a-fim* de um caminho ℓ, e seu respectivo tráfego escoado $\bar{A}_i^{\ell}$ para a i-ésima classe pertencente àquele caminho. O valor do tráfego realmente escoado $\bar{a}_i^s$ por enlace é computado por meio da equação (4.12).

4.3.2 Métodos de cálculo

O sistema não linear de equações de Ponto Fixo não admite soluções analíticas, de maneira que resoluções por métodos numéricos devem ser empregadas. Dentre os métodos usados destacam-se três deles.

Método de Newton

Esta técnica utiliza o desenvolvimento de primeira ordem do vetor $\mathbb{F}(X)$ próximo ao ponto atual X_0 para computar a nova estimativa. Cada componente do vetor pode ser expandida numa série de potências, que expressa em notação matricial se torna

$$\mathbb{F}(X) = \mathbb{F}(X_0) + \mathbb{J}\ [X - X_0] \tag{4.18}$$

onde $\mathbb{J}$ consiste na matriz jacobiana, a qual determina o gradiente a ser utilizado no cálculo. Reescrevendo de modo a explicitar a solução para a variável vetor X

$$X = \mathbb{J}^{-1}[\mathbb{F}(X) - \mathbb{F}(X_0)] + X_0 \tag{4.19}$$

Impondo que a função $\mathbb{F}$ valha zero no ponto de solução e adequando os índices para tornar aparente a recursividade da operação

$$X_{(n+1)} = -\mathbb{J}^{-1}\mathbb{F}(X_{(n)}) + X_{(n)} \tag{4.20}$$

O procedimento em si consiste no uso de uma estimativa inicial $X_{(0)}$ para o vetor X que é empregada no cálculo da primeira iteração; em seguida, o valor obtido de X é utilizado como novo valor na próxima iteração, e assim sucessivamente até a convergência.

Para o sistema de equações de ponto fixo, a função $\mathbb{F}$ a ser utilizada é obtida pela substituição das equações (4.16) em (4.17). Esta equação resultante é expressa na forma $f = 0$ e igualada à função de cálculo $\mathbb{F}$ tal que

$$\mathbb{F}_i^s(B_i^s) = \mathcal{K}\{a_i^s(B_i^s); c_i; C^s\} - B_i^s \tag{4.21}$$

Para uma única classe de serviço, a matriz jacobiana $J^{S\times S}$ é descrita por

$$J(s_x, s_y) = \left[\frac{\partial \mathbb{F}^{s_x}}{\partial B^{s_y}}\right] \tag{4.22}$$

No caso multiclasses, cada combinação possível de pares de classes produz uma matriz jacobina expandida, cujas dimensões têm tamanho dado pelo produto de k classes por S enlaces resultando em uma matriz ampliada $\mathbb{J}^{kS\times kS}$ composta por $(k)^2$ matrizes $J^{S\times S}$ tal que

$$\mathbb{J} = \begin{bmatrix} J_{11}(s_x, s_y) & J_{12}(s_x, s_y) & \cdots & J_{1k}(s_x, s_y) \\ J_{21}(s_x, s_y) & J_{22}(s_x, s_y) & \cdots & J_{2k}(s_x, s_y) \\ \vdots & \vdots & & \vdots \\ J_{k1}(s_x, s_y) & J_{k2}(s_x, s_y) & \cdots & J_{kk}(s_x, s_y) \end{bmatrix} \tag{4.23}$$

Na matriz jacobiana $\mathbb{J}$ (k classes em S enlaces) os elementos linha $(i_x s_x)$ e coluna $(i_y s_y)$ são associados a todas as combinações possíveis de acoplamento entre classes e enlaces e podem ser expressos como

$$\mathbb{J}(i_x s_x, i_y s_y) = \left[\frac{\partial \mathbb{F}_{i_x}^{s_x}}{\partial B_{i_y}^{s_y}}\right] \tag{4.24}$$

que é a matriz a ser empregada no cálculo da equação (4.20).

A principal vantagem do método de Newton é a rápida convergência, quadrática quando na vizinhança da solução. A desvantagem do método advém da necessidade do recálculo e inversão da matriz jacobiana da equação (4.24) a cada passo da iteração. Além disso, as variáveis devem permanecer limitadas ao intervalo $(0 \leq X \leq 1)$ durante o processo iterativo, sob risco de não haver convergência, o que pode requerer o uso de alterações adicionais ao método [25].

Minimização

Este método consiste no emprego de técnicas de minimização não-linear da função objetivo $Z(X)$ tal que satisfaça o critério

$$\min Z = \sum_i \sum_s \mathbb{F}_i^s(X)^2 \; ; \qquad 0 \leq X \leq 1 \tag{4.25}$$

Tal solução, caso exista, será encontrada quando $\mathbb{F}_i^s(X) = 0$ que é a equação original desejada. A vantagem deste método é a possibilidade de utilização de técnicas numéricas eficientes de minimização para rápida convergência, onde as únicas restrições sejam os limites nas variáveis da função objetivo [25]. Por outro lado, a eficiência computacional deste método quando empregado no cálculo das equações de ponto fixo é baixa, limitando a aplicação a sistemas cujo número de variáveis seja pequeno.

Relaxação

O método mais comum e com menores requisitos computacionais é o método da *relaxação*, também conhecido como *substituição repetida*. A forma original do problema não-linear de ponto fixo $X = \mathbb{F}(X)$ permite que os resultados obtidos pela substituição de um $X_{(0)}$ inicial sejam substituídos recursivamente até que a convergência $X_{(m)}$ seja atingida

$$\begin{aligned} X_{(1)} &:= \mathbb{F}(X_{(0)}); \\ X_{(2)} &:= \mathbb{F}(X_{(1)}); \\ &\vdots \\ X_{(m)} &:= \mathbb{F}(X_{(m-1)}); \end{aligned}$$

O valor inicial pode ser tomado como $X_{(0)} = 0$ para os problemas de ponto fixo onde $X = B^s$. A vantagem deste método é que o cálculo e inversão de matrizes jacobianas (gradientes) não são necessários. Também garante que o vetor de variáveis permaneça limitado no intervalo $0 \leq X \leq 1$. A desvantagem é que a convergência nem sempre é garantida, exceto quando os autovalores de $\mathbb{J}$ forem todos positivos no ponto-solução ou quando a função $\mathbb{F}$ for um mapa de contração [25]. Para a maioria dos casos práticos contudo, este método pode ser aplicado sem que apresente problemas de convergência.

4.4 Cálculos de redes

O método de ponto fixo como ferramenta de análise permite avaliar o desempenho de redes multisserviços do tipo SVC com resultados bastante satisfatórios. Para tanto, é feito uso do próprio método modificado para sistemas multiclasses em conjunto com uma função $\mathcal{K}$ de cálculo de bloqueio eficiente, como o algoritmo recursivo de Kaufman. O procedimento de computação utilizado foi implementado com uso do software *MATLAB*. Vide apêndice para validação numérica com o exemplo de rede apresentado em [16]. A resolução pelo método de relaxação aplicada ao problema de ponto fixo multiclasses também apresentou boa estabilidade numérica para convergência.

4.4.1 Rede simples em Y

Seja a rede simples da Figura 4.2 composta por três enlaces em "Y", com duas classes de serviço com requisitos $c_1 = 4$ e $c_2 = 1$.

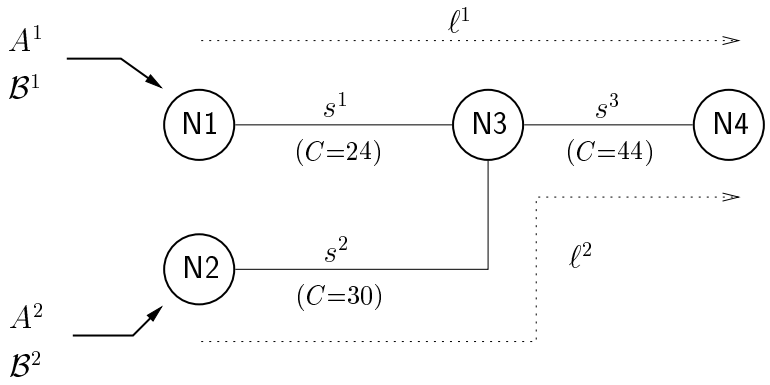

Figura 4.2: Rede simples em Y e duas classes

O conjunto de caminhos factíveis $\mathcal{F}$ é composto por dois caminhos que são enumerados $\ell^1 = \{N1 \to N3 \to N4\}$ e $\ell^2 = \{N2 \to N3 \to N4\}$, gerando a matriz de incidência I dada por

$$I = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 1 \end{bmatrix} \tag{4.26}$$

Há tráfego da primeira classe de 3 erlangs apenas pelo segundo caminho e tráfego da segunda de 16 erlangs pelo primeiro caminho. As matrizes de tráfego ordenadas por caminhos correspondentes são descritas como

$$A^1 = \begin{bmatrix} 0 \\ 16 \end{bmatrix} \qquad A^2 = \begin{bmatrix} 3 \\ 0 \end{bmatrix} \tag{4.27}$$

Os enlaces s^1, s^2, s^3 têm as respectivas capacidades $C^1 = 24$, $C^2 = 30$ e $C^3 = 44$, sem reservas ou privilégios de acesso para nenhuma classe. As equações de ponto fixo obtidas a partir das expressões (4.16) e (4.17) são

$$\begin{aligned}
a_1^1 &= 0 \\
a_2^1 &= (1 - B_2^1)^{-1}[16(1 - B_2^1)(1 - B_2^3)] \\
a_1^2 &= (1 - B_1^2)^{-1}[\ 3(1 - B_1^2)(1 - B_1^3)] \\
a_2^2 &= 0 \\
a_1^3 &= (1 - B_1^3)^{-1}[\ 3(1 - B_1^2)(1 - B_1^3)] \\
a_2^3 &= (1 - B_2^3)^{-1}[16(1 - B_2^1)(1 - B_2^3)] \\
[B_1^1, B_2^1] &= \mathcal{K}\{\ \ 0, a_2^1\ ;\ 4, 1\ ;\ 24\ \} \\
[B_1^2, B_2^2] &= \mathcal{K}\{\ a_1^2,\ 0\ ;\ 4, 1\ ;\ 30\ \} \\
[B_1^3, B_2^3] &= \mathcal{K}\{\ a_1^3, a_2^3\ ;\ 4, 1\ ;\ 44\ \}
\end{aligned}$$

Os resultados para bloqueio fim-a-fim $\mathcal{B}_i$ e tráfego escoado $\bar{A}_i$ por classe de serviço são mostrados na Tabela 4.1

Tabela 4.1: Resultados fim-a-fim para rede da Figura 4.2

Caminho	$\mathcal{B}_1$	$\mathcal{B}_2$	$\bar{A}_1$	$\bar{A}_2$
ℓ^1	–	2.1 %	–	15.67
ℓ^2	5.5 %	–	2.84	–

Os resultados foram calculados pelo método de relaxação. O sistema de equações converge após 6 iterações, para uma tolerância relativa máxima de 1 ppm entre passos consecutivos (`reltol` $\leq 10^{-6}$). O número de operações executadas pelo algoritmo é 29812 flops, incluindo a computação dos bloqueios pelo método de Kaufman.

Os resultados para tráfego escoado $\bar{a}_i$ e bloqueio por enlace B_i por classe de serviço são mostrados na Tabela 4.2.

Tabela 4.2: Resultados por enlace para rede da Figura 4.2

Enlace	B_1	B_2	$\bar{a}_1$	$\bar{a}_2$
s^1	–	1.4 %	–	15.67
s^2	1.9 %	–	2.84	–
s^3	3.7 %	0.7 %	2.84	15.67

Devido a interação entre os tráfegos, o enlace de maior bloqueio para uma classe não é necessariamente o mesmo para outra classe com requisitos distintos, conforme pode ser observado para s^3 na Tabela 4.2. A continuidade dos tráfegos fim-a-fim $\bar{A}_i$ através da rede também pode ser constatada comparando-se as duas tabelas.

4.4.2 Rede com caminhos diversos

Redes mais elaboradas e com várias opções de caminhos para um mesmo interesse de tráfego são tratadas adequadamente desde que a resolução de alternativas seja feita apenas por balanceamento de carga (i.e. sem transbordo). A Figura 4.3 ilustra um exemplo de rede, onde coexistem três classes de serviço com quatro interesses de tráfego distintos.

As classes presentes no sistema possuem requisitos $c_1 = 8$, $c_2 = 4$ e $c_3 = 1$. A capacidade de cada enlace é dada pelo vetor

$$C^s = \left[\begin{array}{cccccccccc} 32 & 64 & 64 & 64 & 64 & 96 & 96 & 96 & 96 \end{array} \right]$$

Os interesses de tráfego enumerados por caminhos ℓ são descritos pelas matrizes A que contêm os tráfegos das classes

$$A^1 = \left[\begin{array}{c} 0 \\ 6 \\ 32 \end{array} \right] \quad A^2 = \left[\begin{array}{c} 2 \\ 8 \\ 16 \end{array} \right] \quad A^3 = \left[\begin{array}{c} 5 \\ 0 \\ 0 \end{array} \right] \quad A^4 = \left[\begin{array}{c} 0 \\ 0 \\ 20 \end{array} \right]$$

O processo de encaminhamento por si é um procedimento além do escopo deste trabalho, podendo tornar-se bastante complexo dependendo das restrições escolhidas. Suponha-se,

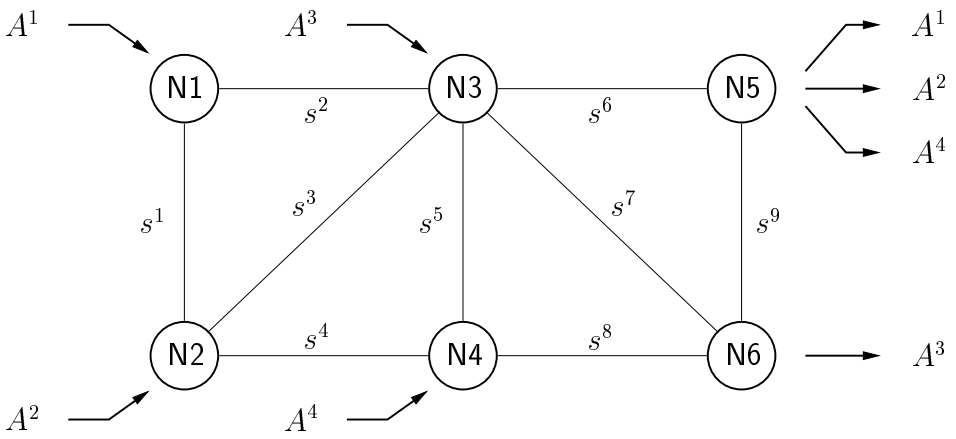

Figura 4.3: Rede com 4 interesses de tráfego e 3 classes

sem perda de generalidade, que um total de nove caminhos enumerados factíveis foi obtido por um algoritmo qualquer e são os únicos definidos para a rede. Isto gera a correspondente matriz de incidência I

$$
\begin{aligned}
\ell^1 &= \{s^2, s^6\} \\
\ell^2 &= \{s^2, s^7, s^9\} \\
\ell^3 &= \{s^1, s^4, s^8, s^9\} \\
\ell^4 &= \{s^3, s^6\} \\
\ell^5 &= \{s^4, s^8, s^9\} \\
\ell^6 &= \{s^7\} \\
\ell^7 &= \{s^5, s^8\} \\
\ell^8 &= \{s^5, s^6\} \\
\ell^9 &= \{s^8, s^9\}
\end{aligned}
\qquad
I = \begin{bmatrix}
0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 \\
1 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \\
0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 \\
0 & 1 & 1 & 0 & 1 & 0 & 0 & 0 & 1
\end{bmatrix}
$$

Os quatro conjuntos de caminhos factíveis $\mathcal{F}_{x,y}$ de um par origem-destino de cada interesse de tráfego, bem como a partição de carga associada (arbitrária neste exemplo), são descritos por

$$
\begin{aligned}
\mathcal{F}_{1,5} &= \{\ell^1, \ell^2, \ell^3\} &\Longrightarrow\quad \tau_{1,5} &= \{0.5, 0.3, 0.2\} \\
\mathcal{F}_{2,5} &= \{\ell^4, \ell^5\} &\Longrightarrow\quad \tau_{2,5} &= \{0.5, 0.5\} \\
\mathcal{F}_{3,6} &= \{\ell^6, \ell^7\} &\Longrightarrow\quad \tau_{3,6} &= \{0.8, 0.2\} \\
\mathcal{F}_{4,5} &= \{\ell^8, \ell^9\} &\Longrightarrow\quad \tau_{4,5} &= \{0.5, 0.5\}
\end{aligned}
$$

A resolução do sistema de ponto fixo pelo método da relaxação converge após 9 iterações, para uma tolerância relativa de 1 ppm entre passos consecutivos ($\texttt{reltol} \leq 10^{-6}$). O número de operações executadas é 719010 flops, incluindo também o cálculo das probabilidades de bloqueio por Kaufman. A escolha optimal da partição de carga, se desejada, pode se tornar um procedimento complexo, envolvendo a solução do problema para diversas escolhas distintas de $\tau_{x,y}$ e seleção dos melhores resultados. Para efeito do exemplo apresentado considere-se apenas os valores dados como finais e não sujeitos a nenhum tipo de re-dimensionamento adicional.

Os resultados para bloqueio fim-a-fim $\mathcal{B}_i$ e tráfego escoado $\bar{A}_i$ por classe de serviço para

cada um dos conjuntos de caminhos $\mathcal{F}_{x,y}$ definidos são dados pela Tabela 4.3.

Tabela 4.3: Resultados fim-a-fim para rede da Figura 4.3

Caminhos	$\mathcal{B}_1$	$\mathcal{B}_2$	$\mathcal{B}_3$	$\bar{A}_1$	$\bar{A}_2$	$\bar{A}_3$
$\mathcal{F}_{1,5}$	–	5.2 %	1.1 %	–	2.84	15.83
$\mathcal{F}_{2,5}$	9.9 %	3.7 %	0.8 %	1.81	7.70	11.91
$\mathcal{F}_{3,6}$	1.2 %	–	–	4.94	–	–
$\mathcal{F}_{4,5}$	–	–	0.4 %	–	–	19.91

Não existem reservas para nenhuma das classes em qualquer um dos enlaces utilizados pelos caminhos ℓ. Os bloqueios fim-a-fim para interesses de tráfegos com caminhos alternativos têm seu valor calculado levando em conta os pesos dados pelas partições de carga τ utilizadas no subconjunto $\mathcal{F}_{x,y}$ correspondente.

Os resultados dos bloqueios computados por classe de serviço por enlace B_i^s e tráfego escoado a_i^s são apresentados na Tabela 4.4. Também neste caso a continuidade do tráfego escoado se preserva, embora seja menos evidente que no caso da rede simples. Note que nem todo enlace possui tráfego de todas as classes de serviço.

Tabela 4.4: Resultados por enlace para rede da Figura 4.3

Enlaces	B_1	B_2	B_3	$\bar{a}_1$	$\bar{a}_2$	$\bar{a}_3$
$s1$	–	0.18 %	0.02 %	–	1.14	6.33
$s2$	–	3.24 %	0.63 %	–	4.55	25.33
$s3$	1.29 %	0.45 %	0.09 %	0.94	3.90	5.97
$s4$	6.80 %	2.66 %	0.55 %	0.87	4.94	12.27
$s5$	0.04 %	–	$\approx$ 0 %	0.98	–	9.96
$s6$	4.97 %	1.98 %	0.42 %	0.93	6.75	31.76
$s7$	1.07 %	0.42 %	0.09 %	3.95	1.70	9.49
$s8$	1.79 %	0.69 %	0.14 %	1.86	4.94	22.22
$s9$	4.26 %	1.67 %	0.35 %	0.87	6.65	31.72

A quantidade média de recursos ocupados (carga) por enlace w^s é dada por

$$w^s = \sum_{i=1}^{k} \bar{a}_i^s c_i \tag{4.28}$$

para o tráfego escoado $\bar{a}_i^s$ e quantidade c_i de recursos requisitados por classe. A Figura 4.4 mostra os histogramas de ocupação para cada enlace s em relação ao total de recursos disponíveis C^s dos mesmos.

Um subproduto interessante do método de ponto fixo é o cálculo do atraso máximo fim-a-fim. Este atraso é acumulado ao longo dos caminhos devido ao tempo de propagação em cada um dos enlaces (d^s) e ao tempo de permanência no *buffer* presente em cada nó (d^N). O atraso de propagação depende da distância física entre os pontos que o enlace conecta, enquanto o atraso de armazenamento é função do tamanho da fila do servidor e da taxa de serviço.

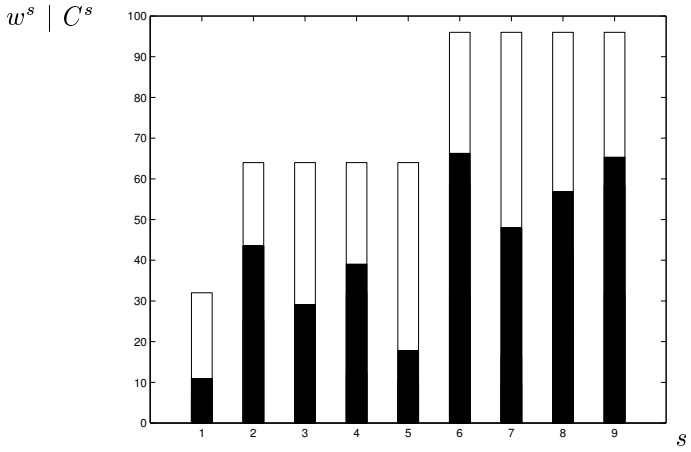

Figura 4.4: Ocupação de recursos por enlace para rede da Figura 4.3

No exemplo da rede da Figura 4.3 sejam os atrasos encontrados no sistema com os seguintes valores (em milissegundos)

$$d^s = \left[\begin{array}{ccccccccc} 0.20 & 0.20 & 0.20 & 0.20 & 0.20 & 0.50 & 0.50 & 0.50 & 0.50 \end{array}\right]$$
$$d^N = \left[\begin{array}{cccccc} 6.00 & 6.00 & 3.00 & 3.00 & 1.50 & 1.50 \end{array}\right]$$

Estes valores resultam nos atrasos máximos acumulados fim-a-fim por caminho ℓ (em milissegundos) conforme descrito na Tabela 4.5 para a rede da Figura 4.3. É importante observar que tanto o nó de origem quanto o nó destino também contribuem com o atraso acumulado.

Tabela 4.5: Atraso por caminho para rede da Figura 4.3

ℓ^1	ℓ^2	ℓ^3	ℓ^4	ℓ^5	ℓ^6	ℓ^7	ℓ^8	ℓ^9
11.2	13.2	19.4	11.2	13.2	5.0	8.2	8.2	7.0

Uma vez que um interesse de tráfego fim-a-fim pode escoar por qualquer um dos caminhos alternativos que componham cada subconjunto $\mathcal{F}_{x,y}$, o atraso máximo esperado d^M para cada um dos quatro interesses (pior caso) vale em milissegundos, respectivamente

$$d^M = \left[\begin{array}{cccc} 19.4 & 13.2 & 8.2 & 8.2 \end{array}\right]$$

4.5 Conclusão

Neste capítulo foi apresentado o método de ponto fixo generalizado para sistemas multiclasses, utilizando a notação matricial, com a análise de alguns exemplos ilustrativos.

O processo de determinar caminhos factíveis na rede a partir do conjunto de caminhos possíveis deve ser executado tanto na análise quanto na síntese de redes, além de ser importante no processo de operação em tempo real de uma rede, onde rotas têm de ser determinadas repetidamente. Neste caso, os métodos de encaminhamento podem se tornar

bastante complexos, caso sejam utilizados algoritmos para cálculo em sistemas distribuídos, notadamente em sistemas hierárquicos contendo recursos adicionais de roteamento (como por exemplo o *crank-back* do PNNI).

O método de equações de ponto fixo aplicado para sistemas multiclasses apresenta bons resultados quando utilizado conjuntamente com o cálculo de bloqueio pelo método de Kaufman. O uso de outras funções, tais como a forma-produto (teorema do produto) também é possível, embora neste caso a complexidade de cálculo aumente em demasia para um grande número de classes e recursos. A utilização híbrida, em que ambos os métodos são empregados, é uma boa alternativa para redes que apresentem reservas de recursos em alguns (mas não em todos) dos enlaces constituintes.

O cálculo por relaxação (substituição repetida) possui boa eficiência numérica e é aplicável a grande parte dos sistemas reais, convergindo com rapidez na maioria dos casos. O cálculo por meio do método de Newton apresenta o inconveniente da computação e inversão das matrizes jacobianas a cada passo da iteração, o que pode ser complexo dado a grande dimensão destas ($kS \times kS$), inviabilizando o método para sistemas com grande número de classes ou enlaces.

Embora o modelo de ponto fixo tenha sido definido para sistemas com balanceamento de carga apenas, o tratamento de transbordo também pode ser equacionado. Isso requer, contudo, extensões adicionais ao modelo por meio da resolução de vários problemas de ponto fixo, um para cada combinação de possibilidades de transbordo. Uma ponderação destes resultados parciais deve ser feita levando em conta todas as probabilidades condicionais de transbordo na rede, o que pode tornar o problema complexo, mesmo para um número não muito elevado de pontos de transbordo.

Capítulo 5

Conclusão

5.1 Contribuições

Na determinação da capacidade de um enlace o método de cálculo de banda efetiva foi analisado em suas duas abordagens mais comuns, a probabilidade de perda de células/pacotes e o atraso devido à permanência no *buffer*. Alguns modelos de três, dois e um parâmetros foram discutidos para o cálculo de banda efetiva, bem como alguns métodos de agendamento de pacotes no processo de transmissão. Simulações de eventos discretos foram usadas como metodologia de validação dos resultados. Algumas contribuições na escolha de modelos mais adequados, bem como do tamanho de *buffer*, foram apresentadas.

Sistemas multiclasses comutados foram caracterizados e uma descrição analítica na forma de espaço de estados foi apresentada. Também o cálculo da probabilidade de bloqueio em sistemas multiclasses comutados foi descrito para diferentes métodos, e o conceito de reserva de recursos visando a melhoria de desempenho do sistema foi discutido em duas finalidades específicas, a redução do número de recursos necessários no dimensionamento e a proteção contra sobrecarga. Contribuições na adoção das políticas de compartilhamento e algoritmos de dimensionamento foram também apresentadas.

A extensão do problema do bloqueio em enlaces individuais para redes foi realizada pela combinação do método de cálculo de ponto fixo com o método de cálculo de Kaufman para sistemas multiclasses. Uma descrição matricial do problema foi proposta, e métodos numéricos de resolução, especialmente o da substituição recursiva, foi empregada para obtenção de resultados ilustrativos.

5.2 Trabalhos futuros

Como trabalhos futuros, são propostos os seguintes tópicos:

- Exploração exaustiva dos métodos de banda efetiva para diversos modelos de fontes de tráfego, com validação por simulação, a fim de estabelecer limites confiáveis para os métodos;
- Elaboração de algoritmos mais eficientes de cálculo de bloqueio para sistemas com reservas de recursos, exatos ou aproximados, com validade estendida para políticas de acesso mais elaboradas;

- Implementação de transbordo para o método de cálculo de rede multiclasses, bem como a avaliação do impacto de diferentes algoritmos de encaminhamento no desempenho global da rede.

Apêndice A

Modelos de Simulação

Este apêndice inclui os modelos para simulação de eventos discretos executados no software *ARENA*, versão 5.0. Todas as execuções do programa foram computadas para um número mínimo de eventos da ordem de 10^7, com as adaptações pertinentes para que 1 unidade na escala de tempo do simulador corresponda a 10^{-3} segundos no tempo do modelo.

A.1 Fila FIFO

A Figura A.1 ilustra o modelo utilizado para a simulação de uma fila com disciplina de agendamento FIFO para um número de fontes N variando na faixa $1 \leq N \leq 5$.

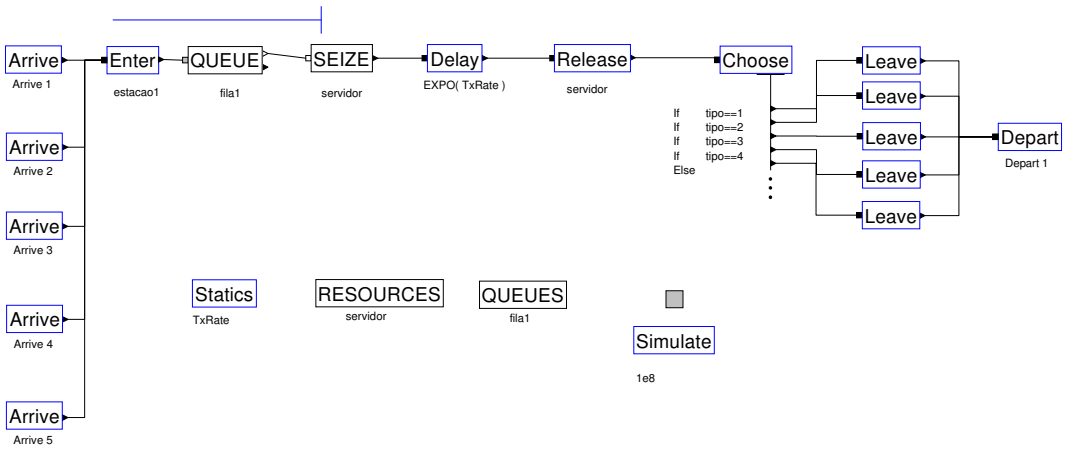

Figura A.1: Modelo para fila FIFO

Este modelo é utilizado para simular filas com diferentes tipos de atendimento (M/M/1 e M/D/1) para uma única fonte. Também é usado para verificar o ganho de multiplexação para várias fontes nas diferentes estratégias de dimensionamento empregadas em modelos de um parâmetro (conjunta e combinada). Fontes tipo fluido markoviano para diferentes parâmetros de pico são validadas com este modelo.

Os blocos `Arrive` modelam os processos de chegada, que podem ser poissonianos ou nulos (i.e. sem chegadas) nos exemplos. O bloco `Enter` registra a entrada de um elemento no sistema e inicializa as estatísticas a serem contabilizadas. A fila tipo FIFO é representada no bloco `QUEUE`. O bloco `SEIZE` é o componente encarregado de ocupar o servidor para uso

do elemento que se encontra na primeira posição da fila. O bloco `Delay` modela o tempo de atendimento do servidor, que pode ser constante ou exponencial negativo neste modelo. Após o serviço, o bloco `Release` libera o servidor para ser utilizado novamente pelo processo de *seize*. Após isto, o bloco `Choose` separa por atributo de qual fonte o elemento que foi servido é proveniente, sendo encaminhado em seguida para o bloco `Leave` adequado onde são contabilizadas as estatísticas necessárias (número de eventos, tempo de permanência, etc.). Finalmente, o bloco `Depart` encerra o trânsito dos elementos no sistema.

O bloco `Statics` é utilizado para atribuir valores para as constantes usadas (no caso a taxa de serviço). A fila e seu respectivo tamanho são declarados no bloco `QUEUES`, e o bloco `RESOURCES` define o (único) servidor disponível para uso do sistema. Os parâmetros de simulação são definidos no bloco `Simulate`.

A.2 Fila WRR e cíclica exaustiva

A Figura A.2 ilustra o modelo utilizado para simular os agendamentos WRR com filas de comprimento fixo (WRR-estático) e o atendimento cíclico exaustivo (C-exaust), para um número de fontes no intervalo $1 \leq N \leq 3$, mais uma possível quarta fonte interferente (tipo UBR, melhor esforço).

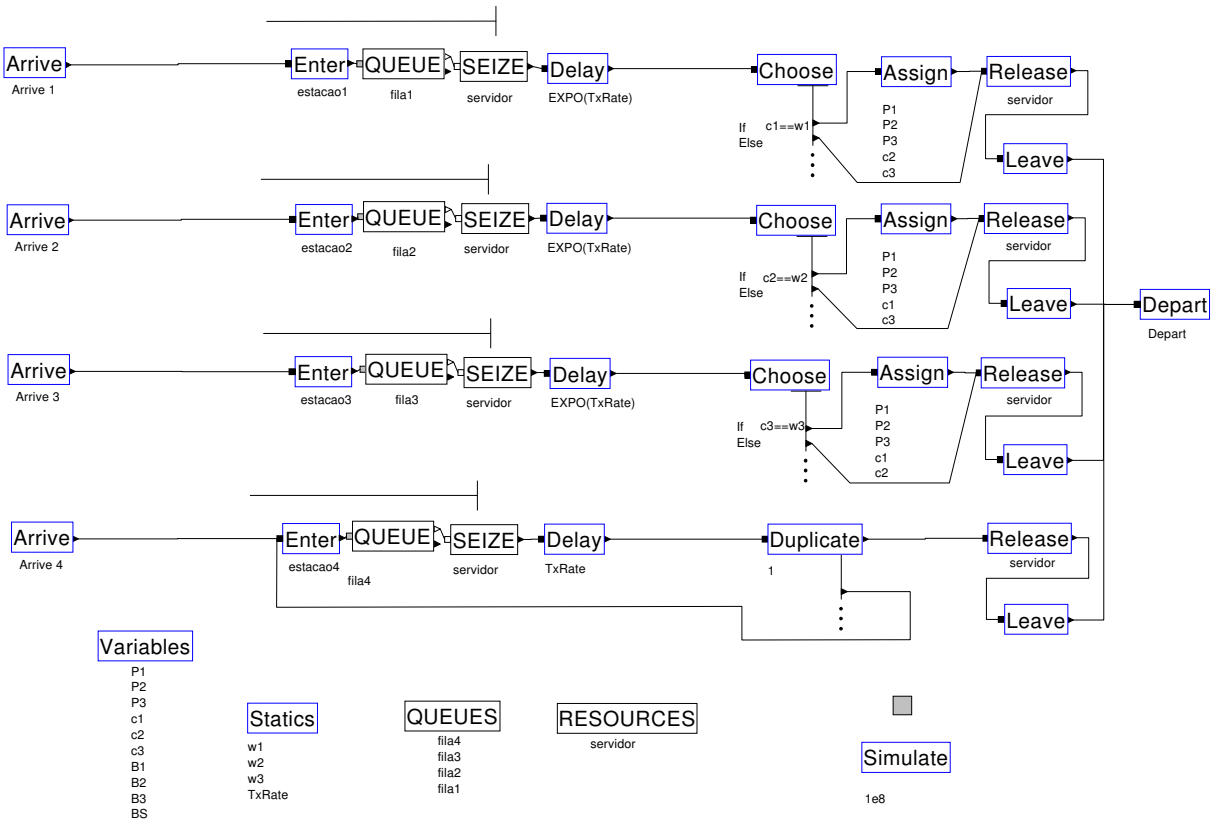

Figura A.2: Modelo para fila WRR e cíclica exaustiva

Os blocos `Arrive` modelam os processos de chegada, que podem ser poissonianos ou nulos (i.e. sem chegadas) nos exemplos. Os blocos `Enter` registram a entrada de um elemento no sistema e inicializam as estatísticas a serem contabilizadas. São modelados

quatro canais, os três primeiros para o processo cíclico e o quarto, com prioridade mais baixa, para o tráfego interferente. Cada uma das filas é do tipo FIFO, e são representadas pelos blocos `QUEUE`. Cada um está associado a um bloco `SEIZE` que é encarregado de ocupar o servidor para uso do elemento que se encontra na primeira posição da fila correspondente. Por haver apenas um servidor, apenas uma fila é atendida por vez de maneira não-preemptiva. Os blocos `Delay` associados a cada fila modelam o tempo de atendimento do servidor, que é do tipo exponencial negativo neste modelo. Após o serviço, os blocos `Choose` e `Assign` constituem um contador de elementos servidos. As variáveis `'c'` contam o número de elementos servidos de uma fila, e são comparadas às constantes `'w'` que representam os correspondentes pesos. Quando este valor é atingido, as variáveis `'P'`, que designam a prioridade para atendimento pelo servidor, têm seu valor alterado de modo cíclico para simular o efeito de varredura do WRR. Satisfeito o limite de peso para a fila 1, a prioridade passa para a 2. Satisfeito o limite de peso desta, a prioridade passa para a 3 e, finalmente, para a fila 1 fechando o ciclo. Após a atualização das variáveis de controle, os blocos `Release` liberam o servidor para ser utilizado novamente por algum dos processos de *seize*. Em seguida, os blocos `Leave` contabilizam as estatísticas adequadas (número de eventos, tempo de permanência, etc.). O bloco `Depart` encerra o trânsito dos elementos no sistema.

Para simular a fila cíclica exaustiva, o sistema de contagem é alterado de modo que a comparação efetuada no bloco `Choose` é agora com o número de elementos presentes na correspondente fila naquele instante. Apenas quando este número atingir zero o bloco `Assign` passa a prioridade de varredura (variáveis `'P'`) para a fila subseqüente, também de maneira cíclica.

O quarto canal gera o tráfego interferente, que pode variar de zero a infinito. O bloco `Duplicate` é ativado apenas nesta última situação, gerando um elemento persistente (*greedy*) de modo a sempre haver demanda por serviço no canal de baixa prioridade (UBR). Elementos desta fila possuem a mais baixa prioridade e são só atendidos quando não houver elementos presentes nas demais filas, de maneira não-preemptiva.

Todas as variáveis do sistema são declaradas no bloco `Variables` e o bloco `Statics` é utilizado para atribuir valores para as constantes usadas (taxa de serviço e pesos). As filas e seu respectivo tamanho são declarados no bloco `QUEUES`, e o bloco `RESOURCES` define o (único) servidor disponível para uso do sistema. Os parâmetros de simulação são definidos no bloco `Simulate`.

A.3 Fila WRR com buffer compartilhado

A Figura A.3 ilustra o modelo utilizado para simular o agendamento WRR com filas de comprimento variável (WRR-dinâmico) com uma quantidade de *buffer* compartilhada entre todas. O modelo comporta um número de fontes no intervalo $1 \leq N \leq 3$, mais a possível quarta fonte interferente (tipo UBR).

O funcionamento é semelhante ao modelo da disciplina WRR descrito no item anterior, com o sistema de atribuição cíclica de prioridades operando com as mesmas variáveis. É acrescentado em cada um dos três canais cíclicos um conjunto de blocos para gerenciar o comprimento (dinâmico) das filas. Variáveis `'B'` são criadas para controlar o tamanho admissível de cada uma das filas, uma das quais (`BS`) contém o valor total de *buffer* dis-

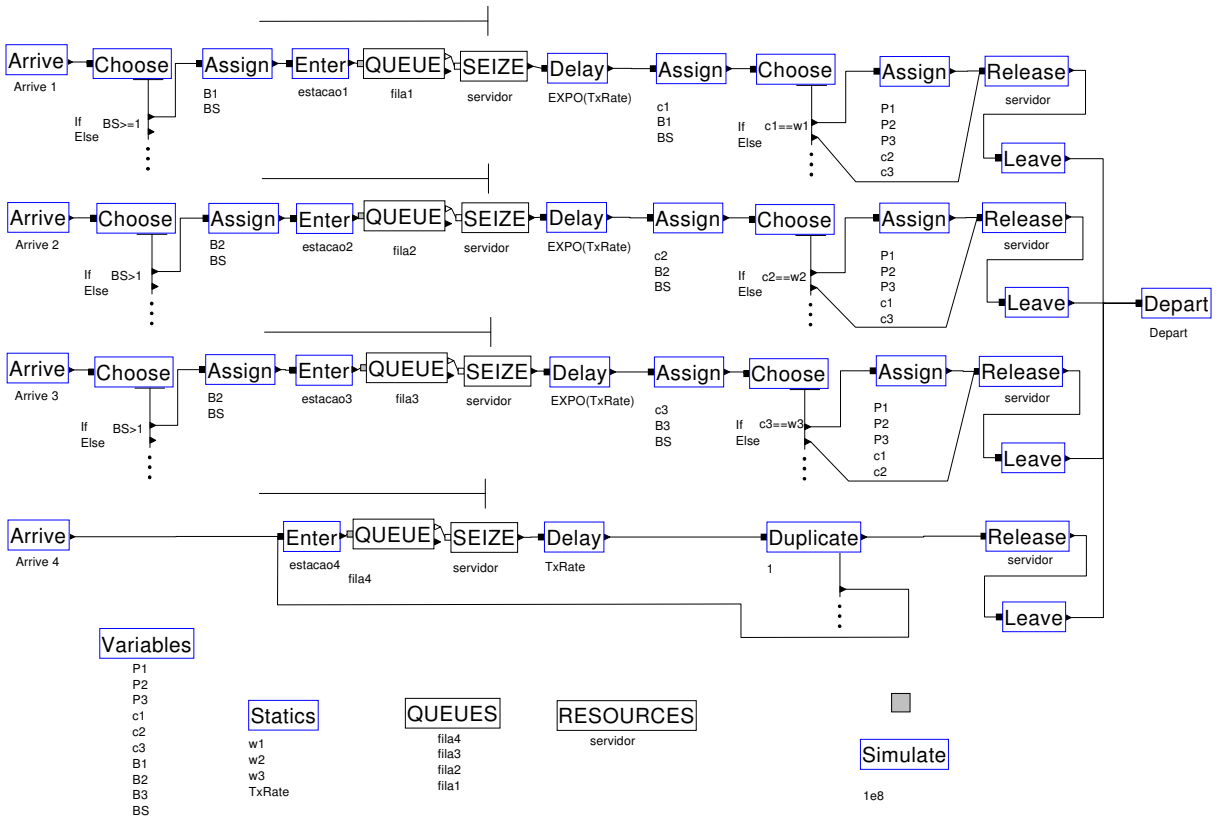

Figura A.3: Modelo para fila WRR com buffer compartilhado

ponível ao sistema. Os blocos `Choose` na entrada de cada canal verificam se há *buffer* disponível para ser alocado e só permite o ingresso de elementos se esta condição for satisfeita. Caso contrário, o elemento é eliminado. O bloco `Assign` que vem logo a seguir decrementa a variável `'B'` correspondente, indicando que aquela fila recebeu um acréscimo no seu comprimento variável. De maneira análoga, os blocos `Assign` que estão situados após os processos de *delay* realizam a função de liberar recursos que foram alocados para uma fila em particular de volta para o valor total de *buffer* disponível (`BS`).

As variáveis adicionais deste modelo também são declaradas no bloco `Variables`.

Apêndice B

Validação do Ponto Fixo

O algoritmo de ponto fixo implementado no software *MATLAB* foi validado com auxílio do exemplo dado em [16] para uma rede em estrela, conforme ilustrado na Figura B.1.

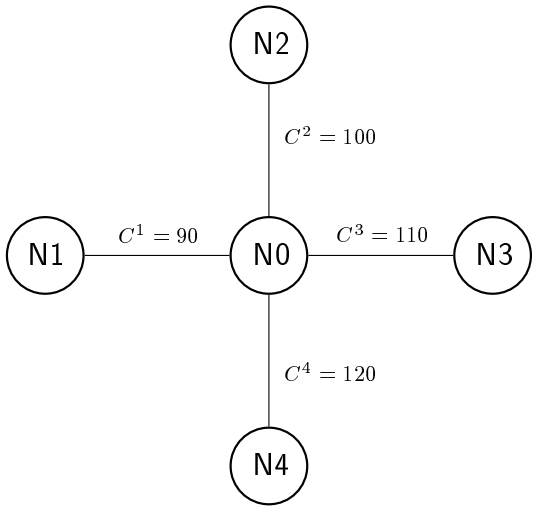

Figura B.1: Modelo para validação – rede em estrela

Este modelo possui duas classes de serviço definidas, com $c_1 = 5$ e $c_2 = 1$. As capacidades dos enlaces são dadas por

$$C^s = \left[\begin{array}{cccc} 90 & 100 & 110 & 120 \end{array} \right]$$

com interesse de tráfego para ambas as classes entre todos os braços da estrela exceto para o nó central, definindo um conjunto $\mathcal{F}$ de seis caminhos enumerados

$$\begin{aligned} \ell^1 &= \{N1 \leftrightarrow N0 \leftrightarrow N2\} \\ \ell^2 &= \{N1 \leftrightarrow N0 \leftrightarrow N3\} \\ \ell^3 &= \{N1 \leftrightarrow N0 \leftrightarrow N4\} \\ \ell^4 &= \{N2 \leftrightarrow N0 \leftrightarrow N3\} \\ \ell^5 &= \{N2 \leftrightarrow N0 \leftrightarrow N4\} \\ \ell^6 &= \{N3 \leftrightarrow N0 \leftrightarrow N4\} \end{aligned}$$

A Tabela B.1 mostra os resultados para todos os interesses de tráfego e classes, comparando os resultados simulados pelo método de Monte Carlo, com intervalo de confiança de 95% conforme [16] e os resultados calculados pelo método de ponto fixo implementado neste trabalho. O perfil de tráfego adotado é do tipo leve (*light*).

Tabela B.1: Resultados fim-a-fim para rede da Figura B.1 – Tráfego leve

	classe 1			classe 2		
Caminho	A_1^ℓ	simulação	ponto fixo	A_2^ℓ	simulação	ponto fixo
ℓ^1	1.6	0.38 ∼ 0.39	0.39	9	0.05 ∼ 0.05	0.05
ℓ^2	1.6	0.34 ∼ 0.35	0.35	9	0.05 ∼ 0.05	0.05
ℓ^3	1.6	0.34 ∼ 0.34	0.34	9	0.04 ∼ 0.05	0.05
ℓ^4	1.6	0.05 ∼ 0.05	0.06	9	0.01 ∼ 0.01	0.01
ℓ^5	1.6	0.05 ∼ 0.05	0.05	9	0.01 ∼ 0.01	0.01
ℓ^6	1.6	0.01 ∼ 0.01	0.01	9	0.00 ∼ 0.00	0.00

A Tabela B.2 mostra os resultados para a mesma situação, mas para um perfil de tráfego do tipo médio (*moderate*).

Tabela B.2: Resultados fim-a-fim para rede da Figura B.1 – Tráfego moderado

	classe 1			classe 2		
Caminho	A_1^ℓ	simulação	ponto fixo	A_2^ℓ	simulação	ponto fixo
ℓ^1	2	2.29 ∼ 2.30	2.35	10	0.34 ∼ 0.35	0.36
ℓ^2	2	1.96 ∼ 1.98	1.99	10	0.29 ∼ 0.30	0.30
ℓ^3	2	1.90 ∼ 1.92	1.92	10	0.28 ∼ 0.29	0.29
ℓ^4	2	0.49 ∼ 0.50	0.53	10	0.07 ∼ 0.07	0.07
ℓ^5	2	0.43 ∼ 0.44	0.46	10	0.06 ∼ 0.06	0.06
ℓ^6	2	0.08 ∼ 0.08	0.09	10	0.01 ∼ 0.01	0.01

A Tabela B.3 mostra os resultados para situação semelhante, com um perfil de tráfego do tipo intenso (*heavy*).

Tabela B.3: Resultados fim-a-fim para rede da Figura B.1 – Tráfego intenso

	classe 1			classe 2		
Caminho	A_1^ℓ	simulação	ponto fixo	A_2^ℓ	simulação	ponto fixo
ℓ^1	3	28.0 ∼ 28.1	28.6	15	5.70 ∼ 5.70	5.70
ℓ^2	3	23.3 ∼ 23.4	23.8	15	4.60 ∼ 4.60	4.60
ℓ^3	3	21.4 ∼ 21.4	21.6	15	4.10 ∼ 4.20	4.20
ℓ^4	3	13.2 ∼ 13.2	13.8	15	2.40 ∼ 2.40	2.50
ℓ^5	3	10.9 ∼ 11.0	11.4	15	1.90 ∼ 1.90	2.00
ℓ^6	3	4.80 ∼ 4.80	5.40	15	0.08 ∼ 0.08	0.90

Apêndice C

Glossário de Símbolos

Este apêndice contém a listagem geral de símbolos usados ao longo deste trabalho.

Capítulo 2 – Banda Efetiva (1/2)

$\mathcal{A}(t)$	Processo de chegada de elementos em função do tempo
$\mathcal{A}_i(t)$	Processo de chegada de elementos da fonte i em função do tempo
b; b_i	Limiar de *buffer*; limiar de *buffer* para fonte i
c; c_i	Banda efetiva; banda efetiva para fonte i
c_D	Banda efetiva calculada por atraso
c_L	Banda efetiva calculada por perda
C	Banda efetiva agregada
C_I	Estimativa (parcial) para banda efetiva agregada – Método Guérin
C_{II}	Estimativa (parcial) para banda efetiva agregada – Método Guérin
d	Atraso máximo admissível no *buffer*
d_i	Atraso máximo admissível para classe i
$\tilde{d}$	Atraso máximo verificado no *buffer*
$E\{\cdot\}$	Operador esperança
$E\{W\}$	Tempo médio de espera na fila
$E\{Q\}$	Número médio de elementos na fila
f	Variável auxiliar
g	Variável auxiliar
$H(\delta)$	Função geradora de Momentos Logarítmicos
i	Índice genérico
j	Índice genérico
K	Número máximo admissível de elementos em espera no modelo M/M/1/K
l	Variável auxiliar
$\log(\cdot)$	Função logaritmo neperiano
$\log_{10}(\cdot)$	Função logaritmo na base 10
L_b	Perda verificada em sistema com *buffer* de tamanho b
m	Média (genérica) de função estatística
n	Número de elementos gerados num intervalo de tempo t

Capítulo 2 – Banda Efetiva (2/2)

N	Número de fontes agregadas em um único servidor
p_n	Probabilidade de que existam n elementos no *buffer*
$\bar{p}_i$	Tempo médio de duração de surto
$\hat{p}_i$	Tempo máximo de duração de surto
$P(n,t)$	Matriz de probabilidades de transição em modelo MMPP
q	Vetor de probabilidades estacionárias de fase
Q	Número de elementos numa fila genérica
R	Matriz de transição de fase em modelo MMPP
t	Variável de tempo
T_i	Parâmetro de pico – Modelo Guérin
T_{on}	Tempo de permanência no estado ativo – Modelos *On-Off*
T_{off}	Tempo de permanência no estado inativo – Modelos *On-Off*
u	Taxa de chegada de tráfego interferente
$\mathcal{W}$	Função W de Lambert
x	Fase atual de fonte MMPP
y	Fase subseqüente de fonte MMPP
z	Variável da transformada Z
α	Variável auxiliar
β	Variável auxiliar
δ	Parâmetro de perda definido como $(-\log(\varepsilon)/b)$
Δ	Parâmetro de perda alternativo $(-\log(\varepsilon))$
$\tilde{\Delta}$	Parâmetro de perda alternativo $(-\log_{10}(\varepsilon))$
ε	Valor máximo admissível da perda devido a congestionamento de *buffer*
ε_i	Valor máximo de perda ε para a fonte i
ζ	Relação entre a variância e a média da função de chegada
θ	Parâmetro de agregação – Modelo Neame-Zukerman
λ; λ_i	Taxa média de ingresso no sistema; taxa média da classe i
$\hat{\lambda}$; $\hat{\lambda}_i$	Taxa máxima de ingresso no sistema; taxa máxima da classe i
Λ	Matriz de taxas de transmissão por fase em modelo MMPP
μ, μ_i	Taxa de atendimento de servidor; taxa de atendimento para classe i
ξ	Relação de pico PCR/SCR
ρ	Ocupação de servidor (λ/μ)
σ^2	Variância (genérica) de função estatística
Φ	Número de fases em modelo MMPP

Capítulo 3 – Bloqueio em Enlaces (1/2)

a_i	Tráfego oferecido λ_i/μ_i pela classe i
a_i^t	Tráfego transbordado sobre a classe i
$\tilde{a}_i$	Tráfego oferecido com sobrecarga pela classe i
B	Vetor de probabilidades de bloqueio
B_i	Probabilidade de bloqueio para a classe i
B^N	Vetor de bloqueios naturais
B^R	Vetor de bloqueios alterados por reserva
$B^\star$	Valor objetivo de bloqueio para dimensionamento
$\tilde{B}^\star$	Valor objetivo de bloqueio com tolerância para dimensionamento
c	Vetor de quantidade de recursos requisitados pelas classes
c_i	Componente do vetor c para a classe i
C_i	Recursos disponíveis exclusivamente para a classe i
C	Total de recursos disponíveis
D	Matriz de descritores de tráfego
D_i	Descritores de tráfego para a classe i
$\hat{D}$	Matriz abreviada de descritores de tráfego
$\hat{D}_i$	Descritores abreviados de tráfego para a classe i
g	Variável auxiliar
$G(\cdot)$	Função de normalização do método do Teorema do Produto
i	Classe de serviço qualquer $i = 1, \ldots, k$
J	Matriz de sensibilidade ao tráfego
$J_{(a)}$	Matriz de sensibilidade J calculada para o ponto a
j	Índice auxiliar
k	Número de classes de serviço
L	Vetor indicativo de presença de sobrecarga
L_i	Indicativo de presença de sobrecarga na classe i
m	Índice auxiliar
n	Vetor de estado da quantidade de recursos ocupados num dado instante
n_i	Componente do vetor n para a classe i
n_i^+	Estado adjacente (superior) ao estado n em relação à classe i
n_i^-	Estado adjacente (inferior) ao estado n em relação à classe i
$\mathcal{N}_i$	Tamanho da população da classe i
$\tilde{p}(\nu)$	Função (contínua) de distribuição de probabilidade de estado ν
$p^*(\nu)$	Função re-normalizada de distribuição de probabilidade de estado ν
P_n	Probabilidade do estado n (discreto)
$\tilde{P}_j$	Probabilidade não-normalizada de estado j (unidimensional)
$\tilde{P}_n$	Probabilidade não-normalizada de estado n (multidimensional)
q_i	*Quantum* de recursos $\lfloor C/c_i \rfloor$ disponíveis para a classe i
R_i	Limite máximo para a variável de estado n_i devido à presença de reservas
$\mathcal{S}^L$	Algoritmo de dimensionamento para reserva parcial tipo θ (limiar)
$\mathcal{S}^N$	Algoritmo de dimensionamento para compartilhamento total
$\mathcal{S}^S$	Algoritmo de dimensionamento para reserva parcial tipo σ (simples)
$\mathcal{S}^T$	Algoritmo de dimensionamento para reserva plena

Capítulo 3 – Bloqueio em Enlaces (2/2)

$T_0(\cdot)$	Valor assintótico (limite zero) calculado – Método de LeGall
$T_\infty(\cdot)$	Valor assintótico (limite infinito) calculado – Método de LeGall
$\mathcal{U}^k$	Conjunto universo de todos os estados n com k componentes
w	Vetor ocupação (carga) de recursos
w_i	Ocupação (carga) de recursos para classe i
α	Porcentagem de sobrecarga de tráfego em um enlace
$\Gamma(\cdot)$	Função Gama
$\delta_i^-(n)$	Transição devido a partida de elemento da classe i, no estado n
$\delta_i^+(n)$	Transição devido a chegada de elemento da classe i, no estado n
ϵ	Tolerância de dimensionamento no bloqueio objetivo
ζ_i	Relação entre a variância e a média da função de chegada para a classe i
θ_i	Limiar de admissão para a classe i
θ_i^t	Limiar de admissão para um tráfego transbordante sobre a classe i
λ_i	Média do tempo entre chegadas de requisições para a classe i
$\Lambda(x)$	Função região de maior probabilidade de estados
$\Lambda_i^+(x)$	Função região de maior probabilidade de estados na direção de i
μ_i	Inverso da média do tempo de permanência para a classe i
ν	Variável contínua de estado da quantidade de recursos ocupados
σ_i	Valor de reserva do tipo estático (sem marcação) para classe i
$\Psi(\cdot)$	Função auxiliar
Ω	Espaço de estados n admissíveis a um sistema
Ω^B	Sub-espaço de Ω que contém os estados bloqueantes
Ω_i^B	Sub-espaço de Ω dos estados bloqueantes para a classe i

Capítulo 4 – Bloqueio em Redes (1/1)

$a(x,y)$	Elemento de tráfego entre os pontos x, y
$a_i(x,y)$	Elemento de tráfego entre os pontos x, y para classe i
a_i^s	Tráfego oferecido da classe i ao enlace s
$\bar{a}_i^s$	Tráfego escoado da classe i pelo enlace s
$a_i^{s,\ell}$	Tráfego oferecido da classe i ao enlace s devido ao caminho ℓ
$\bar{a}_i^{s,\ell}$	Tráfego escoado da classe i pelo enlace s devido ao caminho ℓ
$A(x,y)$	Matriz de interesses de tráfego
$A_i(x,y)$	Matriz de interesses de tráfego para a classe i
A_i^ℓ	Elemento da matriz de tráfego oferecido enumerada por caminhos, classe i
$\bar{A}_i^\ell$	Elemento da matriz de tráfego escoado enumerada por caminhos, classe i
B_i^s	Bloqueio da classe i no enlace s
$\mathcal{B}_i^\ell$	Bloqueio fim-a-fim da classe i para o caminho ℓ
c_i	Vetor de quantidade de recursos requisitados pela classe i
C^s	Quantidade de recursos disponíveis no enlace s
d^M	Vetor de atrasos verificados fim-a-fim
d^N	Vetor de atrasos nos nós de comutação
d^s	Vetor de atrasos de propagação nos enlaces
$\mathcal{F}$	Conjunto de todos os caminhos factíveis na rede
$\mathcal{F}_{x,y}$	Conjunto de todos os caminhos factíveis entre os pontos x, y
$\mathbb{F}(X)$	Função para cálculo das equações de ponto fixo
i	Classe de serviço qualquer $i = 1, \ldots, k$
I	Matriz de incidência para enlaces e caminhos
$I_{s,\ell}$	Elemento da matriz de incidência I
$J(s_x, s_y)$	Matriz jacobiana para única classe de serviço – Método de Newton
$\mathbb{J}(X)$	Matriz jacobiana para cálculo de ponto fixo – Método de Newton
k	Número total de classes de serviço
$\mathcal{K}\{\cdot\}$	Função de cálculo de bloqueio multiclasses
ℓ	Caminho enumerado através da rede
$\ell_{x,y}$	Caminhos enumerados entre os pontos x, y
$\ell_{x,y}^r$	r-ésimo caminho entre os pontos x, y – Balanceamento de carga
$\ell_{x,y}^v$	v-ésimo caminho entre os pontos x, y – Transbordo
L	Número de caminhos na rede
N	Número de nós da rede
$\mathcal{P}$	Conjunto de todos os caminhos possíveis na rede
$\mathcal{P}_{x,y}$	Conjunto de todos os caminhos possíveis entre os pontos x, y
$\mathcal{R}$	Matriz de topologia descritiva da rede
s	Enlace qualquer da rede $s = 1, \ldots, S$
S	Número total de enlaces existentes
w^s	Ocupação de recursos (carga) no enlace s
X	Vetor-solução do sistema de equações de ponto fixo
Z	Função auxiliar – Método de Minimização
$\tau_{x,y}$	Vetor de partição de carga para o interesse de tráfego x, y
$\tau_{x,y}^r$	r-ésima componente do vetor de partição de carga

Referências Bibliográficas

[1] R. G. Addie, M. Zukerman, and T. D. Neame. Broadband traffic modeling: simple solutions to hard problems. *IEEE Communications Magazine*, pages 88–95, August 1998.

[2] J. M. Aein. A multi-user-class blocked-calls-cleared demand access model. *IEEE Transactions on Communications*, 26(3), March 1978.

[3] J. M. Aein and O. S. Kosovich. Satellite capacity allocation. In *Special Issue of the Proceedings of IEEE*, volume 65-3, pages 332–342, March 1977.

[4] ATM Forum. *Private network-network interface (PNNI)*, 1.0 edition, May 1996.

[5] A. Baiocchi and N. Bléfari-Melazzi. Steady-state analysis of the MMPP/G/1/K queue. *IEEE Transactions on Communications*, 41(4):531–534, April 1993.

[6] R. E. Bellman. *Dynamic Programming*. Princeton University Press, Princeton-NJ, 1957.

[7] L. R. Bercho, I. S. Bonatti, P. L. D. Peres, and A. Lopes. Robust link dimensioning in multi-rate loss networks. In *Proceedings of ITS2002 – International Telecommunications Symposium*, Natal-RN, Brazil, 2002.

[8] A. W. Berger and Y. Kogan. Dimensioning bandwidth for elastic traffic in high-spped data networks. *IEEE/ACM Transactions on Networking*, 8(5):643–654, October 2000.

[9] J. Y. Le Boudec. Network calculus made easy. Technical Report EPFL-DI 96/218, Laboratoire de Réseaux de Communication, Ecole Polytechnique Fédérale de Lausanne (EPFL), December 1996.

[10] J. Y. Le Boudec and R. Nagarajan. A CAC algorithm for VBR connections over a VBR trunk. In *Proceedings of 15th International Teletraffic Congress - ITC*, volume 1, pages 59–70, Washington, DC, USA, 1997.

[11] C. S. Chang and J. A. Thomas. Effective bandwidth in high-speed digital networks. *IEEE Journal on Selected Areas in Communications*, 13(6), August 1995.

[12] S. Chen and K. Nahrstedt. On finding multi-constrained paths. In *IEEE ICC'98 Proceedings*, June 1998.

[13] S. Chen and K. Nahrstedt. An overview of quality of service routing for next-generation high-speed networks: problems and solutions. *IEEE Network Magazine*, pages 64–79, November/December 1998.

[14] G. L. Choudhury, K. K. Leung, and W. Whitt. An algorithm for product-form loss networks based on numerical inversion of generating functions. *IEEE Transactions on Communications*, 1994.

[15] G. L. Choudhury, D. M. Lucantoni, and W. Whitt. Squeezing the most out of ATM. *IEEE Transactions on Communications*, 44(2):203–217, February 1996.

[16] S. P. Chung and K. W. Ross. Reduced load approximations for multirate loss networks. *IEEE Transactions on Communications*, 41(8):1222–1231, August 1993.

[17] I. Cidon, R. Rom, and Y. Shavitt. Analysis of multi-path routing. *IEEE/ACM Transactions on Networking*, 7(6):885–896, December 1999.

[18] R. M. Corless, G. H. Gonnet, D. E. G. Hare, D. J. Jeffrey, and D. E. Knuth. On the Lambert W function. *Advances in Computational Mathematics*, 5, 1996.

[19] L. E. N. Delbrouck. On the steady-state distribution in a service facility carrying mixtures of traffic with different peakedness factors and capacity requirements. *IEEE Transactions on Communications*, COM-31(11):1209–1211, November 1983.

[20] E. Dijkstra. A note on two problems in connection with graphs. *Numerische Mathematik*, 1:269–271, 1959.

[21] A. I. Elwalid and D. Mitra. Effective bandwidth of general markovian traffic sources and admission control of high speed networks. *IEEE Transactions on Networking*, 1(3):329–343, June 1993.

[22] A. Farago. Blocking probability estimation for general traffic under incomplete information. In *IEEE International Conference on Communications, ICC 2000*, volume 3, pages 1547–1551, New Orelans-LA, USA, June 2000.

[23] R. F. Farmer and L. Kaufman. On the numerical evaluation on some basic traffic formulae. *Networks*, 8:153–186, 1978.

[24] F. Le Gall and J. Bernussou. Blocking probabilities for trunk reservation policy. *IEEE Transactions on Communications*, COM-35(3):313–318, March 1987.

[25] A. Girard. *Routing and Dimensioning in Circuit-Switched Networks*. Addison-Wesley, 1990.

[26] D. Gross and C. M. Harris. *Fundamentals of Queueing Theory*. John Wiley & Sons, Inc., Toronto, Canada, 1974.

[27] R. Guérin, H. Ahmadi, and M. Naghshineh. Equivalent capacity and its application to bandwidth allocation in high-speed networks. *IEEE Transactions on Selected Areas in Communications*, 9(7):968–981, September 1991.

[28] R. Guérin and A. Orda. QoS-based routing in networks with inaccurate information: theory and algorithms. In *IEEE INFOCOM '97 Proceedings*, Japan, April 1997.

[29] F. Hartleb and G. Hasslinger. Comparison of link dimensioning methods for TCP/IP networks. In *IEEE Global Telecommunications Conference, GLOBECOM '01*, pages 2240–2247, November 2001.

[30] S. A. Johnson. A performance analysis of integrated communication systems. *British Telecom Technical Journal*, 3(4), October 1985.

[31] M. Katenevis, S. Sidiropoulos, and C. Courcoubetis. Weighted Round-Robin cell multiplexing in a general purpose ATM switch chip. *IEEE Journal on Selected Areas in Communications*, 9(8):1265–1279, October 1991.

[32] J. S. Kaufman. Blocking in a shared resource environment. *IEEE Transaction on Communications*, 29(10):1474–1481, 1981.

[33] G. Kesidis, J. Walrand, and C. S. Chang. Effective bandwidths for multiclass Markov fluids and other ATM sources. *IEEE Transactions on Networking*, 1(4):424–428, August 1993.

[34] Q. Ma and P. Steenkiste. Quality-of-service routing with performance guarantees. In *Proc. 4th Int. IFIP Workshop QoS*, May 1997.

[35] G. Malicsko, G. Fodor, and M. Pióro. Link capacity dimensioning and path optimization for networks supporting elastic services. In *IEEE International Conference on Communications, ICC 2002*, volume 4, pages 2304–2311, 2002.

[36] R. S. Mata. Dimensionamento de enlaces em redes com integração de serviços. Master's thesis, FEEC – Unicamp, Campinas-SP, Brasil, Abril 2002.

[37] M. Naraghi-Pour. Bandwidth and buffer dimensioning for guaranteed quality of service in wireless ATM networks. In *IEEE Wireless Communications and Networking Conference, WCNC 2000*, volume 2, pages 730–735, Chicago-IL, USA, September 2000.

[38] T. Neame, M. Zukerman, and N. F. Maxemchuk. On link dimensioning based on estimating the mean rate by the sustainable rate. *IEEE Communications Letters*, 5(2):73–75, February 2001.

[39] T. D. Neame, M. Zukerman, and R. G. Addie. A tradeoff between increased network speed and reduced efective bandwidths. *IEEE/ACM Transactions on Networking*, pages 979–984, 2001.

[40] J. C. Oliveira. Dimensionamento de enlaces em redes de telecomunicações ATM. Master's thesis, FEEC – Unicamp, Campinas-SP, Brasil, Fevereiro 1998.

[41] R. Guérin and L. Gun. A unified approach to bandwidth allocation and access control in fast packet-switched networks. In *Proc. IEEE Infocom '92*, Florence, Italy, May 1992.

[42] M. Ritter and P. Tran-Gia. Multi-rate models for dimensioning and performance evaluation of ATM networks - COST 242. Technical report, University of Wurzburg, June 1994.

[43] J. Roberts, U. Mocci, and J. Virtamo, editors. *Broadband network teletraffic – performance evaluation and design of broadband multiservice networks – Final report of action COST 242.* Springer-Verlarg, Berlin, Heidelberg, New York, 1996.

[44] J. W. Roberts. Teletraffic models for the telecom 1st integrated services network. In *10th Int. Teletraffic Congress Proceedings*, page 1.1.2, 1983.

[45] T. L. Saaty. *Elements of queueing theory with applications.* McGraw Hill, New York, 1961.

[46] H. F. Salama, D. S. Reeves, and Y. Viniotis. A distributed algorithm for delay-constrained unicast routing. In *IEEE INFOCOM '97*, Japan, April 1997.

[47] K. G. Shin and C. C. Chou. A distributed route-selection scheme for establishing real-time channel. In *6th IFIP Int. Conference High Performance Networking*, pages 319–329, September 1995.

[48] J. T. Virtamo. Reciprocity of blocking probabilities in multiservice loss systems. *IEEE Transactions on Communications*, 36(10):1174–1175, October 1988.

[49] Z. Wang and J. Crowcroft. QoS routing for supporting resource reservation. *IEEE JSAC*, 1996.