Universidade Federal do Pará

Centro Tecnológico

Programa de Pós-Graduação em Engenharia Elétrica

Modelagem e análise de desempenho de esquemas de alocação de recursos em redes móveis celulares

Glaucio Haroldo Silva de Carvalho

Orientador:

Prof. Dr. João Crisóstomo Weyl A. Costa

Belém
2005.

Modelagem e análise de desempenho de esquemas de alocação de recursos em redes móveis celulares

Glaucio Haroldo Silva de Carvalho

Orientador:

Prof. Dr. João Crisóstomo Weyl A. Costa

Tese de Doutorado submetida à Banca Examinadora do Programa de Pós-graduação em Engenharia Elétrica da Universidade Federal do Pará como requisito para obtenção do título de "Doutor em Engenharia Elétrica".

Universidade Federal do Pará

Belém
2005

Modelagem e análise de desempenho de esquemas de alocação de recursos em redes móveis celulares

Glaucio Haroldo Silva de Carvalho

Banca examinadora

.....

Prof. Dr. João Crisóstomo Weyl Albuquerque Costa (UFPA)- Orientador

.....

Prof. Dr. Carlos Renato Lisboa Francês (UFPA)-Co-Orientador

.....

Prof. Dr. Evaldo Gonçalves Pelaes (UFPA)-Membro

.....

Prof. Dr. Gervásio Protásio dos Santos Cavalcante (UFPA)-Membro

.....

Prof. Dr. Aldebaro Barreto da Rocha Klautau Jr. (UFPA)-Membro

.....

Prof. Dr. Michel Daoud Yacoub (UNICAMP)-Membro externo

.....

Prof. Dr. Solon Venâncio de Carvalho (INPE)-Membro externo

Visto:

.....

Prof. Dr. Roberto Célio Limão

Coordenador do PPGEE/CT/UFPA

Aos meus pais

Nelcy e Lilian

meus irmãos

George e Glauco,
e minha esposa

Mary,
com amor...

Agradecimentos

A frase "Sem orientação não há aprendizado"do filme **O Tigre e o Dragão** ganhador de quatro Oscars do diretor Ang Lee traduz o meu sentimento após o término deste trabalho. Assim, gostaria de agradecer imensamente à todas as pessoas que colaboraram nesta minha caminhada. Especialmente, àquele que vislumbrou tudo isso em meados de 2001. Dessa maneira, fica o meu sincero agradecimento ao Prof. Dr. João Crisóstomo Weyl A. Costa pelo suporte, orientação, esforço e amizade. A esse professor e pesquisador fica também o meu muito obrigado por insistentemente lutar para dotar cada vez mais de competência a Amazônia.

Seguindo uma ordem cronológica, fica meu muito obrigado ao Prof. Dr. Carlos Renato Lisboa Francês que me apresentou os conceitos de análise de desempenho, sendo dessa forma, uma peça chave para a definição da minha linha de pesquisa. Porém, dentro de um processo de maturidade científica um dos maiores ensinamentos é o suporte as idéias (mesmo quando essas parecem não muito claras) e o incentivo para executá-las. Assim, novamente, me reporto a esse professor para agradecer pelo apoio, paciência e amizade.

Devo muito ainda ao Prof. Dr. Solon Venâncio de Carvalho que possibilitou o meu estágio no Laboratório Associado de Computação e Matemática Aplicada no INPE de São José dos Campos. Fica o meu obrigado pela orientação e suporte técnico dentro dessa fantástica área de pesquisa da modelagem que compreende os processos Markovianos. Além disso, agradeço pela amizade balizada nas conversas sobre os mais diversos assuntos.

Gostaria de agradecer também ao Programa de Pós-Graduação em Engenharia Elétrica (PPGEE) da Universidade Federal do Pará por possibilitar o desenvolvimento deste trabalho. Agradeço ainda aos professores e funcionários desse programa que tomaram parte deste trabalho.

Devo também um obrigado aos colegas de estudo do LEA e LACA pelas conversas sempre muito bem humoradas e que tornam o desenvolvimento de uma Tese de Doutorado uma tarefa mais agradável. Em especial aos amigos André Mendes Cavalcante, Josiane do Couto Rodrigues (*fia*), Marco José de Sousa, Mauro Margalho Coutinho, Claudomiro de

Souza Sales, Diego Lisboa Cardoso, Edvar da L. Oliveira, Claudia da Silva Batista, Elaine Sena Lelis, etc. Com certeza, como sempre, esqueci alguem, porém, mesmo assim, fica um muito obrigado. Gostaria de agradecer à Michelle Bitar Lelis dos Santos, Roberto Menezes Rodrigues, pela ajuda técnica, amizade e paciência que tiveram quando permitiram a minha ajuda no desenvolvimento dos seus trabalhos.

Fica o meu obrigado também aos integrantes, pesquisadores e funcionários do Laboratório Associado de Computação e Matemática Aplicada (LAC) no INPE de São José dos Campos por propiciar me um ambiente agradável de trabalho. Em particular aos amigos Marcos Antonio Pereira e Diego Fonseca de Souza por tornarem a sala de estudo um ambiente bem agradável. Agradeço ainda a secretária Maria Cristina Peloggia de Araújo sempre muito solícita e agradável. Agradeço ainda a pesquisadora Dr. Rita de Cássia M. Rodrigues.

Agradeço ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) pelo suporte financeiro durante os anos de desenvolvimento deste trabalho.

Agora nada disso seria realmente possível sem a minha família maravilhosa. Assim, ao meu papai Nelcy e minha mamãe Lilian, que felicidade e que honra ser filho de vocês. Meu pai que apesar de não estar mais aqui na terra entre os "vivos", tem sua presença constante e latente no meu coração. A esse homem que sempre "batalhou"muito em prol de sustentar uma família com dignidade, incentivando e ensinado os filhos, um beijão... que saudades!.

A minha mamãe, sempre preocupada comigo, principalmente aqui em São Paulo, uma mulher exemplar que abdicou da sua vida para educar os filhos. Um verdadeira **Mãe**, que sempre praticou os principais conceitos sobre como lidar com as adversidades da vida com ética.

É importante que se diga que o maior ensinamento é o exemplo no dia-a-dia. Assim, a meus pais um muito obrigado pelas várias lições. Aos meus dois manos George e Glauco. Como caçula da família fui criado por todos e particularmente pelos meus dois irmão amados. Agradeço por sempre estarem do meu lado e pela união depois daquele fatídico 3/02/2001. Amo vocês!

Agradeço à minha menina, eterna namorada, companheira, risonha, carinhosa, paciente, guerreira e depois dos sete anos de namoro, quase três meses de noivado, e aproximadamente quarenta minutos de atraso, minha esposa. Que felicidade te ter na minha vida, é impressionante que mesmo nos dias mais cansativos podes me animar com um afago, uma risada ou até mesmo com um daqueles maravilhosos pratos que preparas. Hoje, ao fechar essa porta da minha vida, abrir-se-ão outras que ainda desconheço. Porém, ao teu lado vou tranqüilo, vou com calma, pois, tenho a plena certeza de ter ao meu lado um anjo de DEUS. Te amo!

Sem tu senhor, meu Deus e fiel amigo, não sou nada, nem mesmo pó. Então te

agradeço todos os dias por me conceder o dom da vida, a graça de poder viver com pessoas boas, honestas, e colocar no meu caminho adversidades e ao mesmo tempo me proveres de meios para vencê-las e assim me tornar um homem melhor. Pai divino, mesmo sem merecermos como filhos, agradeço te, pois, por meio do teu filho o mundo ficou puro novamente, do teu espírito ganhamos a sabedoria e do teu amor ganhamos a vida.

Resumo

A modelagem e a análise de desempenho de esquemas de alocação de recursos em redes móveis celulares é o assunto deste trabalho. Nesse sentido, são abordados temas como: geração de modelos de desempenho a partir de uma especificação de alto nível feita usando o método formal *Statecharts*; modelos analíticos para a avaliação de redes hierárquicas multicamadas; estudo do controle de admissão de chamadas e o procedimento de adaptação de largura de banda tão bem como políticas ótimas para esquemas de alocação de recursos adaptativos. Todos os modelos propostos são investigados usando a teoria de Markov.

Abstract

Modelling and performance evaluation of resource allocation schemes are studied in this thesis. In this sense, the following subjects are investigated: generation of performance models from the high level specification Statecharts; performance models for hierarchical cellular mobile networks; call admission control and bandwidth adaptation as well as optimal policy for adaptive resource allocation schemes. All performance models are built in using the theory of Markov chain.

Sumário

Lista de Figuras

Lista de Tabelas

Glossário

3GPP	-	*Third Generation Partnership Project*
8-PSK	-	*8-state Phase Shift Keying*
AIV	-	Algoritmo de Iteração de Valores
AJ	-	Adaptação Justa
AMPS	-	*Advanced Mobile Phone System*
ANSI	-	*American National Standards Institute*
B3G	-	*Beyond 3G*
BB	-	*Bandwidth Broker*
BLER	-	*Block Error Ratio*
RF	-	*Radio Frequency*
BMAP	-	*Batch Markovian Arrival Process*
BSC	-	*Base Station Controller*
CA	-	Com adaptação
CAC	-	Controle de Admissão de chamadas
CDMA	-	*Code Division Multiple Access*
CS	-	*Coding Scheme*
D-AMPS	-	*Digital AMPS*
DBMAP	-	*Discrete Batch Markovian Arrival Process*
DCA	-	*Dynamic Channel Assignment*
DMAP	-	*Discrete Markovian Arrival Process*
DSL	-	*Digital Subscriber Line*
DTBR	-	*Dual threshold bandwidth reservation*

ECSD - *Enhanced Circuit Switched Data*

EDGE - *Enhanced Data rates for GSM Evolution*

EGPRS - *Enhanced GPRS*

EM - Estação Móvel

ERB - Estação Rádio Base

ETSI - *European Telecommunications Standards Institute*

FCA - *Fixed Channel Assignment*

FCC - *Federal Communications Commissions*

FDD - *Frequency Division Duplex*

FDMA - *Frequency Division Multiple Access*

FTP - *File Transfer Protocol*

GMSK - *Gaussian Minimum Shift Keying*

GPRS - *General Packet Radio Service*

GPS - *General Position Sytem*

GSM - *Global System for Mobile Communications*

HSCSD - *High Speed Circuit Switched Data*

IMT2000 - *International Mobile Telecommunications 2000*

IP - *Internet Protocol*

IPP - *Interrupted Poisson Process*

IR - *Incremental Redundancy*

IS - *Interim Standard*

ITU - *International Telecommunications Union*

JDC - *Japanese Digital Cellular*

LA - *Link Adaptation*

LC - *Load Control*

LOTOS - *Language of Temporal Ordering Specification*

LQC - *Link Quality Control*

MAN - *Metropolitan Area Network*

MAP - *Markovian Arrival Process*

MMPP - *Markov Modulated Poisson Process*

MPEG - *Moving Pictures Experts Group*

NMT - *Nordic Mobile Telephone*

nG - Enésima Geração de telefonia móvel celular

NS - *Network Simulator*

OSI/ISO - *Open System Interconnection/ International Standards Organisation*

PDC - *Personal Digital Cellular*

PDCH - *Packet Data Channel*

PLC - *Power Line Communications*

PMD - Processo Markoviano de Decisão

PSMD - Processo Semi-Markoviano de Decisão

QoS - *Quality of Service*

RRM - *Radio Resource Management*

SA - Sem adaptação

SDL - *Specification and Description Language*

SDU - *Service Data Unit*

SIR - *Signal-to-Interference Ratio*

SLA - *Service Level Agreement*

SMS - *Short Message Service*

TCP - *Transmission Control Protocol*

TDD - *Time Division Duplex*

TDMA - *Time Division Multiple Access*

UMTS - *Universal Mobile Telecommunications System*

UTRA - *UMTS Terrestrial Radio Access*

VoIP - Voz sobre IP

WCDMA - *Wide band CDMA*

WLAN - *Wireless Local Area Network*

WWW - *World Wide Web*

Introdução

0.1 Motivação

Nas últimas décadas o mundo vem acompanhando uma revolução nas telecomunicações. O advento das tecnologias de comunicações móveis, o desenvolvimento de serviços multimídias e da Internet mudaram a maneira pela qual as pessoas passaram a se relacionar seja no âmbito profissional ou pessoal.

A comunicação global através de uma única rede, antes possível apenas por meio de um computador pessoal, hoje é realizada perfeitamente com a computação móvel. A capacidade de se poder comunicar em qualquer lugar, momento, e ainda em movimento tem funcionado como uma alavanca para as pesquisas nessa área da engenharia e computação.

O grande passo para a construção desse cenário foi dado quando através da digitalização dos sistemas de comunicação, pôde-se integrar várias mídias em uma mesma infraestrutura. A partir de então, o polarizado mundo das telecomunicações começou a convergir para uma única plataforma incentivado pela redução dos custos de instalação e manutenção. O ganho mercadológico também teve a sua parcela de responsabilidade, pois, houve um aumento no número de acessos dos usuários finais que se motivaram ao perceber que podiam receber todos os serviços em um único terminal.

Com essa integração despontaram as seguintes questões:

- Como alocar ou gerenciar os recursos de modo que os vários serviços possam ser entregues com qualidade?
- Como especificar os requerimentos de qualidade de serviço (QoS) para cada serviço?
- Quais os fatores que limitam o desempenho de cada serviço?

Essas questões começaram a ser respondidas com o desenvolvimento de políticas de controle de admissão de chamadas (CAC) que controlam o acesso à rede das chamadas de modo que um determinado perfil de QoS seja satisfeito.

Paralelamente a esse desenvolvimento, as comunicações móveis evoluíram de geração em geração por meio da adição de novos serviços e funcionalidades de modo a propiciar aos usuários uma extensão da rede fixa. Hoje, as redes de 2.5G e 3G já estão espalhadas pelo mundo entregando conteúdo *Web*, serviços multimídias em tempo real por meio de um limitado espectro de freqüência. Para fazer frente à tamanha demanda por serviços, teve-se início a utilização de estruturas hierárquicas multicamadas através dos quais o usuário pode se movimentar livremente e ainda garantir o seu acesso.

Para melhorar a QoS dos serviços oferecidos, iniciou-se também o desenvolvimento de aplicações multimídias em tempo real que ajustam dinamicamente a sua taxa de bits através de codificadores e decodificadores adaptativos. Nesse sentido, tecnologias como MPEG-2 e MPEG-4 serão fortemente exploradas na provisão da QoS. Assim, será imperativo que as próximas redes explorem essa capacidade de adaptação e melhorem o desempenho dos seus esquemas de alocação de recursos.

Hoje a sinergia entre CAC e o procedimento de adaptação de largura de banda torna a alocação de recursos mais eficiente. Porém, os projetos desses esquemas ficam cada vez mais complexos.

Nesta tese, esses assuntos são abordados profundamente por meio da modelagem de esquemas de alocação de recursos. Como solução para esse problema se destacam duas possíveis linhas: a simulação e a cadeia de Markov.

A simulação é uma ferramenta eficiente para o estudo de sistemas complexos. Porém, o preço a ser pago pela sua escolha reflete-se no tempo computacional quando a exatidão requerida nos resultados é alta.

Na outra face da mesma moeda está a cadeia de Markov. A sua capacidade de produzir soluções rápidas e precisas dado que as considerações feitas na confecção do modelo seguem a propriedade do esquecimento sempre impulsionou seu uso na modelagem de sistemas com cunho em desempenho. Assim, neste trabalho, a cadeia de Markov é considerada como solução para os modelos propostos.

O aumento na complexidade dos esquemas de alocação de recursos em redes móveis torna cada vez mais importante o emprego de uma ferramenta de especificação que torne o projeto desses esquemas mais confiáveis, isto é, mais claros e consistentes. Esse objetivo é normalmente conseguido por meior de um método formal e de preferência visual.

O uso de métodos formais facilita a verificação, validação e documentação do projeto. Neste trabalho, propõe-se o uso de uma técnica para a modelagem da alocação de recursos em redes móveis celulares através de um método formal. A partir dessa especificação pode-se gerar então a solução analítica para o problema em estudo. O método formal escolhido foi o

Statecharts devido à sua peculiaridade na representação de sistemas complexos reativos. Vale ressaltar que o uso dessa metodologia não é novo, porém, a sua aplicação na modelagem de redes móveis celulares sim.

0.2 Contribuições

As contribuições técnicas deste trabalho são:

- Amplo estudo sobre modelagem e análise de desempenho de redes móveis celulares destacando:

    1. Aspectos inerentes a especificação de sistemas complexos reativos e sua aplicação em redes móveis celulares. Nesse contexto, destaca-se uma abordagem sobre modelagem usando a ferramenta de especificação formal *Statecharts*;
    2. O controle de admissão de chamadas (CAC) em redes planares e hierárquicas. Propondo modelos analíticos para a avaliação de desempenho dessa estrutura;
    3. Mecanismo de adaptação de largura de banda e sua atuação juntamente ao CAC. Propondo modelos analíticos para a representação desses procedimentos no mecanismo de alocação de canal;
    4. Otimização de esquemas de alocação de recursos adaptativos que consideram o CAC juntamente ao mecanismo de adaptação de largura de banda em redes móveis celulares usando o Processo Semi-Markoviano de Decisão (PSMD). Propondo modelo para a busca de políticas estacionárias ótimas.

As outras contribuições deste trabalho são relativas às atividades desenvolvidas juntamente aos professores do Programa de Pós-Graduação em Engenharia Elétrica (PPGEE) da UFPA. Assim, sob supervisão dos professores Dr. João Crisóstomo Weyl Albuquerque Costa, Dr.Carlos Renato Lisboa Francês e Dr. Evaldo Gonçalves Pelaes realizaram-se as seguintes atividades de docência:

1. Aulas na disciplina "Redes Móveis Celulares"na graduação do curso de Engenharia Elétrica na Universidade Federal do Pará e no Programa de Pós-graduação em Engenharia Elétrica;
2. Aulas na disciplina "Processo Estocástico"no Programa de Pós-graduação em Engenharia Elétrica sobre o tema assunto teoria de filas;

3. Aulas na disciplina "Análise de Desempenho"no Programa de Pós-graduação em Engenharia Elétrica sobre o assunto teoria de filas, redes de filas e leis operacionais;

4. Participação na definição da dissertação de Mestrado da Engenheira Michelle Bitar Lelis dos Santos, intitulada "Análise de Desempenho de redes GSM/GPRS", que resultou nas publicações (91)(92);

5. Participação na definição da dissertação de Mestrado do Engenheiro Roberto Menezes Rodrigues intitulada "Análise da Qualidade de Serviço em Redes GSM/GPRS através de Esquemas de Compartilhamento de Recursos", que resultou nas publicações (89)(86);

6. Orientação do Trabalho de Conclusão de Curso da Engenheira Elaine Sena Lelis intitulada "Ambiente Computacional para a Análise de Desempenho de Redes Móveis Celulares", que resultou na publicação (93).

0.3 Organização do trabalho

No capítulo 1 são apresentados e discutidos os conceitos básicos sobre redes móveis celulares e alocação de recursos. Esse capítulo é finalizado com a revisão bibliográfica sobre os assuntos abordados e os trabalhos desenvolvidos nesta tese.

A modelagem com cunho em desempenho de sistemas e ainda uma apresentação sobre os conceitos Markovianos fundamentais para a compreensão dos modelos apresentados neste trabalho são assuntos do capítulo 2.

No capítulo 3 apresentam-se os modelo *Statecharts*/Markov para a representação do controle de admissão de chamadas em redes móveis GSM/GPRS. Para mostrar a flexibilidade da técnica usada são propostos três modelos. Posteriormente, uma análise é realizada para verificar o desempenho de cada esquema.

Redes móveis hierárquicas são o assunto do capítulo 4. Nele é proposto um esquema de alocação de recursos que não impacta na QoS dos serviços de voz e ainda melhora o desempenho do serviço de dados. A rede analisada é parametrizada usando a tecnologia (E)GPRS.

O CAC e o mecanismo de adaptação de largura de banda são propostos no capítulo 5. A união desses procedimentos, como mencionado anteriormente, serão a chave para o esquema de alocação de recursos em redes móveis celulares de terceira e quarta gerações, 3G e 4G[2]. Assim, modelos markovianos são propostos para a investigação desses esquemas.

[2]Beyond 3G (B3G) é um outro termo usado para especificar as futuras redes

No capítulo 6 é apresentado uma otimização em esquemas de alocação de recursos adaptativos. O modelo proposto, usando um Processo Semi Markoviano de Decisão (PSMD), busca uma política ótima que minimize a probabilidade de bloqueio, freqüência de adaptação e a insatisfação do usuário a longo prazo.

Por fim, nas conclusões são analisados e discutidos os resultados mostrados no decorrer deste trabalho. Além disso, são sugeridos alguns temas para cada linha de pesquisa abordada nesta tese.

Capítulo 1

Redes Móveis Celulares

1.1 Preliminares

Em um sistema móvel celular, a área na qual o serviço será implantado é dividida em regiões menores chamadas de células. O acesso sem fio é provido em cada célula por meio da Estação Rádio Base (ERB), que dependendo da tecnologia de rede, pode ser denominada de *Base Transceiver Station* (BTS) como no GSM, Nó B como no WCDMA, etc. O assinante do serviço, usuário, comunica-se pelo enlace de subida com a rede usando uma Estação Móvel (EM), que pode variar de um simples telefone celular à aparelhos multifuncionais contendo câmera digital, aplicativos de videoconferência, etc.. Por sua vez, a ERB, comunica-se com o usuário através do enlace de descida. Quando um usuário se movimenta através das células é realizado um procedimento de *hand off* ou *handover*, o qual é responsável pela continuidade do serviço. Se existirem recursos de rádio suficientes na célula na qual o assinante está entrando, a sua chamada é aceita, caso contrário, é bloqueada. A Fig.(1.1) mostra uma típica rede móvel celular destacando os elementos mencionados acima.

Devido a escassez do espectro de freqüência destinado à comunicação móvel é necessário que os recursos de rádio sejam compartilhados de forma a aceitar o maior número possível de assinantes. O que é desejável, uma vez que, isso resulta em uma maximização da utilização dos canais e conseqüentemente em uma maximização da receita da operadora. Há três formas básicas de se realizar o múltiplo acesso:

- Múltiplo Acesso por Divisão de Freqüência (FDMA);
- Múltiplo Acesso por Divisão de Tempo (TDMA);
- Múltiplo Acesso por Divisão de Código (CDMA).

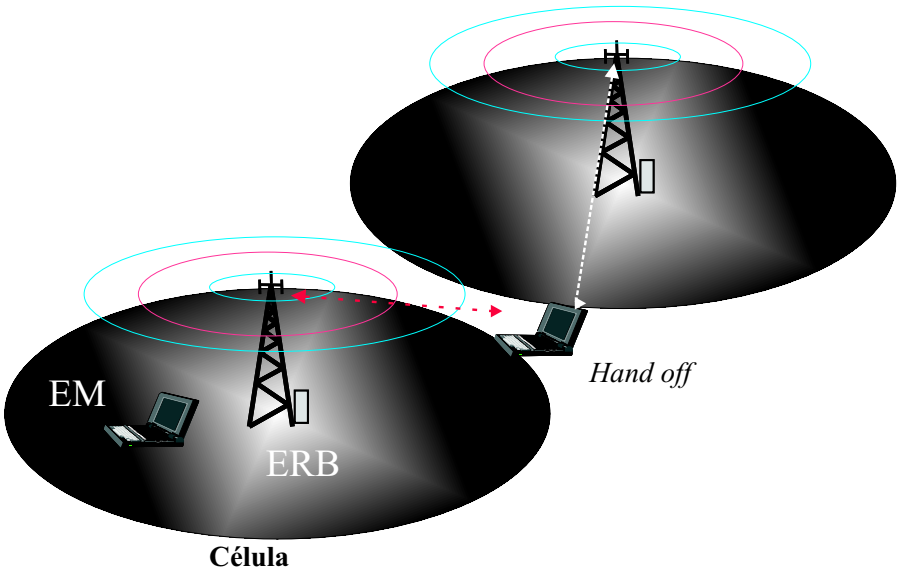
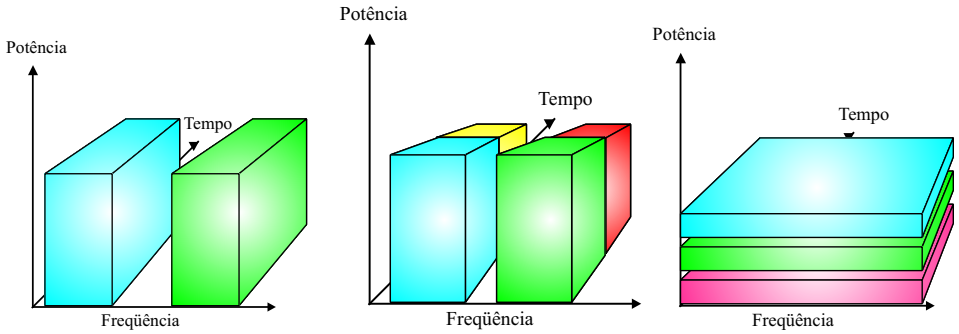


Figura 1.1: Sistema celular: célula, ERB, EM e *hand off*.

No FDMA, o espectro de freqüência é dividido em um determinado número de canais que são atribuídos aos usuários. No TDMA, cada canal FDMA é dividido em porções de tempo chamadas de *slots*, que são atribuídos aos usuários. O número de *slots* em cada canal de rádio depende da tecnologia de rede implantada pela operadora de serviço. Por exemplo, o D-AMPS possibilita três *slots* de tempo por canal de rádio, enquanto que o GSM, oito. O CDMA é uma técnica de múltiplo acesso no qual todos os usuários compartilham a mesma faixa de freqüência para a transmissão do seu serviço, sendo que, a distinção entre cada um é feita através de um código. A Fig.(1.2) mostra os três esquemas de múltiplo acesso descritos.



(a) (b) (c)

Figura 1.2: Múltiplo acesso: (a) FDMA, (b) TDMA e (c) CDMA.

A Fig.(1.3) mostra um ambiente celular multicamada, onde células de diferentes tamanhos ou áreas de cobertura são organizadas em uma estrutura hierárquica. A cobertura global é provida por uma rede de satélites. Assim, por exemplo, uma rede de três satélites de

órbita geo estacionária pode enviar e receber sinais através de quase toda a Terra fornecendo uma comunicação global com poucos *hand off* (1). Todavia, o longo atraso na transmissão torna essa solução inadeqüada para a prestação de serviços em tempo real. A utilização de satélites com órbitas baixas e fixas é uma alternativa para a solução desse problema (2). As macrocélulas possuem raio na ordem de 1 à 10 km e são usadas para cobrir áreas urbanas com uma densidade de tráfego baixa ou média. Da mesma forma, as microcélulas também são usadas em ambientes externos, porém com cobertura reduzida, possuindo raio na ordem de 200m à 2km. Sua densidade de tráfego varia de média para alta. Diferente das anteriores, ambientes picocélulares são usados para a cobertura de ambientes internos tipicamente na faixa de 10 à 200m de raio.

Além de possibilitar uma comunicação sem fio por uma extensa área geográfica, uma arquitetura multicamada hierárquica aumenta a capacidade da rede móvel celular. Isso porque, com células menores, a distância de reuso de freqüência é reduzida. Então, o mesmo canal de rádio pode ser usado mais vezes resultando em um aumento da capacidade da rede.

Outra vantagem dessa arquitetura está relacionada à otimização da eficiência espectral (melhor uso dos recursos de rádio) por meio de dois mecanismos: o transbordo e o retorno. Em situações de sobrecarga na rede, uma chamada que solicitou um serviço em uma célula onde não existem recursos de rádio disponíveis pode ser transbordada para células em camadas superiores ou inferiores. Caso não existam recursos nessas células, essa chamada é bloqueada (2)(3). Uma vez que existe uma rota alternativa para o escoamento do tráfego oferecido, a probabilidade de bloqueio nessas redes é bem menor que em uma rede planar (2). Apresentando recursos disponíveis na camada em que essa chamada originalmente solicitou serviço, ela pode retornar, liberando os canais de rádio da célula para onde a mesma foi transbordada (2).

1.2 Evolução

1.2.1 Primeira Geração (1G)

A primeira geração de redes móveis celulares foi caracterizada pela transmissão do serviço de voz de forma analógica. Os primeiros sistemas se tornaram operacionais no início da década de 80. Na Escadinávia o *Nordic Mobile Telephone* (NMT) foi comercialmente lançado em 1981 operando nas faixas de 450 MHz e 900 MHz. Em 1982, no Reino Unido, foi desenvolvido o *Total Access Communications System* (TACS), o qual operava na faixa de 900 MHz. Ainda em 1982, foi lançado nos Estados Unidos o *Advanced Mobile Phone*

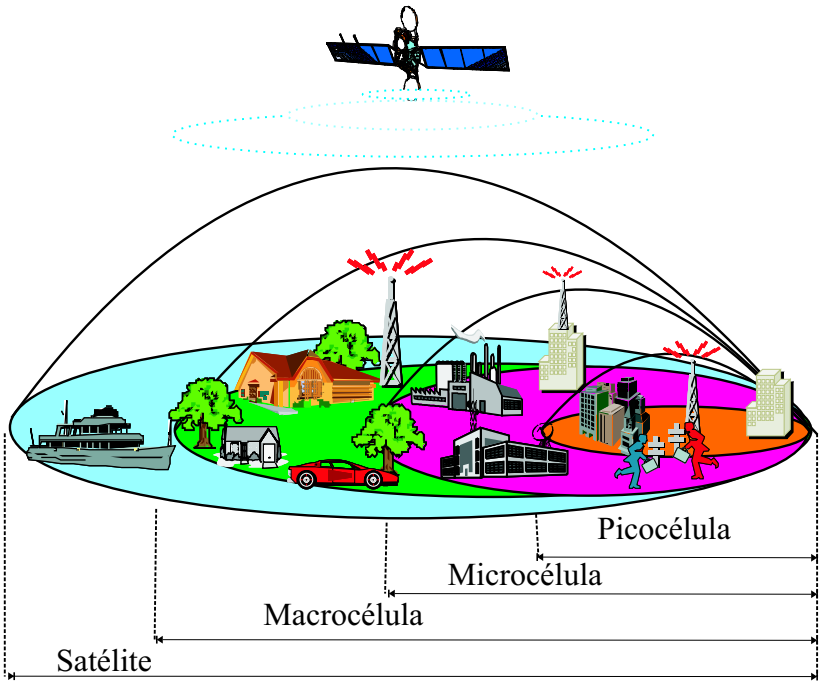


Figura 1.3: Rede hierárquica para cobertura global

System (AMPS) operando entre as faixas de 800 MHz e 900MHz. Outros sistemas de primeira geração são: o *C-System* na Alemanha, operando nas faixas de 450MHz e 900MHz; o JTCAS implantado no Japão (4).

A 1G foi um sucesso no sentido de revelar a existência de uma demanda por serviços sem fio apesar de problemas como: incompatibilidade entre os sistemas dos diversos países, baixa qualidade do serviço conversacional, ausência de segurança, capacidade limitada, freqüente queda de chamadas, etc. (5).

1.2.2 Segunda Geração e mais (2G - 2,5G)

A segunda geração de redes móveis celulares, começou a ser concebida quando os países da Europa perceberam que a incompatibilidade entre os diversos sistemas segmentava o mercado e limitava o movimento dos usuários. Assim, em 1982, foi formado um grupo chamado de *Group Speciale Mobile* (GSM) para criar um sistema que permitisse operar em toda a Europa de forma unificada. Posteriormente, por razões mercadológicas, o GSM passou a se chamar *Global System for Mobile Communications*.

Desde que a primeira rede foi lançada comercialmente, o GSM se transformou no principal padrão de comunicação sem fio do mundo (6). No fim de janeiro de 2004 havia aproximadamente cerca de um bilhão de assinantes em mais de duzentos países do mundo (7).

Inicialmente ele foi projetado para cobrir grandes áreas, macrocélula, operando na faixa de 900MHz. Posteriormente, mais duas versões nas faixas de 1800 MHz e 1900 MHz foram incorporadas a esse padrão para operar na Europa e na América, respectivamente. Ambas usando terminais com uma baixa potência em ambientes microcelulares (6). Sendo uma tecnologia totalmente digital, o GSM possibilitou o *roaming*, isto é, que um assinante de uma rede possa usar os serviços de uma outra rede GSM no mundo. A Fig.(1.4) mostra a abrangência mundial do GSM (8).

Figura 1.4: Abrangência do GSM.

No Estados Unidos, o desenvolvimento da 2G trilhou os seguintes caminhos: O Digital-AMPS (DAMPS) foi lançado como o sucessor do AMPS seguindo uma linha TDMA. A atualização consistia em introduzir três *slots* de tempo por freqüência do AMPS aumentando a capacidade e melhorando o desempenho da rede. O D-AMPS também foi conhecido como IS-54, o qual com a adição de novos serviços passou a se chamar IS-136. O *US Federal Communications Commission* (FCC) permitiu que o GSM entrasse no mercado Norte Americano operando na faixa de 1900 MHz. O IS-95 ou CdmaOne foi um outro padrão adotado. Operando em modo dual (analógico e digital), ele utiliza uma banda de 1,25MHz que pode ser usado para o acesso simultâneo de vários usuários (6). Além disso, células adjacentes podem usar a mesma banda de freqüência, o que simplifica o planejamento. Outras novidades inseridas pelo IS-95 foram: o sincronismo entre as ERBs através do GPS; o *soft handoff*, onde a desconexão com a célula de origem somente é feita quando ele completa a conexão com a célula de destino (5).

O Japão desenvolveu uma tecnologia chamada de *Personal Digital Cellular*(PDC) que originalmente foi denominada de *Japanese Digital Cellular* (JDC). Esse padrão é bastante

similar ao IS-54/136 utilizando o mesmo esquema de modulação, duração do quadro TDMA, etc.. Porém, o espaçamento do canal e os codificadores de voz são diferentes (5).

Na 2G as redes móveis celulares ofereciam além do serviço de voz, serviços de dados com baixas taxas (9,6 Kbits/s e 14,4 K bits/s), fax, Serviço de Mensagens Curtas-(*Short Message Service*- SMS), além de outros serviços suplementares. Todavia, a popularização de serviços como correio eletrônico, WWW, ftp, e a saturação do mercado de voz, tornaram a convergência entre a telefonia móvel celular e a Internet imprescindível (4). Porém, o desenvolvimento de uma nova geração exigiria anos e altíssimos investimentos. Assim, foi necessário a criação de novas tecnologias que servissem como "pontes"entre a 2G e a Terceira Geração (3G). Alguns desses sistemas, chamados de 2,5G, são *High Speed Circuit Switched Data* (HSCSD), *General Packet Radio Service* (GPRS) e IS-95B.

O HSCSD e o GPRS são evoluções diretas do GSM. A primeira, também comutada a circuito, prove taxas de até 57,6 kbit/s considerando uma operação *multislot*. Contudo, a ausência de multiplexagem estatística, garantias de QoS durante operações de *hand off* e o alto custo do serviço limitaram a aceitação dessa tecnologia.

O GPRS emprega a tecnologia de comutação por pacote para melhorar e simplificar o acesso sem fio à redes de dados externas, como Internet, Intranet, X.25 e X.75, proporcionando aos usuários altas taxas de transmissão e o estabelecimento de uma sessão em alguns segundos (6)(9)(10). A conexão com redes externas é baseada no protocolo IP. Ele utiliza o conceito de alocação sob demanda, onde os canais de rádio ociosos, podem ser usados pelo GPRS para aumentar a QoS. Contudo, mediante uma chegada de um serviço de maior prioridade, esses recursos são liberados. A tarifação no GPRS é baseada no volume de informação trafegada na inteface aérea, diferente dos serviços de dados comutados a circuito, onde a cobrança é baseada na duração da conexão.

Do ponto de vista tecnológico, a comutação por circuito degrada o desempenho do sistema devido à retenção do canal de rádio por toda a duração do serviço. Esse tipo de alocação de recurso é ineficiente para o tráfego em rajada, como o tráfego Internet. Nesse caso, a comutação por pacote traz resultados muito melhores, pois os canais de rádio são alocados somente quando há transferência de dados através da interface aérea (9)(10). Uma vez que a transmissão é terminada, o canal de rádio é imediatamente liberado.

Um outro aspecto importante em relação ao serviço de dados é a assimetria do tráfego ou o desbalanceamento entre as cargas dos enlaces de subida e descida. Isso acontece pois, enquanto a solicitação de um serviço *Web* é encaminhada pelo enlace de subida, todo conteúdo do serviço, normalmente formado por um documento HTML, applet Java, imagens, etc., é escoado através do enlace de descida. Na comutação por circuito os canais desses

enlaces são alocados conjuntamente e permanentemente durante todo o tempo do serviço, o que, devido a assimetria do tráfego, diminui as suas utilizações. O GPRS permitiu a alocação independente desses canais de rádio.

A integridade dos dados através da interface aérea é garantida por meio dos seguintes esquemas de codificação CS-1 (9,05 Kbit/s), CS-2 (13,4 Kbit/s), CS-3 (15,6 Kbit/s) e CS-4 (21,4 Kbit/s).

Nesse cenário de evolução da 2G para 3G uma outra tecnologia despontou como uma solução para sistemas GSM e IS-136. O EDGE é uma tecnologia de acesso TDMA que possibilita a oferta de serviços 3G nos espectros existentes de 800, 900, 1800 e 1900 MHz. Devido a ao reuso de freqüência, não é necessário a aquisição de novas licenças, o que o torna uma solução de baixo custo. Além disso, ele pode ser usado com o UMTS por operadoras que possuam uma rede GSM/GPRS para fornecer serviços 3G em células onde a implantação do UMTS não for economicamente viável (11). A Fig.(1.5) mostra o cenário mundial do EDGE (7).

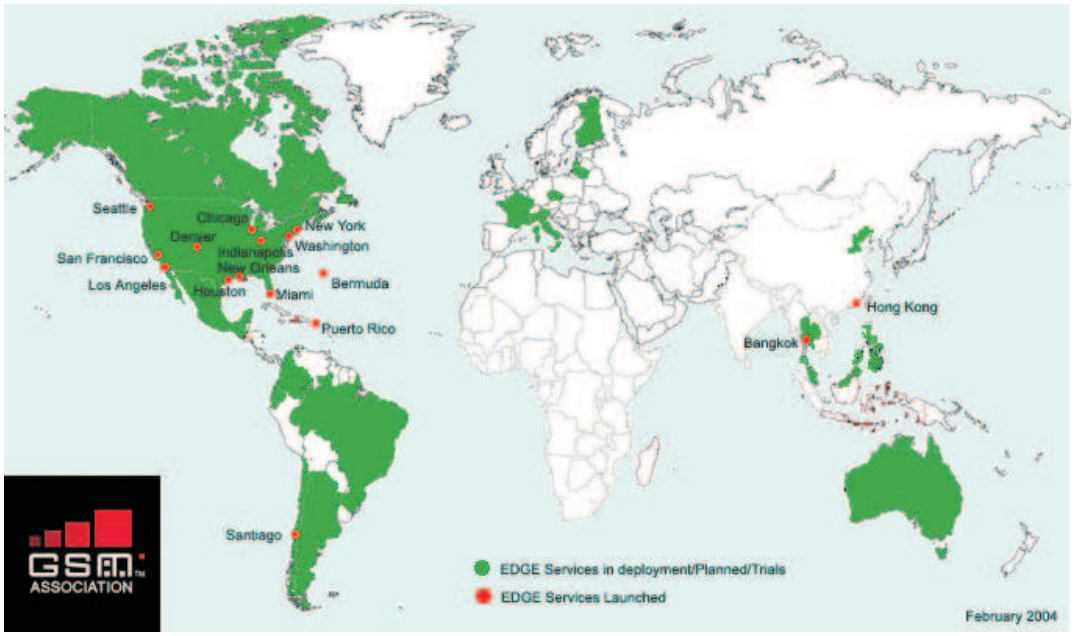


Figura 1.5: Cenário mundial do EDGE

A sua padronização define mudanças na interface aérea da rede GSM/GPRS, as quais aprimoram a prestação do serviço de dados na parte comutada a circuito e a pacote. Essas passam a se chamar *Enhanced* CSD (ECSD) e *Enhanced* GPRS (EGPRS).

As altas taxas de transmissão alcançadas pelo EDGE são conseguidas por meio da modulação *eight-state phase-shift keying* (8-PSK) e novos esquemas de codificação que possibilitam a adaptação da proteção de dados (redundância) à qualidade do canal (12). Além

disso, ele aperfeiçoa o mecanismo de controle de qualidade do enlace (*Link Quality Control*-LQC) usado pelo GPRS introduzindo um esquema ARQ chamado incremento da redundância (*Incremental Redundancy* - IR). No GPRS o LQC usa somente o procedimento adaptação de enlace (Link adaptation- LA), que estima regularmente a qualidade do canal, e então, seleciona o esquema de codificação adequado. No EDGE é ainda selecionado a modulação, 8-PSK ou GMSK, de forma a maximizar a taxa de transmissão. No IR a informação é enviada com pouca proteção visando uma alta taxa. Caso a decodificação não ocorra, será enviado mais redundância. Quando os mecanismos LA/IR trabalham de forma híbrida, a taxa de *bits* é maximizada durante as variações da SIR, balanceando a taxa de erro de bloco (BLER). Isso resulta na utilização de um esquema de codificação de canal em uma grande parte da célula (13).

O impacto da introdução do EDGE em redes GSM/GPRS está limitado a interface aérea, mais especificamente a camada física e de enlace. A sua implantação pode ser feita gradualmente através dos equipamentos apropriados, transmissores e receptores, nas células ou em setores das mesmas. Com isso toda a *core network* é reaproveitada. Uma célula com o EDGE contém quatro tipos de canais físicos (12):

1. Canais físicos de voz e dados comutados a circuito GSM/CSD;
2. Canal físico de dados comutados a pacote GPRS;
3. Canais físicos GSM/CSD/ECSD;
4. Canais físicos GPRS/EGPRS.

O procedimento de gerência dos recursos de rádio deve atuar dinamicamente para definir quais tipos de canais devem ser usados para atender uma dada demanda de voz ou dados. Por exemplo, se o número de usuários de dados aumenta, o número de canais do tipo 2 e 4 deve aumentar automaticamente. Com isso, evita-se a reserva de recursos que normalmente diminui a utilização dos canais de RF.

1.2.3 Terceira Geração (3G)

A 3G de sistemas móveis celulares é um conjunto de normas, especificações e padronizações chamado de *International Mobile Telecommunications-2000* (IMT-2000) definido sob a supervisão do *International Telecommunications Union* (ITU). Dentre os requerimentos do IMT-2000 estão (1) (5)(6)(14):

1. Transmissão de dados em altas taxas:
   - 144 kbit/s em ambientes macrocelulares;
   - 384 kbit/s em ambientes macrocelulares e microcelulares;
   - 2 Mbit/s em ambientes microcelulares e picocelulares.
2. Suporte à transmissão de dados simetricamente e assimetricamente. A oferta de serviços assimétricos como correio eletrônico, WWW, tão bem quanto novos serviços multimídia que demandam uma taxa de transmissão igual entre os enlaces de subida e descida é um dos pilares fundamentais dos sistemas de 3G;
3. Prover uma qualidade do serviço conversacional comparável àquela de redes cabeadas;
4. Aumento da capacidade. Devido à necessidade de atender à crescente demanda por serviços, torna-se imperativo que a utilização dos recursos de rádio seja a melhor possível;
5. Acesso simultâneo a múltiplos serviços. Possibilita que um assinante faça um *download* de um arquivo MP3 durante uma conversação;
6. *Roaming* global;
7. Aumento da segurança. Capacita a oferta de diversas formas de *m-commerce*;
8. Flexibilidade. Suporte de serviços comutados a circuito e a pacote.
9. Compatibilidade com os sistemas de 2G. Isso fez com que as organizações internacionais com diferentes iniciativas se unissem de forma a criar sistemas que tornassem a migração dos sistemas de 2G para os de 3G o mais suave possível, ou seja, com o menor custo e máximo reaproveitamento do investimento feito em equipamentos, instalação, equipe técnica, etc.

1.2.3.1 Alianças da 3G - 3GPP e 3GPP2

O propósito do 3G *Partnership Project* (3GPP) é preparar, aprovar e manter especificações técnicas aplicáveis globalmente além de relatórios técnicos para sistemas de 3G partindo diretamente do GSM. Para isso, as seguintes organizações de desenvolvimento de padrões se uniram em 1998: ARIB e TTC (Japão), ETSI (Europa), T1 (Estados Unidos) e TTA (Coréia).

Seguindo a mesma iniciativa, liderados pelo *American National Standards Institute* (ANSI), em janeiro de 1999 as quatro seguintes organizações internacionais criaram o 3GPP2

com a meta de criar especificações técnicas para sistemas evoluídos do IS-95A e B: ARIB, TIA (Estados Unidos), TTA e TTC. É importante mencionar que as especificações do 3GPP e 3GPP2 não incluem apenas os aspectos da interface aéria, mas também, outros relacionados com a *core network*, interconexão com outras redes, etc.

Assim, apesar das diferentes abordagens em termos de evolução da 2G para a 3G, o consenso geral das alianças foi conceber meios para os quais se tenha uma harmonização e consolidação de especificações de interfaces aéreas banda larga usando o CDMA (14).

1.2.3.2 *Universal Mobile Telecommunication System* (UMTS)

O candidato europeu do IMT-2000 foi o UMTS. O acesso de rádio terrestre do UMTS chamado de UMTS *Terrestrial Radio Access* (UTRA) foi aprovado pelo ITU em maio de 2000. No UMTS a interface aérea é o CDMA de banda larga (WCDMA) com aproximadamente 5 MHz por portadora [1]. Devido ao fator de reuso unitário, as mesmas portadoras podem ser usadas em células adjacentes.

O UMTS suporta dois modos de operação: *Frequency* e *Time Division Duplex* FDD e TDD, respectivamente. No modo FDD a conexão é feita usando diferentes portadoras nos enlaces de subida e descida. Cada portadora de 5MHz é dividida em quadros de 10 ms, sendo que cada quadro é posteriormente fragmentado em 15 *slots* de tempo. A taxa de *chip* especificada é de 3,84 Mcps [2] (15). A cada segundo 3,84 milhões de *chips* são enviados através da interface de rádio. Todavia, o número de bits transmitidos durante o mesmo período é menor. A relação entre o número de *chips* e *bits* transmitidos é chamado fator de espalhamento. No modo TDD, os enlaces de subida e descida são implementados na mesma portadora. Os quinze *slots* de tempo podem ser dinamicamente alocados nas direções de subida e descida. A taxa de *chip* é a mesma usada no modo UTRA-FDD.

No UMTS a provisão da qualidade de serviço (QoS) está relacionada ao tipo de tráfego na rede. Assim, são definidas as seguintes classes de QoS, também chamadas de classes de tráfego, as quais consideram as restrições e limitações da interface aérea (4)(16)(17):

1. Classe Conversacional: Essa classe se aplica a qualquer serviço em tempo real entre pessoas ou grupos de pessoas. Portanto, devem ser mantidos valores baixos para o atraso, variações no atraso (*jitter*), e perda dos quadros. Exemplos são o serviço de

[1] O termo WCDMA é empregado para diferenciar os acessos CDMA do UMTS e do CdmaOne. Uma vez que, o primeiro usa aproximadamente três vez mais largura de banda que o segundo

[2] Um *chip* é um *bit* em um código usado para modular uma informação. O termo é utilizado pois ele não representa nenhuma informação por si só .

telefonia, voz sobre IP (VoIP), videofone, videoconferência, telemetria, jogos interativos, etc. A Tabela 1.1 mostra alguns exemplos de serviços conversacionais tão bem quanto suas características de QoS (16).

2. Classe *Streaming*: Essa classe representa as aplicações de fluxo contínuo de áudio e vídeo. Exemplos dessa classe são: rádio Web, vídeo sob demanda, etc. Devido a ausência de interação humana, há uma menor exigência de baixos atrasos, porém existe a necessidade de sincronização da mídia e o jitter deve ser mantido baixo. A Tabela 1.2 mostra alguns exemplos de serviços da classe *streaming* tão bem quanto suas características de QoS (16).

3. Classe Interativa: Essa classe é definida para aplicações cliente e servidor, onde o usuário final (cliente) representa um homem ou uma máquina. Exemplos dessa classe são: serviço *Web*, servidor de acesso, sistemas de gerência de rede, etc. Apesar de mais tolerantes a atrasos, essa classe de tráfego exige uma baixa taxa de erro de *bit*, de forma que é necessário a utilização de algum mecanismo de retransmissão de pacote. A Tabela 1.3 mostra alguns exemplos de serviços da classe interativa tão bem quanto suas características de QoS (16).

4. Classe de Fundo (*background*): Essa classe está relacionada as aplicações com mais baixa prioridade nas quais os dados são enviados ou recebidos passivamente ou ativamente por um computador e onde o atraso não é um requisito importante. Exemplos de aplicações dessa classe são: correio eletrônico, SMS, transferência de arquivos, etc. A Tabela 1.4 mostra alguns exemplos de serviços da classe de fundo tão bem quanto suas características de QoS (4).

A principal diferença entre essas classes de QoS é a tolerância ao atraso. As duas primeiras são destinadas à aplicações em tempo real, sendo portanto, mais sensíveis a atrasos. As demais possuem uma sensibilidade menor, e assim, são dirigidas a aplicações na qual o tempo não é um fator crítico. De uma forma geral, a classe conversacional é aquela que possui maiores restrições de QoS, enquanto que, a classe de fundo possui as menores restrições.

O UMTS definiu os seguintes atributos de QoS (4)(16)(18):

- Taxa de *bit* máxima: equivalente à taxa máxima;
- Taxa de *bit* garantida:equivalente à taxa média;
- Ordem de entrega: esse parâmetro especifica se os pacotes fora de seqüências são aceitos ou descartados;

Tabela 1.1: Serviços conversacionais em tempo real.

Mídia	Aplicação	Simetria	Taxa de dados (Kbps)	Atraso fim-a-fim (ms)	Variação no atraso (ms)	Perda
Áudio	Serviço de telefonia	bidirecional	4-25	< 150 preferencial <400 limite	< 1	< 3% taxa de erro de quadro
Vídeo	Videofone	bidirecional	32-384	< 150 preferencial <400 limite Sincronizador de lábios<100		< 1% taxa de erro de quadro
Dados	Telemetria	bidirecional	< 28,8	< 250	NA	Zero
Dados	Jogos interativos	bidirecional	< 1	< 250	NA	Zero
Dados	*Telnet*	bidirecional	< 1	< 250	NA	Zero

- Comprimento máximo do SDU: usado para controle de admissão;
- Taxa de erro de *bit* residual: indica a taxa de erro de *bit* não detectada, ou ainda, se nenhuma detecção de erro é solicitada, ele indica a taxa de erro de *bit* do SDU;
- Taxa de erro do SDU: indica a fração de SDUs perdidos ou com erro. Esse atributo é usado em esquemas de detecção de erro;
- Atraso na transferência: indica o máximo atraso para 95^{th} da distribuição do atraso para todos os SDUs dentro da rede UMTS;
- Prioridade de tráfego: é definido para diferenciar os tipos de tráfego pertencentes a classe interativa. Ele é usado no escalonamento nos nós da rede UMTS.

A tabela 1.5 mostra os atributos da QoS para cada classe de tráfego definido pelo UMTS. Deve ser notado que a própria classe de tráfego é definida como um atributo (17).

1.2.3.3 CDMA2000

O outro candidato à 3G de sistemas móveis celulares é o CDMA2000. Esse padrão segue a linha de evolução do CdmaOne. Enquanto o WCDMA é totalmente assíncrono o CDMA2000, assim como seu antecessor, é síncrono. A taxa de *chip* no CDMA2000 varia em

Tabela 1.2: Serviços *streaming* em tempo real.

Mídia	Aplicação	Simetria	Taxa de dados (Kbps)	Atraso na inicialização (s)	Variação no atraso no transporte (s)	Perda na camada de sessão
Áudio	Voz/música Média/alta qualidade	unidirecional	5-128	< 10	< 2	< 1% taxa de perda do pacote
Vídeo	Filmes em tempo real	unidirecional	20-384	< 10	< 2	< 2% taxa de perda do pacote
Dados	Transferência de dados em grande quantidade, informação de sincronização	unidirecional	< 384	< 10	NA	Zero
Dados	Jogos interativos	unidirecional		< 10	NA	Zero

diferentes graus de multiplicidade em relação a taxa básica de 1,2288 Mcps. Inicialmente, esses valores são de 1.2288 Mcps à 3.6864 Mcps o que é muito próximo da taxa do UTRAN.

Na primeira fase, o CDMA2000 ou 1x suporta taxas de 384 kbits/s para transmissão de dados e duplica a capacidade de atendimento de serviços de voz. Na segunda fase, duas variações do 1x são propostas: o 1xEV-DO e 1xEV-DV. O primeiro, suporta somente a transmissão de dados (DO-data only), permitindo taxas acima de 2MHz. O problema dessa tecnologia é o emprego de uma técnica de compartilhamento de recursos de rádio chamada de partição completa. Assim se a rede está congestionada com serviços de voz e com uma carga de tráfego de dados baixa, os recursos ociosos de rádio que são destinados para o escoamento do tráfego de dados não podem ser usados para atender as chamadas de voz diminuindo a utilização dos recursos de rádio. A outra tecnologia EV-DV é mais flexível misturando o tráfego de voz e dados na mesma portadora (4)

O CDMA2000 não define explicitamente classes de QoS, tal como feito pelo UMTS, contudo, na prática, ele suporta os mesmos serviços com os mesmos atributos (16)(4).

Segundo o ***CDMA Development Group***, o CDMA2000 foi a primeira rede de 3G comercialmente implantada em outubro de 2000. Atualmente cerca de 94 operadoras trabalham com o CDMA2000 1X e 13 com o 1xEV-DO na Ásia, América e Europa. Sendo que mais 21 redes 1x e 17 redes 1xEV-DO estão agendadas para 2004 (19).

Tabela 1.3: Serviços Interativos.

Mídia	Aplicação	Simetria	Taxa de dados (Kbps)	Atraso unidirecional (s)	Variação no atraso (ms)	Perda
Áudio	Mensagem de voz	unidirecional	4-13	< 1 por *playback* <2 por *record*	< 1	< 3% taxa de erro de quadro
Dados	*Web browsing*-HTML	unidirecional		< 4 por página	NA	Zero
Dados	Serviços de transação de alta prioridade ex.:ATM, comércio eletrônico	bidirecional		< 4	NA	Zero
Dados	correio eletrônico	unidirecional		< 4	NA	Zero

Tabela 1.4: Serviços de fundo.

Mídia	Aplicação	Taxa de dados (Kbps)	Atraso unidirecional (s)	Variação no atraso (s)	Perda
Dados	serviço de mensagens curtas (SMS)		< 30	NA	Zero
Dados	FAX		< 30	NA	Zero
Dados	correio eletrônico (entre servidores)		variável	NA	Zero

1.2.4 Quarta Geração (4G)

Tradicionalmente, a concepção e o desenvolvimento da geração vindoura de uma tecnologia se inicia de 10 à 20 anos antes de sua implantação, tal qual ocorrido anteriormente na transição entre as segunda e terceira gerações de redes móveis celulares. Assim, enquanto a 3G está consolidada[3], a pesquisa na 4G se encontra em um estágio embrionário. Porém, seguindo metas que parecem refletir um consenso geral expresso através da palavra integração.

Essa integração tem um caráter abrangente, pois, ela pode ser visualizada ou contextualizada sob vários aspectos ou expectativas como as dos usuários, serviços (aplicações), terminais móveis e das redes. De uma forma geral, a 4G deve ser capaz de (20) (21)(22):

[3]Esse termo é usado no sentido de que as tecnologias estão definidas, embora, existam várias questões em aberto.

Tabela 1.5: Atributos de QoS para cada classe de tráfego.

Classe de tráfego	Conversacional	*Streaming*	Interativa	Fundo
Taxa de*bits* máxima	x	x	x	x
Taxa de*bits* garantida	x	x		
Ordem de entrega	x	x	x	x
Comprimento máximo do SDU	x	x	x	x
Taxa de erro de *bit* residual	x	x	x	x
Taxa de erro do SDU	x	x	x	x
Atraso na transferência	x	x		
Prioridade de tráfego			x	

- Atender aos anseios dos usuários. Tal feito é bastante complexo pois requer que vários serviços e terminais correspondam às diferentes características tais como: social, educacional, cultural, tão bem como, dos aspectos inerentes de cada indivíduo como visão, audição, tato, voz;

- Para corresponder à essas expectativas, os terminais devem prover diferentes interfaces (teclados alfa-numéricos, reconhecimento de locutor, etc.); possuir diferentes características físicas (formatos, tamanhos, peso, etc.) e tecnológicas (antena inteligentes, *software radio*, *codecs* adaptativos, tempo de vida da bateria, operando em baixa e alta velocidade, etc.), que refletirá diferentes custos. Além disso, será imperativo que o terminal possa operar ininterruptamente através de diversas redes.

- Prover o suporte a mobilidade do usuário por meio de uma tecnologia de transporte "transparente". Em outras palavras, o usuário deve ter o acesso ao serviço, através do seu terminal personalizado, independente da rede, legada ou não (1G, 2G, 3G, WLAN, MAN) e do seu perfil de mobilidade. Além disso, todas as redes e sub-redes deverão ser capazes de interoperar entre si independentemente da tecnologia de acesso, via rádio ou cabeado (AMPS, CdmaOne, GSM, GPRS, EDGE, WCDMA, CDMA2000, IEEE 802.11, HiperLAN/2, HPNA, PLC, USB, IEEE 1394, xDSL, HFC, Powerline, *Bluetooth*). Outra característica desejável é que a rede seja flexível o bastante para se ajustar dinamicamente às condições do canal (modulação/ demodulação, codificação, antenas inteligentes, etc.) de tráfego e ao serviço. Observe que algumas características da rede também são comuns as dos terminais. Além disso, a rede deve suportar o tráfego multimídia e ainda ser capaz

de prover altas taxas e garantias de QoS as aplicações (Serviço diferenciado-*Diffserv*). Além disso, devido a característica hierarquizada desse sistema, a rede deve ser capaz de suportar o *handoff*, transbordo e retorno entre as diversas camadas da rede.

- Nesse leque de requerimentos se inclui ainda os serviços ou aplicações que devem ser multiplataforma (capacidade de operar em diferentes terminais), personalizadas (adequadas ao usuário e pelo usuário), adaptáveis as características da rede (diferentes *codecs* podem ajustar a sua taxa de transmissão de acordo com o tráfego da rede levando a uma maximização da utilização dos recursos de rádio) e inteligentes o suficiente para negociar qual a melhor rede ou camada da mesma lhe proverá o acesso.

1.3 Alocação de recursos

Na modelagem com cunho em desempenho de esquemas de alocação de recursos em redes móveis celulares, o termo "recurso"se refere aos canais de rádio disponibilizados por uma Operadora de Serviço e que são distribuídos entre às células durante o planejamento do sistema para o atendimento de uma dada demanda. Para aumentar a capacidade do sistema, esses canais normalmente são alocados aos assinantes por meio de alguma técnica de múltiplo acesso.

Como o espectro de freqüência destinado à telefonia móvel celular é um recurso limitado e dispendioso, a alocação de recursos, sob responsabilidade da gerência de recursos de rádio (RRM-*Radio Resource Management*), deve ser feita de forma a estabelecer um compromisso entre utilização dos canais e a satisfação dos usuários.

A solução dessa equação é uma tarefa complexa, visto que os objetivos expostos acima são conflitantes, pois, as operadoras almejam atender o maior número possível de usuários, e esses, esperam a melhor qualidade possível no serviço. Assim, achar um esquema que equilibre esses objetivos é a meta da análise de desempenho.

1.3.1 Tipos de alocação de recursos

Dependendo da forma como os canais são distribuídos, tem-se os seguintes tipos de alocação de canal: fixa, dinâmica e híbrida.

1.3.1.1 Alocação de canal fixa (FCA)

Nesse esquema, a área a ser atendida pela operadora de serviço é dividida em células, sendo que, cada uma recebe um determinado número de canais de acordo com um padrão de reuso, o que resulta em uma capacidade de transmissão fixa.

Uma forma simples de implementá-la é através da distribuição uniforme dos canais entre as células. Assim, a probabilidade de bloqueio de uma chamada em uma célula é a mesma da rede. Obviamente, essa alocação somente será eficiente quando o tráfego oferecido também for distribuído uniformemente (23).

Porém, via de regra, o tráfego oferecido em uma rede móvel celular apresenta flutuações espaciais e temporais, o que, resultará em uma pobre utilização dos recursos de rádio, caso uma alocação fixa uniforme seja usada.

Em uma alocação de canal não uniforme, a quantidade de canais de rádio alocados para cada célula é função da demanda da rede e do perfil de QoS. Conseqüentemente, células com uma alta carga de tráfego receberão mais canais, e vice-versa. A função da análise de desempenho, durante o planejamento da rede, é identificar quantos canais devem ser atribuídos para cada célula.

Devido à natureza fixa dessa alocação, a utilização dos canais de rádio será melhor quanto maior for o tráfego na célula. Caso contrário, os mesmos ficaram ociosos e indisponibilizados para o uso em outras células.

1.3.1.2 Alocação de canal dinâmica (DCA)

Nesse caso, diferente do anterior, todos os canais são disponíveis para todas as células (23). Quando uma nova chamada solicita uma conexão em uma célula, um canal, dentre um conjunto de canais disponíveis, será alocado, desde que ele satisfaça uma dada restrição. Como, normalmente, mais de um canal satisfaz tal condição, um critério de otimização deve ser empregado para selecionar o melhor canal dentre os eleitos.

A função objetivo a ser otimizada pode ser a probabilidade de bloqueio local, global, a ocupação do canal, etc. Assim, os esquemas DCA diferem entre si pela escolha dessa função (23).

Devido às variações temporais e espaciais, sob uma carga de tráfego baixa e média, o esquema de alocação dinâmico utiliza os canais de rádio mais eficientemente.

Por outro lado, quando o tráfego é intenso, obtém-se a máxima eficiência espectral por meio da máxima eficiência espacial, ou seja, quando os canais forem atribuídos na mínima

distância de reuso possível. Nesse caso, a FCA prove melhores resultados (23).

1.3.1.3 Alocação de canal híbrida (HCA)

A alocação de canal híbrida é uma combinação de ambas descritas anteriormente (23)(24). Nesse caso, um conjunto de canais será alocado permanentemente para uma dada célula, enquanto que, uma outra parte, será compartilhada entre as células.

Quando uma chamada solicita uma conexão em uma célula, e todos seus canais nominais estão ocupados, um canal pertencente ao conjunto dos dinâmicos, será alocado para a chamada. Assim, o bloqueio somente ocorrerá quando todos os recursos de rádio fixo e dinâmico estiverem ocupados.

1.3.2 Gerência de tráfego

Gerência de tráfego constitui um conjunto de políticas e mecanismos que possibilita uma rede atender eficientemente os diversos tipos de tráfego durante os momentos de congestionamento devido à sobrecarga.

Nesse conjunto estão os procedimentos que evitam e previnem o congestionamento e os que detectam o congestionamento e restauram o equilíbrio na rede (25). Fazem parte do primeiro grupo, o controle de admissão de chamadas (CAC) e o agendamento, enquanto que, o controle de fluxo, adotado pelo TCP, faz parte do segundo.

O congestionamento é classificado como um problema de "compartilhamento de recursos"(25). Em redes de 1G e 2G os canais de rádio eram compartilhados entre as novas chamadas de voz e os *hand off* de chamadas de voz. A partir da 2.5G, com a introdução do tráfego de dados, os recursos de rádio passaram ainda a ser divididos com esse novo serviço.

A chegada das 3G e 4G desponta, a *priori*, como solução para esse problema devido ao aumento na capacidade da rede. Porém, a capacidade banda larga estimulará o surgimento de novos serviços. Esses, principalmente multimídia, demandarão uma grande quantidade de recursos. Conseqüentemente, essas redes experimentarão os mesmos problemas encontrados nas gerações anteriores, contudo, agravados pela característica auto similar encontrada no tráfego de dados e nas aplicações multimídia em tempo real.

Embora esse tema esteja fora do escopo deste trabalho, é importante ressaltar que a auto-similaridade impacta diretamente no dimensionamento do sistema causando um aumento no atraso médio e na perda de pacote. O que clama por um aumento da capacidade de armazenamento do *buffer* (26).

Nesse cenário, medidas de detecção e restauração tornam-se ineficientes tendo sua atuação demorada no restabelecimento do equilíbrio na rede. Além disso, pequenas variações na carga de tráfego causam uma grande degradação da QoS.

Particularmente, nesse caso, é mais apropriado prevenir e evitar o congestionamento. Além disso, mesmo em situações onde ele é inevitável, seu efeito pode ser minimizado de forma que se possa usar uma técnica para detecção e restauração mais eficientemente. Por esse motivo, neste trabalho, adota-se como mecanismo de gerência de tráfego os procedimentos que evitam e previnem o congestionamento.

Como mencionado anteriormente, nesse conjunto estão o CAC e o agendamento. A diferença entre eles reside no fato de que a primeira aloca recursos restringindo o acesso à rede de serviços de forma que um determinado nível de QoS seja satisfeito, e a segunda, aloca os recursos para os serviços (fontes de tráfego) que já se encontram na rede (27).

Neste trabalho, o procedimento escolhido para a alocação de recursos é o CAC. De uma forma geral, ele possibilita um uso eficiente dos recursos da rede enquanto satisfaz os requerimentos de QoS. Sua função é impor limites as chamadas, aceitando ou rejeitando novas solicitações de serviço. Essa decisão pode ser tomada baseada em (25)(28):

1. Modelos: nesse caso, modelos matemáticos são usados para estimar a situação da rede a *priori*. Assim, quanto mais elaborado o modelo, melhor será a estimação, e, via de regra, mais difícil a sua aplicabilidade.

2. Medidas: nesse caso, a condição da rede é observada e uma decisão é tomada. Desde que observar a condição da rede é uma função trivial, o emprego dessa categoria de CAC é mais difundido. A medida observada pode ser a carga da rede em um dado momento ou uma medida de QoS.

Neste trabalho considera-se que o CAC decidirá a admissão ou a rejeição de um cliente baseado na ocupação dos recursos de rádio.

Tradicionalmente, quando se projeta uma rede multiserviço, o CAC assume uma das três políticas (29)(30)(31) (32) [4]:

- Acesso total ou sem fronteira (*complete sharing, no boundary*): nessa política, os canais de rádio são completamente compartilhados entre todas as classes de serviço que acessam à rede. Para garantir a QoS de um serviço em detrimento a outro, normalmente é

[4] É importante mencionar que existem várias nomenclaturas associadas com as políticas de admissão. As apresentadas neste trabalho são as comumente encontradas na literatura.

aplicada uma prioridade preemptiva. Assim, esse serviço se comportará de acordo com a fila M/M/c/c, onde c é o número total de canais. Para mitigar o efeito dessa prioridade, os serviços menos privilegiados são acomodados em filas de espera. A Fig.(1.6.a) mostra como essa política de acesso é realizada para um sistema no qual os canais de rádio são totalmente compartilhados entre duas classes de serviço com taxas de chegada de Λ_1 e Λ_2.

- Acesso compartilhado ou fronteira fixa (*complete partition, fixed boundary*): nessa política, os canais de rádio são divididos entre todas as classes de serviço que acessam à rede. Nesse caso, cada classe de serviço pode ser modelada como uma fila M/M/c1/c1 e M/M/c2/c2, onde c1 e c2 são o número de canais dedicados para cada partição. Dentre as políticas de compartilhamento do espectro de rádio, essa, é a menos eficiente, pois, a partição fixa de recursos pode ocasionar em um ambiente onde há uma grande variação na carga de tráfego, ociosidade nos canais de rádio pertencentes a uma partição e sobrecarga dos canais da outra. A Fig.(1.6.b) mostra como essa política de acesso é realizada para um sistema no qual os canais de rádio são divididos entre duas classes de serviço com taxas de chegada de Λ_1 e Λ_2. Dependendo do número de canais dedicados e dos valores das taxas de chegada, os canais alocados em cada partição podem ser subutilizados ou sobre utilizados.

- Acesso restrito ou parcial, fronteira móvel (*Restrict or Partial Sharing, moveable boundary*): essa política é uma combinação das duas anteriores, pois, uma parte dos canais de rádio é completamente compartilhada, enquanto que, a outra é dedicada à uma classe de serviço. Na partição compartilhada, há, normalmente, a prioridade de uma classe em relação à outra. A divisão entre os canais é realizada pelo uso de um *threshold*. Em ambientes multiserviços, múltiplos *thresholds* podem ser usados para discriminar as várias classes de serviço. A Fig.(1.6.c) mostra como essa política de acesso é realizada para um sistema no qual os canais de rádio são alocados entre duas classes de serviço com taxas de chegada de Λ_1 e Λ_2. Nesse caso, na partição compartilhada, a classe de serviço 1 possui prioridade preemptiva sobre a 2. Essa, por sua vez, utiliza os canais dedicados mais os canais ociosos não usados pela classe 1 na partição compartilhada. Assim, diz-se que a alocação de canal para a classe 2 ocorre "sob demanda". Se esses canais não forem suficientes para garantir a qualidade de serviço da classe 2 e, não for possível aumentar o número de canais totais, pode-se utilizar um *buffer* para acomodar as novas solicitações de serviço dessa classe, Fig.(1.6.d).

O emprego do *buffer* não está limitado somente a uma classe de serviço. De fato, cada classe de serviço pode ser acomodada em um de forma a aumentar a QoS prestada. Outra

possibilidade, é o uso de um único *buffer* para todas as classes de serviço, sendo que, pode-se utilizar múltiplos *thresholds* para discriminá-las.

Recentemente, algumas variações ou extrapolações das políticas acima descritas têm sido estudas na literatura, principalmente quando o número de classes de serviço é maior que dois. Por exemplo, a Fig.(1.6.e) mostra uma topologia recentemente proposta por Bin Li *et al.* para integração de voz e dados em redes móveis celulares chamada de DTBR (*Dual threshold bandwidth reservation*) (32). Nele existem três classes de serviço: classe 1 (*hand off*), classe 2 (novas chamadas de voz) e classe 3 (dados). Utilizando dois *thresholds*, K_1 e K_2 ($K_1 > K_2$), ele segrega a região de acesso da seguinte forma: quando o número de canais ocupados é menor que K_2, todas as classes de serviço podem acessar a rede; caso contrário, somente as classe 1 e 2 têm acesso ao sistema; e finalmente, quando o número de chamadas é maior que K_1 somente é permitido o acesso da classe 1. Esse esquema é montado sobre a política de acesso total.

Desde que as políticas de controle discutidas tem como cerne as classes de serviço, e suas políticas de admissão são baseadas no número de clientes na rede e nos instante de chegada e término de uma nova solicitação por serviço, elas podem ser modeladas como cadeias de Markov a tempo contínuo.

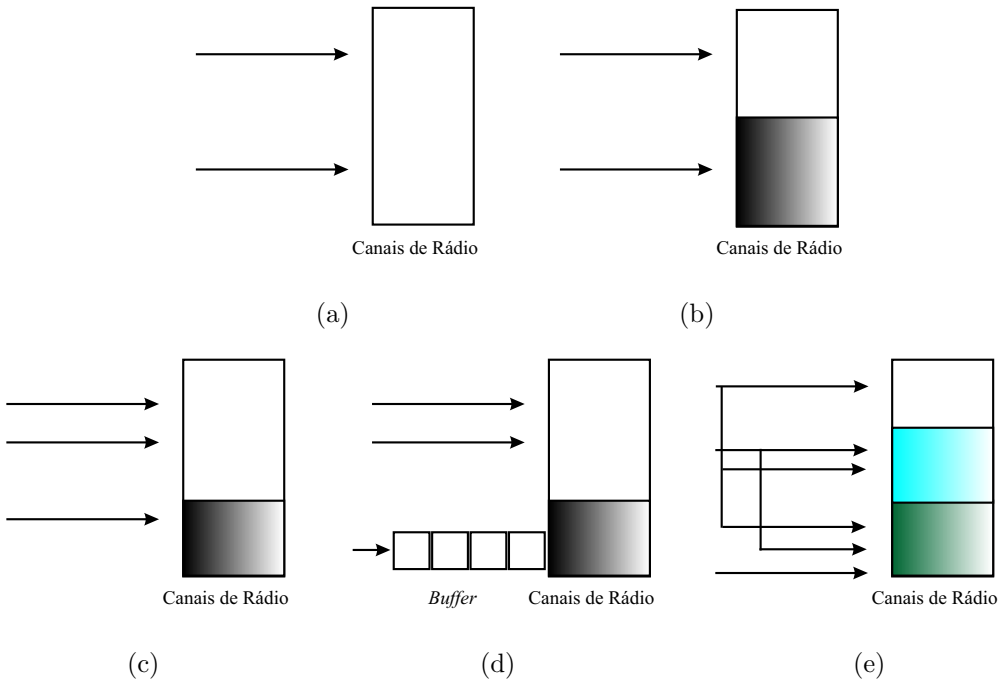


Figura 1.6: Políticas de acesso: (a) Acesso total, (b) Acesso compartilhado, (c) Acesso restrito, (d) Acesso restrito com *buffer*, (e) DTBR.

O crescente aumento na demanda por serviços multimídia em redes móveis celu-

lares tem tornado a provisão de garantias de QoS uma tarefa cada vez mais importante. Nesse contexto o CAC, como mencionado, exerce um papel fundamental controlando o acesso das chamadas. Todavia, problemas como mobilidade do usuário, escassez dos recursos de rádio e variação do canal de rádio motivaram, recentemente, o desenvolvimento de aplicações multimídia onde a largura de banda de uma aplicação pode ser dinamicamente ajustada, adaptando-se as condições da rede (28)(32)(33)(34). Exemplos de aplicações desse tipo são os serviços audiovisuais com codificação MPEG-2, MPEG-4 e H.263+ (28)(35).

Nesse sentido, os esquemas de alocação de recursos em redes de 3G e 4G devem ser projetados de forma a suportar de maneira eficiente o tráfego de aplicações multimídia adaptativas.

De uma forma geral, o esquema de alocação de recurso deve ser concebido de maneira a perceber o congestionamento e atuar de forma a ajustar a largura de banda destinada à aplicação multimídia em tempo real adaptando-a as condições da rede.

Ao atuar juntamente ao CAC na composição de um mecanismo de alocação de recursos, o mecanismo de adaptação de largura de banda (AB) torna-se uma ferramenta poderosa na redução da probabilidade de bloqueio de chamadas multimídia em tempo real, e, conseqüentemente, no melhoramento da QoS da rede.

Em ambientes não adaptativos, quando a disponibilidade de largura de banda da rede é menor que a exigida pelo serviço, o serviço é bloqueado, mesmo que existam recursos para prover o acesso com uma menor qualidade, até que haja disponibilidade de recursos para o aumento da banda destinada a essa aplicação.

1.3.3 Modelos de tráfego

1.3.3.1 Processos de chegada

O modelo de fonte de tráfego tradicionalmente usado na análise de fila é o processo de Poisson, o qual representa o número de ocorrência de um evento em um intervalo de tempo. Em uma rede móvel celular, ele é normalmente empregado na caracterização dos eventos de chegada de novas chamadas de voz e *hand off* em ambientes planares ou multicamadas, sessões de dados, fluxo de pacotes IP, etc.

Em um processo Poissoniano, a chegada de um cliente ocorre de forma "suave". Assim, ele não é capaz de modelar um padrão de chegada em grupo, tradicionalmente conhecido como **rajada**.

Para entender melhor o efeito da rajada no desempenho do sistema, cita-se o tra-

balho de Herman e Koen quando compararam o dimensionamento de um *buffer* usando uma fonte comportada, como uma Poissoniana e uma com rajada. O resultado, chamado de "assinatura"da rajada, indica que a utilização de um processo como o de Poisson subdimensiona o *buffer*.

Uma forma de representar a rajada é a utilização de uma fonte *ON-OFF*. Como uma fonte chaveada, a informação é enviada em uma sucessão de períodos ativos, chamados *ON*, separados por períodos de inativos, *OFF*. O tempo de duração dos períodos *ON-OFF* são variáveis aleatórias distribuídas exponencialmente. Esse comportamento pode ser representado por um processo de Poisson Interrompido (IPP) o qual é um caso particular do processo de Poisson modulado por uma cadeia de Markov (MMPP), pois a geração das chamadas somente ocorre em um dos estados.

O MMPP é descrito como um processo onde existe a geração de chamadas durante todos os estados do sistema, porém, com taxas diferentes. O caso mais comum, é aquele com dois estados, Fig.(1.7). Cada estado corresponde a diferentes taxas de chegada ou geração de chamadas (μ_1 para o estado 1 e μ_2 para o 2); com diferentes tempos de permanência (σ_1 para o estado 1 e σ_2 para o 2). Esse MMPP é caracterizado por um gerador infinitesimal e uma matriz de taxas, dados pelas Eq.(1.1) e Eq.(1.2), respectivamente. As principais características do MMPP são (36):

1. A capacidade de capturar taxas de chegada variantes no tempo e correlações entre os tempos entre chegada de chamada;
2. A superposição e a separação de um processo MMPP resultam em processos MMPP;
3. Um processo MMPP de $M+1$ estados pode ser obtido através da superposição de M processo IPP idênticos.

$$Q_s = \begin{bmatrix} -\sigma_1 & \sigma_1 \\ \sigma_2 & -\sigma_2 \end{bmatrix} \tag{1.1}$$

$$\Lambda_s = \begin{bmatrix} \mu_1 & 0 \\ 0 & \mu_2 \end{bmatrix} \tag{1.2}$$

Outros processos de chegada Markovianos são ainda definidos na literatura, são eles: MAP (*Markovian Arrival Process*) e sua versão discreta DMAP; o caso geral do MAP incluindo chegada em grupos BMAP (*Batch* MAP) e sua versão discreta D-BMAP. Particularmente, o emprego dos modelos BMAP e D-BMAP têm sido visto como uma excelente

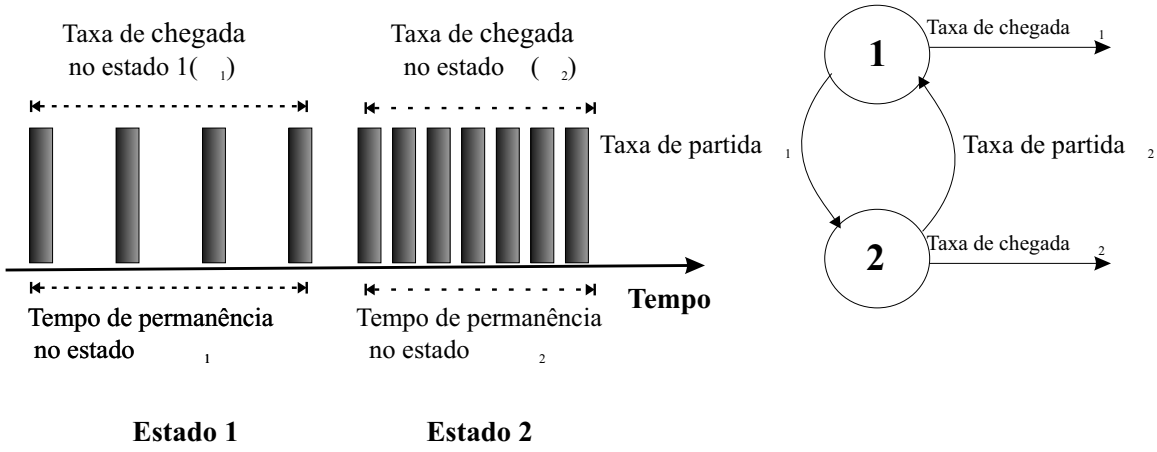


Figura 1.7: Processo MMPP de dois estados

opção para a caracterização do tráfego IP (37)(38) principalmente no que tange a captura do comportamento auto similar, uma vez que, a solução final é matematicamente tratável. Contudo, neste trabalho, são usados apenas os modelos de chegada Poisson e MMPP.

1.3.3.2 Tempos de serviço

Em uma rede móvel celular o tempo de retenção do canal (T_h) é função das variáveis tempo de duração da chamada (T_d) e tempo de residência da estação móvel na célula (T_s). Sendo que a última depende de fatores como tamanho da célula, velocidade, direção no qual o móvel caminha. Por definição, o tempo de retenção do canal é o tempo transcorrido entre os instantes no qual um canal é atribuído para servir uma chamada em uma célula e aquele no qual o canal é liberado pelo completamento da chamada ou por cruzar a fronteira da célula. Dessa forma, se a chamada é terminada na célula onde ele originou o serviço, então (39):

$$T_h = T_d \tag{1.3}$$

caso contrario, a estação móvel sai da célula antes do término da chamada, isto é, ($T_s < T_d$). Assim,

$$T_h = T_s. \tag{1.4}$$

Assim,

$$T_h = min(T_s, T_d). \tag{1.5}$$

Para se obter um modelo matematicamente tratável, considera-se que os tempos de duração da chamada e de residência da estação móvel na célula seguem distribuições exponenciais. Assim, o tempo de retenção do canal também é uma variável aleatória distribuída

exponencialmente. Dessa forma, a partir da Eq.(1.5), tem-se:

$$\begin{aligned} P\{T_h \leq t\} &= P\{T_s \leq t, T_d \leq t\} && (1.6)\\ &= 1 - P\{T_s > t, T_d > t\} && (1.7)\\ &= 1 - e^{-(\mu_s+\mu_d)t} && (1.8)\\ & && (1.9) \end{aligned}$$

A função densidade de probabilidade do tempo de retenção de canal é:

$$f_{T_h}(t) = (\mu_s + \mu_d)e^{-(\mu_s+\mu_d)t} \quad (1.10)$$

onde $\mu_s = 1/T_s$ e $\mu_d = 1/T_d$.

Neste trabalho, de acordo com a maior parte da literatura, todas as políticas de admissão apresentadas consideram que o tempo de retenção do canal é uma variável aleatória distribuída exponencialmente. Porém, deve ser mencionado que outros estudos indicam que o desempenho do sistema é degradado quando essa variável não é exponencialmente distribuída (40).

1.3.4 Medidas de Desempenho

As medidas de desempenho normalmente usadas na avaliação da alocação de recursos e do controle de admissão são parâmetros de QoS. Assim, dependendo da quantidade de classes de serviço e do tratamento dado a cada uma, o número de medidas de um sistema pode variar. Porém, de uma forma geral, elas são:

- Probabilidade de bloqueio: É a probabilidade de uma classe de serviço ser bloqueada devido à indisponibilidade de recursos.
- Probabilidade de preempção: É a probabilidade de uma classe de serviço sofrer preempção da classe com maior prioridade.
- Probabilidade de descarte[5]: É a probabilidade da classe com menor prioridade que sofreu a preempção ser bloqueado devido à indisponibilidade de espaço no *buffer*.
- Tempo médio de espera por serviço ou atraso médio: É o tempo médio que uma classe de serviço aguarda no *buffer* até que um canal seja disponibilizado.
- Vazão ou *throughput*: É a quantidade de chamadas de uma classe de serviço que são escoados pelo sistema por unidade de tempo.

[5]Essa probabilidade é algumas vezes chamada de probabilidade de bloqueio.

1.4 Revisão bibliográfica e trabalhos desenvolvidos

1.4.1 Modelagem e análise de desempenho de redes GSM/GPRS

Políticas de Controle de Admissão (CAC) têm sido amplamente estudas na literatura como uma ferramenta para se obter uma QoS satisfatória em redes multiserviço. Inicialmente, elas foram aplicadas para diferenciar a QoS de *hand off* e das novas chamadas de voz. Posteriormente, com a integração de voz e dados em redes móveis celulares, elas foram designadas para diferenciar a QoS entre essas classes de serviço.

A rede GSM/GPRS é alvo de várias pesquisas nessa área, uma vez que, ela constitui a tecnologia de 2.5G mais difundida no mundo. Em linhas gerais, as políticas normalmente adotadas pela grande maioria dos trabalhos são as de acesso total e restrito (29)(41)(42). Prioridade preemptiva entre os serviços de voz e dados também é usada (41)(42)(43), a exceção do trabalho apresentado na referência (44). O emprego de uma fila de espera para acomodar os pacotes GPRS também é normalmente considerado (42), a exceção, novamente da referência (44) onde as novas chamadas de voz e *hand off* são enfileiradas.

Através das considerações listadas acima, tem-se que para altos valores de tráfego de voz, devido à prioridade preemptiva, a quantidade de recurso destinada ao escoamento do tráfego GPRS é muito pequena aumentando a degradação da QoS de dados (45). Dessa forma, é imperativa a dedicação de canais para uso exclusivo do GPRS para que seja mantida uma prestação de serviço com a mínima qualidade de serviço (41). O número desses canais depende da atual demanda de tráfego na célula a ser dimensionada. Um efeito direto da reserva de recursos é o aumento na probabilidade de bloqueio de voz. Uma forma de reduzir esse bloqueio e aumentar a utilização da rede é o compartilhamento dos canais dedicados entre o serviço de dados e o *hand off* assim como feito em (46).

Um estudo interessante concernente à atribuição de prioridades é apresentado na referência (47). Nele uma prioridade dos serviços de voz sob dados é atribuída somente sob as chamadas GPRS *multislot*. Em outras palavras, quando uma chamada de voz chega no sistema e encontra todos os canais ocupados, um dos canais de uma chamada GPRS *multislot* é liberado para a prestação do serviço de voz. Essa solicitação de serviço somente será negada caso existam somente chamadas *singleslot*. O resultado mostrou que devido à sua flexibilidade, o emprego do esquema com liberação de canal reduz a probabilidade de bloqueio de voz e aumenta a utilização de canal. Contudo, essa melhora é obtida em detrimento do aumento da probabilidade de bloqueio e do tempo de transmissão do pacote GPRS.

Uma outra conclusão relevante está relacionada ao emprego do *buffer* para os pa-

cotes de GPRS. Os resultados mostram que o seu emprego para todos os tipos de pacotes GPRS, sejam novos ou que sofreram preempção, reduz consideravelmente a sua probabilidade de bloqueio. O emprego da fila de espera somente para os pacotes que sofrem preempção é adequado somente para aplicações em tempo real desde que o atraso seja relativamente pequeno (46).

A caracterização da fonte de tráfego de um pacote GPRS e seu impacto no sistema também foi bastante estudada. Os modelos de tráfego normalmente usados são o MMPP ou IPP (48)(41)(45). Nesse sentido, observou-se que o aumento na rajada da fonte de tráfego provoca um aumento no atraso médio na entrega dos pacotes (45).

Essas conclusões tornam-se pontos de partida para três estudos que são conduzidos no Capítulo 3 deste trabalho. Os modelos propostos utilizam as políticas de acesso total e restrito. As demais configurações são descritas nos parágrafos posteriores.

O primeiro modelo estudado considera que os canais de rádio são completamente compartilhados entre o serviço de voz GSM e os pacotes GPRS[6]. Uma prioridade preemptiva é atribuída ao serviço de voz sobre o serviço GPRS. Para minimizar o efeito dessa alta prioridade de voz, os pacotes GPRS são enfileirados no *buffer* enquanto aguardam por serviço.

O segundo modelo proposto é similar ao anterior, contudo, alguns canais de rádio (PDCH) são dedicados para o uso exclusivo do GPRS. Essa reserva de canais tem como objetivo garantir a QoS dos serviços GPRS para altas demandas de tráfego de GSM.

O último modelo considera um esquema de alocação de recursos onde as chamadas de voz são encaminhadas para o *buffer*, enquanto que, os pacotes são admitidos desde que exista a disponibilidade de recursos de rádio. Além disso, diferente dos anteriores, não há prioridade das chamadas de voz sobre os pacotes GPRS.

O objetivo desse capítulo está, entretanto, além dos modelos apresentados. Sua ênfase encontra-se principalmente nos aspectos que concernem a especificação comportamental do sistema, ou seja, na busca por um método que provenha uma representação clara e consistente, resultando em um modelo cuja especificação não possua dubiedades e que não gere contradições.

Para isso, propõe-se o uso de um método formal com características visuais para a especificação do sistema a ser modelado, e a partir dessa especificação, a geração de uma solução analítica usando uma cadeia de Markov a tempo contínuo, da qual se possa extrair as medidas de desempenho desejadas.

Métodos formais são métodos matemáticos usados no projeto e no desenvolvimento

[6]Neste trabalho se considera que todo o tráfego GSM é proveniente do serviço de voz.

de sistemas (49)-(53). Eles compreendem um amplo e diverso conjunto de técnicas como SDL (*Specification and Description Language, Z.100*), Estelle (*Extended Finite Machine Language*, ISO 9074), LOTOS (*Language of Temporal Ordering Specification*, ISO 8807), Redes de Petri, *Statecharts*, etc., os quais podem ser usados em diversos ramos da engenharia. Algumas aplicações clássicas de métodos formais são no projeto e no desenvolvimento de *softwares* e *hardwares* (51)(52). Particularmente em comunicações, o seu emprego foi impulsionado pelo desenvolvimento de sistemas abertos como o OSI/ISO, o qual são caracterizados por serviços e interfaces bem definidas (53)(54). A análise de desempenho de sistemas é um outro ramo de atuação de métodos formais. Redes de Petri (e suas variantes), e mais recentemente *Statecharts*, têm sido usados para avaliação de desempenho de sistemas (55)(56)(57).

No contexto deste trabalho, a aplicação do método formal é feita somente na especificação do comportamento do sistema, a qual, como será descrito posteriormente, se resume a primeira fase de um processo de modelagem. Porém, deve ser mencionado que o seu emprego se estende a qualquer fase do processo de desenvolvimento de um sistema como (49)(52)(53): na especificação (funcionalidade do sistema), na documentação (permitindo a interação entre as várias pessoas da equipe de desenvolvimento, etc.), na verificação (constatação se o modelo satisfaz sua especificação), e na validação (eliminação de erros).

Embora um método formal seja um método matemático, a principal dificuldade de sua aplicação não está propriamente na matemática, e sim nos aspectos pertinentes à modelagem (50). Dessa forma, uma especificação usando um método formal deve explicitamente responder, entre outras, às seguintes questões: quais são os elementos e eventos que norteiam o comportamento do sistema, e quais as inter-relações entre eles?

Os principais benefícios obtidos no emprego de um método formal na especificação de um sistema são a clareza e a consistência. Entende-se por clareza, como uma especificação que possui somente um significado, isto é, uma especificação sem ambigüidades. Uma especificação consistente é coerente no sentido de que nada contraditório pode ser derivado a partir da especificação (49)(50).

A característica visual do método formal a ser usado é fundamental para a garantia da intuição sobre o comportamento do sistema. Todavia, sem definições e regras claras, calcadas em bases matemáticas, essa intuição certamente conduzirá o analista a um entendimento ambíguo. Redes de Petri e *Statecharts* despontam na literatura como prováveis opções para resolver esse problema. A primeira, amplamente usada e consolidada na literatura, apresenta problemas na representação de atividades paralelas quando o sistema analisado é muito grande (58). Por outro lado, o *Statecharts* se apresenta como uma interessante ferramenta para especificação de sistemas complexos, pois, combina o formalismo matemático com a característica

visual.

A associação entre cadeia de Markov e o *Statecharts* para a geração de soluções analíticas e de simulações foi alvo de pesquisa em (55) e (56). Contudo, a aplicação dessa técnica foi direcionada a problemas de servidor de arquivo. Assim, cientes da necessidade de modelos matemáticos que permitam representações absolutamente confiáveis, tem-se com a primeira contribuição deste trabalho, o uso do *Statecharts* na modelagem da gerência de recursos de radio de uma rede móvel celular.

Os *templates* propostos, embora direcionados para a aplicação em uma rede GSM/GPRS, não são limitados a essa tecnologia, sendo assim, podem ser aplicados como CAC em outras tecnologias de rede sem fio ou cabeadas.

1.4.2 Modelagem e análise de desempenho de redes móveis hierárquicas

Como extensão dos estudos abordados anteriormente em relação ao CAC em redes móveis celulares, neste trabalho apresenta-se ainda modelos de alocação de recursos em redes móveis celulares hierárquica multicamada. Nesse ambiente, o CAC obedece aos mesmos preceitos seguidos em ambientes planares. Assim, diferentes políticas de alocação de recursos podem ser usadas nos diferentes níveis de hierarquia. Dessa forma, todos os mecanismos citados acima podem ainda ser empregados para melhorar a provisão da QoS e a utilização dos recursos de rádio.

A utilização dessa estrutura é uma alternativa viável para as operadoras de redes móveis celulares em face a crescente demanda de usuários por serviços de dados (2). Uma solução para a integração entre os serviços de voz e dados é o esquema chamado prioridade de voz no qual a microcélula escoa o tráfego de ambos serviços, enquanto que, a macrocélula atende somente o tráfego de voz que não é atendido nas microcélulas.

Recentemente Meo e Marsan propuseram um esquema de alocação de recursos em uma rede hierárquica GSM/GPRS chamado reserva dinâmica no qual um *threshold* foi usado para indicar o momento no qual se deve alocar dinamicamente um determinado número de canais para escoar o tráfego GPRS (59). Dessa forma, se a ocupação do *buffer* for superior ao *threshold* e o número de canais de rádio maior que uma dada fronteira, a quantidade de chamadas de voz que ultrapassou a fronteira é forçada a fazer um *handover* da microcélula para a macrocélula.

Essa alternativa é viável apenas em células onde a carga de tráfego GPRS se

mantém normalmente baixa, uma vez que, mediante ao aumento do tráfego haverá um grande número de *handovers* forçados, o que degradará consideravelmente o desempenho do serviço de voz e da macrocélula.

Neste trabalho é apresentada e analisada uma proposta diferente que consiste no roteamento das sessões de dados das microcélulas para a macrocélula. O critério usado para isso se baseia na ocupação do *buffer* e dos recursos de rádio.

Assim, se a ocupação do *buffer* for superior a um dado *threshold*, o tráfego GPRS é roteado para a macrocélula ao invés de forçar o *handover* da chamadas de voz. Como o tráfego IP resultante da fragmentação do conteúdo das sessões de dados utiliza a capacidade sob demanda da rede, não há impacto na macrocélula e QoS das chamadas de voz que é preponderante na rede e na microcélula.

O desempenho desse esquema de alocação de recursos é investigado usando o modelo proposto por Lindemann e Thummler com as seguintes diferenças: neste trabalho o *threshold* é usado para rotear sessões de dados da microcélula para a macrocélula, enquanto que, em (41) esse mecanismo é usado para estabelecer o controle de fluxo de pacotes IP. Além disso, eles não consideraram um ambiente celular multicamada hierárquico, ao passo que, neste trabalho essa consideração é feita.

Outra diferença existente entre os modelos propostos neste trabalho e em (59) e que: no modelo de Meo e Marsan a modelagem atua no nível de blocos de rádio, isto é, o *buffer* armazena os blocos de rádio resultantes da fragmentação dos pacotes IP. Neste trabalho, são armazenados os pacotes IP. Dessa forma, procura-se evitar usar aproximações para representação de uma distribuição de cauda pesada a qual caracteriza essa fragmentação. Além disso, Meo e Marsan consideram que a sessão já está no sistema, enquanto que, o modelo proposto em (41), que é usado neste trabalho, considera que a sessão de dados deve ainda ser admitida para depois gerar o tráfego IP. Contudo, vale ressaltar que, ambos utilizam o modelo de tráfego descrito pelo 3GPP.

O desempenho desse esquema de alocação de recursos foi estudado na análise de uma rede móvel celular hierárquica GSM/(E)GPRS. Os resultados mostram que o esquema proposto possui um desempenho superior ao do esquema de prioridade de voz, constituindo assim, uma boa alternativa para a integração de voz e dados em redes móveis celulares hierárquicas.

No cenário considerado, para experimentos, a rede móvel celular hierárquica modelada possui duas camadas, uma inferior contendo ψ microcélulas, e outra superior correspondente a macrocélula. Essa rede é considerada homogênea, ou seja, todas as células pertencentes a mesma camada são estatisticamente idênticas. Assim, no estado de equilíbrio, o comporta-

mento geral dessa camada pode ser analisado considerando apenas uma célula (2)(60). Além disso, é considerado um transbordo unidirecional, isto é, somente será atendido o transbordo do tráfego de voz das microcélulas para a macrocélula.

1.4.3 Controle de admissão de chamadas e mecanismos de adaptação de largura de banda

Por fim, seguindo a linha de pesquisa atualmente explorada na literatura, estuda-se, neste trabalho, a atuação do CAC juntamente ao mecanismo de alocação de largura de banda. O emprego desses procedimentos no esquema de alocação de canal traz novas perspectivas no dimensionamento do sistema, haja vista que, o mecanismo de adaptação de largura de banda permite a admissão de novos clientes durante o congestionamento por meio da redução da largura de banda das chamadas multimídia em serviço, provendo o acesso da nova chamada com uma qualidade de serviço mais baixa que a solicitada. Porém, ao detectar que a rede possui canais livres, a largura de banda da aplicação é promovida.

Ji-Hoon Lee *et al.* propuseram um esquema de alocação de recursos adaptativo que procura sempre atribuir uma largura de banda acima da média para cada chamada. Quando novos canais de rádios são disponíveis, recursos adicionais são distribuídos entre as chamadas (63).

Nas Ref.(28)(34)(61) (62) o CAC e a reserva de recursos são usados para prover garantias de QoS de forma a minimizar o bloqueio das chamadas de *hand off* e novas chamadas de voz. Em todos eles os experimentos são conduzidos via simulação.

Durante o seu serviço, uma chamada multimídia pode experimentar a redução (degradação) ou a promoção da largura de banda de acordo com a carga da rede. A taxa e a ordem de degradação da largura de banda de um serviço multimídia são estudados em (71)(33). Nesse sentido, um aspecto relevante aponta para a freqüência de comutação de largura de banda durante o tempo de vida de uma chamada multimídia. Através dos resultados, mostrou-se que a freqüência de comutação pode consumir mais recursos e ser pior que uma grande taxa de degradação (redução da largura de banda) (35). Assim, existe um compromisso entre os recursos da rede utilizados por chamadas e a sinalização devido à adaptação de largura de banda.

Em (32) são propostos modelos analíticos usando múltiplos *thresholds* para discriminar os serviços que acessam à rede. Nesses modelos apenas os serviços de dados adaptam sua largura de banda em função da carga da rede. Outra característica desse serviço é a utilização de chamadas elásticas. Essas chamadas representam a ação do controle de fluxo

realizado mediante situações de sobrecarga. No modelo considerado todas as chamadas multimídia requerem apenas um canal para o escoamento do serviço.

A otimalidade no mecanismo de alocação de canal também é alvo de pesquisas. Quando, a política de controle de admissão é estudada isoladamente, o objetivo é minimizar a probabilidade de bloqueio a longo prazo, maximizando o número de chamadas atendidas. Por outro lado, quando o CAC é analisado juntamente ao AB, o objetivo torna-se, além do anterior, controlar a freqüência de comutação (35), satisfazer os usuários (67).

A ferramenta constantemente usada na busca pela política de alocação de recursos ótima é o Processo Semi-Markoviano de Decisão (PSMD). Sua versão discreta, processo Markoviano de Decisão (PMD), não pode ser usado, nesse caso, pois, os tempos entre as épocas de decisão são aleatórios (68)(69). Outros métodos são, contudo, usados como *simulating anneling* (70), programação linear inteira com variáveis 0 e 1 (67).

Em (64) os autores analisaram uma rede CDMA usando um PSMD. Nesse trabalho, a política de decisão atua diretamente no controle de admissão das chamadas de voz. Comparando o resultado da política ótima com aquela que sempre aceita as chamadas quando o sistema possui recursos disponíveis, observou-se que a regra encontrada melhorou o desempenho do sistema especialmente para uma carga de tráfego elevada. Os resultados mostraram ainda que, o comportamento do serviço de dados sofre pouca influência em relação à política ótima e a que sempre aceita. Na mesma linha encontra-se a referência (65) que modela uma rede CDMA como um PSMD com restrições na probabilidade de bloqueio e na relação sinal interferência.

Em (66) os autores também utilizam o PSMD na modelagem da rede. Nesse caso, porém, a solução é obtida via programação linear usando o método SIMPLEX. A vantagem dessa solução consiste no fato de se poder adicionar uma restrição ao modelo, que nesse caso foi a de um nível máximo de probabilidade de bloqueio de *hand off*. Os autores buscaram maximizar os lucros da operadora satisfazendo os parâmetros de QoS.

Em (67) os autores modelaram a alocação de recursos como um problema de programação linear inteira com variáveis 0 e 1. A solução empregada foi o procedimento de relaxação Lagrangeana. Os autores buscaram maximizar o grau de satisfação do usuário aumentando a largura de banda de sua chamada quando necessário.

Fey Yu *et al.* estudaram o problema do CAC juntamente ao AB através de um PSMD (35). Contudo, a solução empregada foi o reforço de aprendizado, usando uma Rede Neural. Dentre os aspectos interessantes do modelo estão: o conhecimento do estado das células vizinhas na admissão de recursos da célula local e o controle de freqüência de comutação da largura de banda. O modelo proposto obteve bons resultados quando comparado a outros.

Nessa direção, são propostos neste trabalho as seguintes políticas de alocação de recursos adaptativas:

- Com Adaptação (CA): Neste esquema, a largura de banda será negociada entre o usuário (aplicação) e a rede durante a etapa de conexão. Se existirem recursos de rádio disponíveis, o serviço será acomodado com a largura de banda máxima, se não, admitir-se-á com a largura de banda mínima. Se a rede estiver sobrecarregada e não existirem recursos de rádio suficientes para acomodá-lo com largura de banda mínima, ele será bloqueado. Considera-se que, uma vez atendido os requerimentos de largura de banda, todos os demais parâmetros de QoS são satisfeitos. Quando uma chamada com largura de banda máxima deixa a rede, uma, com menor largura de banda, é promovida de forma a aumentar a satisfação do cliente e a utilização dos recursos de rádio.

- Adaptação Justa (AJ): Nesse esquema, todos os serviços serão aceitos se possível com largura de banda máxima. Caso contrário, a largura de banda de todas as chamadas com banda máxima em serviço será reduzida para admitir um novo cliente com largura de banda mínima. Quando um cliente com essa largura de banda deixa o sistema, o CAC verifica se é possível promover o perfil de QoS de todos os cliente da rede. Esse esquema de alocação de recursos é denominado de justo, pois, a taxa de transmissão de todos os serviços em tempo real são reduzidas e elevadas igualmente de forma a beneficiar um novo serviço.

- Adaptação ótima: este esquema é modelado como um PSMD. O objetivo é minimizar a probabilidade de bloqueio, freqüência de adaptação e a insastifação do usuário a longo prazo.

Para efeito de comparação de desempenho também é apresentado um modelo de alocação de recursos não adaptativo onde a chamada somente será aceita somente se existirem recursos de rádio suficiente para acomodá-la com a largura de banda máxima, bw_{max}.

Outros trabalhos relacionados à alocação de recursos que refletem aspectos de implementação das idéias discutidas anteriormente são descritos a seguir.

Serviços diferenciados (*Diffserv*) despontam com uma das melhores soluções para a provisão de QoS em redes móveis celulares de próxima geração (72)(73)(74)(75). Nele, o *Bandwidth Broker* (BB) é responsável pela alocação de recursos para um serviço baseado nos termos especificados no contrato de serviço (*Service Level Agreement*- SLA). Tais especificações podem, por exemplo, representar a faixa de largura de banda permitida e a prioridade (ouro, prata e bronze) para cada classe de serviço. Nesse contexto, o CAC simplesmente verifica a

disponibilidade de recursos na rede, e aceita ou rejeita o acesso baseado nos termos acertados. Os modelos propostos neste trabalho podem ser perfeitamente imersos em um esquema de alocação de recursos usando *Diffserv*. Para isso, as especificações de largura de banda acertadas no SLA devem refletir as possíveis taxas alcançadas por cada serviço, isto é, os níveis mínimo e máximo permitidos de redução e promoção de largura de banda de forma a manter uma QoS satisfatória, e que cada classe de QoS definida pelo UMTS seja mapeada nas classes *Diffserv*, por exemplo, as Classes Conversacional e *Streaming* podem representar a classe ouro, a Interativa, prata, e a classe de Fundo, bronze.

Outra aproximação bastante interessante é a utilização de um *framework* de QoS baseado em classes de serviço. Nesse *framework*, a Operadora de Serviço define suas próprias classes de serviço, suas próprias características de QoS, com seus respectivos preços (76). Tal *framework* deve ser flexível de forma a incorporar os efeitos inerentes a rede móvel celular (mobilidade, variação do canal). Além disso, ele deve possuir uma baixa complexidade de implantação e um eficiente meio de comunicação entre o serviço e a operadora de forma a diminuir a carga de sinalização. Usando esse *framework*, o CAC pode alocar os recursos de rádio baseado na especificação de cada classe.

Assim, essas idéias podem ser inseridas no contexto deste trabalho de forma a viabilizar uma melhor negociação de QoS entre a rede e a aplicação.

Capítulo 2

Modelagem e Análise de Desempenho

O objetivo deste capítulo é inserir o conceito e a importância da modelagem com cunho em desempenho de sistemas e prover o mínimo fundamento matemático de forma que os modelos a serem apresentados nos Capítulos posteriores possam ser entendidos.

2.1 Preliminares

A análise ou avaliação de desempenho é um fator fundamental em qualquer estágio do ciclo de vida de um sistema. Durante a etapa de planejamento, ela pode ser usada para investigar o comportamento do sistema de forma a observar as relações existentes entre as variáveis ou verificar se ele corresponderá às exigências especificadas (57)(79). Após a implantação de um sistema, a análise de desempenho pode ser aplicada para encontrar possíveis gargalos e sugerir alternativas para a sua expansão.

De uma maneira geral, o objetivo da análise de desempenho é encontrar uma dada configuração sistêmica que forneça a melhor relação custo-benefício.

2.1.1 Processo de modelagem

Um modelo, dentro do contexto da análise de desempenho, é uma abstração do sistema a ser analisado, o qual inclui, em diferentes níveis de detalhes, dependendo da solução a ser empregada, os principais elementos e eventos condizentes ao comportamento do sistema, e que ao mesmo tempo são relevantes na sua análise.

Neste trabalho, os sistemas apresentados são classificados como sistemas complexos reativos, pois, a sua dinâmica reage continuamente aos efeitos de estímulos internos e externos

(55)(58). Além disso, o seu comportamento, normalmente dirigido a evento, não pode ser representado de uma forma trivial, uma vez que, ele é dirigido por eventos complexos e inter-relacionados. Exemplos de sistemas complexos reativos são: automóveis, interface humano-máquina, protocolos de redes de comunicações, sistemas computacionais, mísseis e aviões (55)(58).

De uma maneira simplificada, o processo de modelagem pode ser dividido em etapas independentes, contudo, harmônicas entre si (56), são elas: especificação, parametrização, solução e apresentação. Essas etapas serão descritas a seguir.

2.1.1.1 Especificação

Nessa etapa é criada uma representação condizente ao sistema real, na qual devem estar contidos os principais componentes e as relações existentes entre eles, que são relevantes à avaliação de desempenho. O método mais usado é o diagrama de transição de estado.

A escolha do diagrama de transição de estado é norteada pela correspondência direta existente entre essa técnica e a solução via cadeia de Markov. Além disso, sua característica visual facilita o entendimento do comportamento do sistema. Porém, quando o sistema é complexo, o que constitui a maior parte das aplicações reais, pequenas variações no seu comportamento causam grandes mudanças na representação do modelo. Assim, de uma forma consensual, um sistema complexo reativo não pode ser descrito por esse tipo de especificação (58).

Outro fator que dificulta a representação de um sistema complexo reativo por meio do diagrama de transição é o crescimento exponencial do espaço de estados (58). Entretanto, quando se trata de modelos simples ou para exemplificações de partes do modelo, o uso dessa técnica é recomendável.

Nessa etapa do processo de modelagem é aconselhável o uso de uma ferramenta de especificação formal, preferencialmente visual, como redes de Petri e *Statecharts*. A rede de Petri, apesar de amplamente difundida e usada na avaliação de desempenho, apresenta um grande problema na representação de atividades paralelas, o que piora à medida que o modelo cresce (58)(56). Por outro lado, uma especificação *Statecharts* pode representar de forma explicita o efeito do paralelismo e a hierarquia entre os estados de um sistema (58)(55).

2.1.1.2 Parametrização

A parametrização consiste dos valores de entrada do modelo apresentado como tempo entre chegada dos eventos, duração de cada evento, etc.

2.1.1.3 Solução

Existem três tipos de soluções empregados na análise de desempenho, medidas ou experimentação, analítica e simulação (79)(80). Vários critérios norteiam a escolha da solução a ser usada como: o estágio do ciclo de vida do sistema, custo, tempo disponível para o desenvolvimento do modelo, nível de detalhamento, etc.

De uma forma geral, medidas ou experimentação são usadas quando o sistema a ser modelado já existe, ou se possui um protótipo do sistema ou se deseja construir um. Um dos grandes problemas desse método é o custo envolvido, uma vez que, normalmente são necessários equipamentos (*software* e *hardware*) ou o desenvolvimentos de protótipos. Além disso, em um sistema crítico, o próprio ato de "coletar as medidas"pode interferir no seu funcionamento e, conseqüentemente, levar o projetista a conclusões erradas. Contudo, uma vez aplicado corretamente, esse método fornece os melhores e mais confiáveis resultados.

A solução analítica é divida em dois grandes grupos: a teoria de filas e a cadeia de Markov. A primeira normalmente fornece uma solução na forma fechada para o problema proposto. Um uso comum da teoria de filas em telefonia, fixa ou móvel, é o emprego das fórmulas de Erlang B e C provenientes das filas M/M/c/c e M/M/c, no dimensionamento de sistemas de perda e espera. Porém, a obtenção de soluções na forma fechada para sistemas complexos torna-se bastante difícil. Nesse caso, o emprego da cadeia de Markov é mais apropriado.

A cadeia de Markov é uma ferramenta poderosa e consolidada para a modelagem de sistema com fins em desempenho devido a fatores como: flexibilidade, baixo custo[1], relativa rapidez na obtenção dos resultados [2].

A confecção de um modelo Markoviano é fundamentada na propriedade do esquecimento encontrada nas distribuições geométricas e exponenciais para o caso da cadeia de Markov a tempo discreto e contínuo, respectivamente. Isto é, o tempo entre as ocorrências dos eventos que causam as mudanças de estado são, por exemplo para o caso a tempo continuo,

[1]Existem pacotes gratuitos na Internet como o *Probabilistic Symbolic Model Checker* (PRISM) para a resolução de sistemas Markovianos (81).

[2]Isso pode ser intepretado de duas formas: a primeira diz repeito à confecção do modelo, e a segunda refere-se ao tempo no qual o resultado é calculado. Nesse caso, esse tempo depende do porte do modelo e da infraestrutura computacional disponível.

exponenciais. Porém, muitos eventos possuem seus comportamentos regidos de maneira que não podem ser representados exponencialmente. Nesse caso, os resultados obtidos podem ser pouco acurados servindo apenas para mostrar as tendências no comportamento do sistema. Assim, a utilização dessas distribuições, nem sempre reflete a realidade. O que torna alguns modelos Markovianos sujeitos a discussões.

Como solução para esse problema, pode-se utilizar distribuições do tipo fase, que permitem aproximar tão bem quanto se queira uma distribuição. O preço a ser pago nesse caso é a complexidade do modelo e o tempo para a obtenção dos resultados.

O outro método de solução, também muito comum, é a simulação. Dentre os tipos citados, esse é mais flexível, pois, trata-se de um programa computacional. Outra vantagem da simulação é o nível de detalhamento que pode ser usado na caracterização do sistema. Vários pacotes, de caráter geral ou específico, gratuitos [3] ou não, estão disponíveis na literatura que usam a simulação como o método de solução.

A desvantagem desse método é o tempo incorrido na obtenção dos resultados que geralmente é muito elevado. O emprego da computação de alto desempenho ou processamento paralelo aparece como solução desse problema.

Uma boa prática, comumente estabelecida na literatura, é o uso de mais de uma técnica de solução para a validação dos resultados obtidos (79).

2.1.1.4 Apresentação dos resultados

A principal meta da análise de desempenho é ajudar engenheiros, analistas, gerentes projetistas, fornecedores ou compradores a tirar conclusões claras, de um conjunto de opções fornecidas, a respeito do sistema ou da configuração sob análise. Conseqüentemente, uma apresentação coerente dos resultados com textos explicativos, gráficos, tabelas, etc., é de extrema importância para a aceitação do modelo proposto (79).

2.2 Especificação *Statecharts*

Statecharts é uma extensão dos tradicionais diagramas de transição de estados, aos quais foram adicionadas características como hierarquia (profundidade), ortogonalidade (representação de atividades paralelas) e interdependência (mecanismos de comunicação). Os elementos que fazem parte da especificação *Statecharts* são: estados, eventos, transições, rótulos

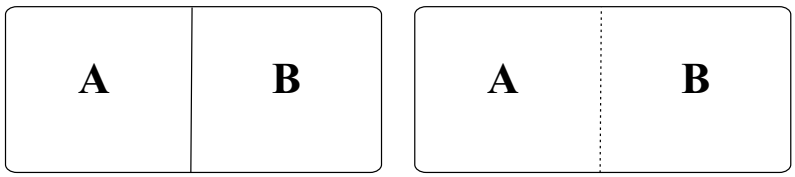
[3] Pacotes gratuitos de caráter geral SMPL (84), CSIM, e específico NS (83)

e expressões. A seguir é apresentada uma descrição de cada um desses elementos (55)(58).

Os estados são usados para descrever certas situações de um determinado sistema como *ON*, *OFF*, *PROCESSING*, *WAITING*, etc. Na sintaxe *Statecharts*, eles são representados por meio de retângulos boleados.

Há dois grupos de estados, os básicos e os não básicos (super estado). O primeiro não pode ser decomposto em subestados (estados filhos), enquanto que, o segundo pode. Dessa forma, expressa-se claramente a hierarquia entre os estados em uma especificação *Statecharts*.

Os estados não básicos podem ser decompostos de duas formas. A primeira é a decomposição do tipo XOR. Nesse caso, somente um dos subestados estará ativado em um determinado instante. Um estado não básico do tipo XOR é dividido usando linhas sólidas como mostrado na Fig.(2.1.a). A outra forma de decomposição estabelece, por outro lado, que todos os subestados do estado não básico estarão ativados ao mesmo tempo. Um estado não básico do tipo AND é dividido usando linhas pontilhadas como mostrado na Fig.(2.1.b).



(a) (b)

Figura 2.1: Decomposição de estados: (a) Tipo XOR, (b) Tipo AND.

Os eventos exercem um papel fundamental dentro do *Statecharts*, uma vez que, eles controlam a dinâmica do sistema provocando a alteração de estado. Opcionalmente, uma condição chamada de *condição de guarda* pode ser anexada ao evento. Nesse caso, mesmo que o evento aconteça, a mudança de estado somente ocorrerá se a condição de guarda for satisfeita.

Um evento chamado ação é usado na sintaxe do *Statecharts* para representar a comunicação entre os subestados de um estado não básico do tipo AND (estados ortogonais)[4]. Assim, quando um dado evento é disparado em algum nível do sistema, um outro evento associado ao primeiro evento (ação), também será instantaneamente disparado em um outro componente do sistema. A sintaxe *Statecharts* usada para representar um evento, condição de guarda e ação é dada por *ev[gc]/ac*.

[4]De uma forma contraditória, a comunicação entre estados ortogonais é chamada de paralelismo. Esse nome é atribuído a esse mecanismo, pois, o estado do superestado (estado pai) é dado como o produto ortogonal dos seus subestados.

Eventos imediatos são possíveis no *Statecharts*. Neles, a mudança de estado ocorre sempre que uma condição for satisfeita. Neste trabalho dois tipos de eventos são usados: *True(condição)*, *False(condição)*, os quais são abreviados para *Tr(condição)* e *Fs(condição)*.

As transições são as setas que representam fisicamente uma mudança de estado. Os rótulos são associados às transições para indicar os eventos que atuam na mudança desse estado. As expressões são a combinação de conectivos lógicos como *and*, *or*, *xor*, para formar eventos, ações, condições de guarda, etc.

Um aspecto importante na observação da dinâmica de um sistema é conhecimento do ponto de partida ou o estado inicial do sistema. O *Statecharts* oferece uma funcionalidade chamada de entrada *default* para representar o estado inicial de um sistema (55)(58). Esse tipo de entrada é representado por uma seta apontado para o estado como mostra a Fig.(2.2). O *Statecharts* ainda apresenta outros tipos de entrada como entrada por História, Condição e Seleção (58)(**?**), contudo, elas estão fora do contexto deste trabalho.

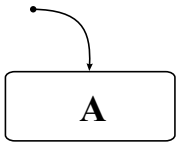


Figura 2.2: Entrada default

2.2.1 Extensão Estocástica do *Statecharts*

A especificação original do *Statecharts* não continha uma extensão estocástica impossibilitando-o de atuar como uma ferramenta de desempenho. Contudo, na referência (58), Harel, o criador do *Statecharts*, sugeriu a sua utilização como uma ferramenta de especificação para cadeias de Markov.

A implementação dessa proposta foi apresentada em (55) na qual foi fornecida uma técnica para geração automática da cadeia a partir da especificação de um sistema feita em *Statecharts*. Seguindo essa mesma linha, em (56) foi apresentada uma extensão do *Statecharts* para geração de uma solução via simulação. Esses dois trabalhos viabilizam através do *Statecharts* as duas formas clássicas de solução para um problema de análise de desempenho.

O caráter estocástico dessa extensão consiste, de uma forma simplificada, no uso de distribuições de probabilidade no acionamento dos eventos. Em uma extensão focada na análise de desempenho onde se norteia a geração de cadeias de Markov, esses eventos devem ser exponencialmente distribuídos.

2.2.2 Alguns aspectos de modelagem

A modelagem de um sistema através do *Statecharts* consiste em caracterizar o seu comportamento como uma composição de subsistemas os quais podem ser representados por subestados AND e XOR. Após a definição dos subsistemas e dos seus tipos, o comportamento de cada um deve ser descrito através da listagem de todas as possíveis transições (eventos e ações) e das mudanças de estados que elas podem causar.

Esse processo de modelagem pode então ser dividido em duas partes: a declaração e a funcional. Na primeira parte os elementos são colocados e distribuídos dentro da especificação de acordo com o seu comportamento, expressando concorrência ou hierarquia. A parte funcional concerne ao comportamento interno de cada subsistema e na sua interação com os outros subsistemas por meio dos eventos e das ações

2.3 Cadeia de Markov

2.3.1 Definição

Um processo estocástico $X_n, n = 1, 2, 3, ...$ é uma cadeia de Markov se para todos os estados $i_0, i_1, ..., i_n, j$, e $n \geq 0$,

$$P\{X_{n+1} = j | X_0 = i_0, X_1 = i_1, ..., X_n = i\} = P\{X_{n+1} = j | X_n = i\} \tag{2.1}$$

Se o termo do lado direito dessa equação é independente de n então o processo possui uma probabilidade de transição, $P\{X_{n+1} = j | X_n = i\} = p_{ij}$, que é estacionária ou homogênea no tempo (68)(77).

A Eq.(2.1) define o que é chamado de propriedade de Markov ou esquecimento (*memoryless*). Ela diz que para qualquer n, o futuro do processo $X_{n+1} = j$ é condicionalmente independente do passado $X_0 = i_0, X_1 = i_1, ..., X_{n-1} = i_{n-1}$, dado o presente $X_n = i$. O caso a tempo contínuo segue o mesmo princípio acima, porém, as probabilidades de transição são dadas por:

$$p_{ij}(t) = P\{X_{t+s} = j | X_s = i\}, \forall t \geq 0 \text{ e } s \geq 0, \tag{2.2}$$

onde, novamente, tem-se uma probabilidade de transição estacionária, pois, a Eq.(2.2) independe de s (78).

As probabilidades de transição da cadeia de Markov são normalmente disposta na forma matricial, **P**, chamada de **matriz de transição**, dada por

$$\mathbf{P} = \begin{bmatrix} p_{00} & p_{01} & \cdots \\ p_{10} & p_{11} & \cdots \\ \vdots & \vdots & \ddots \end{bmatrix}$$

Por definição

$$\begin{aligned} p_{ij} &\geq 0 \qquad \forall\ i, j \in E, \\ \sum_{j \in E} p_{ij} &= 1 \qquad \forall\ i \in E. \end{aligned}$$

onde E é o espaço de estados.

Para uma cadeia de Markov a tempo contínuo, tem-se também uma representação matricial chamada de **matriz de taxas ou gerador infinitesimal**. Contudo, diferente do caso discreto, as entradas dessa matriz são as taxas de transição entre os estados.

$$\mathbf{Q} = \begin{bmatrix} -q_{00} & q_{01} & \cdots & q_{0N} \\ q_{00} & -q_{11} & \cdots & q_{1N} \\ \vdots & \vdots & \ddots & \vdots \\ a_{N0} & q_{N1} & \cdots & -q_{NN} \end{bmatrix}$$

onde os termos da diagonal principal $q_{ii} = \sum_{j=0}^{N} q_{ij}, \forall i$ e $j,\ $ e $j \neq i$ representam a taxa total de saída do estado i.

2.3.2 Classificação dos estados

Sejam as variáveis aleatórias v_i e τ_i o número de visitas e o tempo do primeira visita ao estado i. Os estados de uma cadeia de Markov são classificados com base em $P_i\{\tau_i < \infty\}$ e $E_i\{\tau_i\}$[5] como mostrado na Tabela 2.1.

Um estado i é recorrente se ele for visitado infinitas vezes. Caso contrário, se a realização do processo visitar i um número finito de vezes, de modo que, após a última visita, o processo não entrar nesse estado nunca mais, ele será chamado transiente ou transitório, pois, ele desaparecerá após algum tempo. Note que i é recorrente se (77):

$$E\{v_i\} = \sum_{n=0}^{\infty} p^n(i/i) = \infty,$$

[5] $E\{\cdot\}$ é o operador esperança

Tabela 2.1: Classificação dos estados de uma cadeia de Markov.

	$P\{\tau_i < \infty\} < 1$	$P\{\tau_i < \infty\} = 1$
$E\{\tau_i\} < \infty$	-	Recorrente positivo
$E\{\tau_i\} = \infty$	Transiente	Recorrente nulo

e transiente se $E\{v_i\} < \infty$.

Diz-se que um estado $j \in E$ é acessível ou alcançado ($i \to j$) a partir de um estado $i \in E$ se $p_{ij}^n > 0$ para $n \geq 0$; caso contrário, j é inacessível a partir de i.

O estado j se comunica com i se ($i \to j$) e ($j \to i$). Um subconjunto de estados $C \in E$ é chamado fechado se nenhum estado fora dele ($\overline{C}$)[6] pode ser alcançado por qualquer outro estado dentro dele. Um estado que forma sozinho um conjunto fechado é chamado absorvente. Um conjunto fechado é chamado irredutível se nehum subconjunto próprio dele é fechado. Uma cadeia de Markov é chamada irredutível se seu único conjunto fechado é o conjunto E de todos os seus estados.

Um estado i possui período $d(i)$ se $d(i)$ é o maior divisor comum de $n \geq 1$ tal que $p_{ii}^n > 0$. Se $d(i) = 1$, então i é aperiódico, caso contrário, $d(i) > 1$, então o estado i é periódico.

É importante citar que, a periodicidade, recorrência ou transiencia são propriedades de classe. Isto é, em qualquer classe irredutível, todos os estados são recorrentes positivos ou nulos ou transitórios. No caso da periodicidade, em um conjunto com a mesma condição acima, todos os estados possuem o mesmo período (78).

Uma cadeia de Markov pode ser dividida em um conjunto de estados recorrentes (classes) $C_k, k = 1, 2, ..., m$ com m finito em E e possivelmente infinito quando este é contável. Dessa forma, E pode ser escrito como: $E = C_1 \cup C_2 \cup ... \cup C_m \cup T$, onde T é o conjunto de estados transientes (78). Assim, chama-se de **forma canonica** de P a matriz (77):

$$\mathbf{P} = \begin{bmatrix} P_1 & 0 & 0 & . & . & 0 \\ 0 & P_2 & 0 & . & . & 0 \\ . & & . & & & \\ . & & & . & & \\ 0 & & & & P_m & 0 \\ Q_1 & Q_2 & . & . & Q_m & Q_{m+1} \end{bmatrix},$$

[6] $\overline{X}$ é o complemento de X

onde P_i corresponde às transições entre os estados C_i; Q_i as transições a partir dos estados transientes em T para os estados em C_i; e Q_{m+1} as transições entre os estados em T.

Para um conjunto finito E, usa-se a expressão cadeia única, *unichain*, para as cadeias consistindo de um único conjunto fechado irredutível, e um, possivel vazio, conjunto de estados transientes. Caso contrário, emprega-se o termo de cadeias múltiplas, *multichain*. Essa definição será usada posteriormente para definir a solução do critério de otimalidade usado no processo Semi-Markoviano de Decisão.

2.3.3 Comportamento limite da cadeia de Markov

Se uma cadeia é irredutível, recorrente positiva e aperiódica [7], então exite a probabilidade limite, chamada de distribuição estacionária para a cadeia de Markov, tal que (78):

$$\lim_{n\to\infty} p_{ij}^n = \pi_j > 0, \forall j,$$

o qual é independente do estado inicial i, onde $\{\pi_j, j = 0, 1, 2, ...\}$ é a única solução para

$$\pi_j = \sum_{i=0}^{\infty} \pi_i p_{ij}, \forall j \geq 0 \qquad\qquad \sum_{j=0}^{\infty} \pi_j = 1. \tag{2.3}$$

No caso da cadeia de Markov a tempo contínuo, tem-se que o comportamento limite para uma cadeia finita com N estados é dada por (78):

$$\pi_j Q = 0, \qquad\qquad \sum_{j=0}^{N} \pi_j = 1, \tag{2.4}$$

2.4 Processo Markoviano de Decisão (PMD)

O PMD é um sistema dinâmico que evolui com uma lei de probabilidade de movimento controlada por decisões que são tomadas em pontos no tempo no qual o estado do sistema é observado. Como conseqüência direta dessa decisão, incorre-se em um custo ou uma recompensa e em uma mudança de estado.

Ao observar o estado em um dado momento, uma decisão é tomada. Esses "pontos" no tempo são chamados de instantes ou épocas de decisão. Baseado neles tem-se os processos Markoviano de Decisão a tempo discreto e o Semi-Markoviano de Decisão (PSMD). No primeiro, o sistema é observado e uma ação é tomada em instantes eqüidistantes ou discretos

[7] A cadeia com essa característica é geralmente chamada de *ergódica*.

de tempo $t = 1, 2, ...$ O último é uma generalização do processo Markoviano de decisão por permitir que o tempo até o próximo instante de decisão tenha uma distribuição de probabilidade arbitrária (82). Particularmente, neste trabalho, a distribuição de probabilidade entre os instantes de tomada de decisão são variáveis aleatórias distribuídas exponencialmente, o que remete o PSMD a um PMD a tempo contínuo.

O conjunto de épocas de decisão pode ser finito ou infinito. Quando o mesmo é finito o problema de decisão é chamado horizonte finito, caso contrário, infinito.

Ao tomar uma decisão, atua-se no sistema por meio de uma ação a de um conjunto de ações disponíveis para o estado observado. Dessa forma, a ação exerce um papel fundamental no PMD, pois, ela "dita o rumo" do sistema em cada instante de decisão.

Ao escolher uma ação, deve-se seguir alguma política ou regra que, a *priori*, pode apresentar qualquer qualquer comportamento. Nesse conjunto de políticas existe uma subclasse chamada de deterministica e estacionária que prescreve a mesma ação R_i sempre que o sistema está no estado i em uma época de decisão. Como o caráter Markoviano desse processo reside no fato de que o seu comportamento futuro é independente dos estados e ações passadas dado estado e a ação correntes (82), é intuitivo considerar somente as políticas deterministicas estacionárias (69)(77)

Em um horizonte de planejamento infinitamente longo, dois critérios de otimalidade podem ser usados: custo descontado e custo médio. O primeiro não se aplica neste trabalho visto que a estrutura de custo usada pernanece inalterada com o passar do tempo. Assim, o critério usado é o custo médio esperado por unidade de tempo. Outro fator que norteia a escolha desse critério é o fato de que ele é apropriado em sistemas no qual acontecem muitas transições em um curto espaço de tempo (69), que, particularmente, é o caso do modelo proposto.

2.4.1 O critério do custo médio para uma política estacionária

Para uma política estacionária R, o custo médio a longo prazo é dado por:

$$g_i(R) = \lim_{n\to\infty} \frac{1}{n} V_n(i, R)$$

onde $V_n(i, R)$ é o custo total esperado nas n primeiras épocas de decisão quando o estado i é visitado considerando a política R. De uma forma simplista, o custo médio é dado como á média aritimética do custo total no horizonte considerado.

No caso onde a cadeia de Markov X_n possui um único conjunto fechado de estados, caso *unichain*, sob a política R, o custo médio a longo prazo é independente do estado inicial

(69). Então:

$$g_i(R) = g(R)\ \forall i \tag{2.5}$$

Uma política estacionária é dita ótima se

$$g_i(R^*) \leq g_i(R) \tag{2.6}$$

A teoria dos PMD garante que a política estacionária R^* sempre existe, e que, ela é ótima entre todas as classes de políticas possíveis (69).

2.4.2 Algoritmo de Iteração de Valores (AIV)

A obtenção da política ótima pode ser feita tradicionalmente usando uma das três técnicas: algoritmo de iteração de políticas, valores e programação linear. Recentemente, uma outra aproximação surgiu usando o reforço de aprendizado.

Dentre os métodos tradicionais o algoritmo de iteração de valores (AIV) mostra-se mais efetivo na solução de sistemas Markovianos de decisão de grande porte. Pois, ele calcula o valor do custo total esperado recursivamente ao invês de resolver um sistema de equações lineares a cada iteração como é feito nos demais (69).

Para um PMD a tempo discreto, o algoritmo de iteração de valores computa recursivamente o valor[8]:

$$V_n(i) = \min_{a \in A(i)} \{c_i(a) + \sum_{j \in E} p_{ij}(a) V_{n-1}(j)\},\ i \in E,$$

iniciando de uma função arbitrariamente escolhida $V_0(i), \forall i \in E$.

A quantidade $V_n(i)$ pode ser interpretada como o custo esperado total mínimo com um horizonte de n períodos quando o estado corrente é i e um custo terminal $V_0(j)$ e acarretado no ao sistema quando este pára no estado j (69)(82).

Essa intepretação sugere que para um horizonte suficientemente grande a diferença em um passo $V_n(i) - V_{n-1}(i)$ estará muito próxima do custo médio médio mínimo por unidade de tempo, e quando $n \to \infty$, os limites:

$$m_n = \min\{V_n(i) - V_{n-1}\} \text{ e } \ M_n = \max\{V_n(i) - V_{n-1}\}$$

se aproximão da taxa de custo mínimo.

[8] A prova de converg6encia do algoritmo de iteração de valores é encontrada na referência (69).

Escolhendo $V_0(i)$ tal que $0 \leq V_n(i) \leq \min_a c_i(a), \forall i \in E$, tem-se que $V_1(i) \geq V_0(i) \forall i$, o que implica que cada termo da seqüência não decrescente $\{m_n, n \geq 1\}$ é não negativo. Assim,

$$\frac{M_n - m_n}{m_n} \leq \varepsilon \Rightarrow \frac{g(f^n) - g^*}{g^*} \leq \varepsilon$$

isto é, o custo $g(f^n)$ da política f^n obtido na n-ésima equação não pode diferir mais que a precisão desejada ε do suposto custo mínimo g^* quando $\frac{M_n - m_n}{m_n} \leq \varepsilon$. Abaixo tem-se o pseudo-código do algoritmo de iteração de valores:

1. $INICIO$
2. Para cada estado $i \in E$ e $a \in A(i)$escolher $0 \leq V_n(i) \leq \min_a c_i(a)$
3. Inicializar os valores m_n, M_n e ε
4. Para $n = 1, 2, ...$ fazer{
5. Para cada $i \in E$ fazer {
6. $V_n(i) = \min_{a \in A(i)}\{c_i(a) + \sum_{j \in E} p_{ij}(a) V_{n-1}(j)\}$
7. Determinar f^n como a política estacionária cujas ações minimizam o lado direito da equação acima.
8. }
9. Calcular os limites: $m_n = \min\{V_n(i) - V_{n-1}(i)\}$ e $M_n = \max\{V_n(i) - V_{n-1}(i)\}$
10. O algoritmo pára com a política f^n se $0 \leq \frac{M_n - m_n}{m_n} \leq \varepsilon$
11. }
12. FIM

2.5 Processo Semi-Markoviano de Decisão (PSMD)

Com mencionado anteriormente, o PMD é um processo que observa o estado do sistema em épocas fixas $t = 0, 1, 2, ...$ Porém, nos modelos de desempenho abordados neste trabalho, os tempos entre as ocorrências dos eventos, épocas de decisão, não são idênticos, mas aleatórios. Tais problemas são analisados considerando um processo Semi-Markoviano de Decisão (PSMD).

No PSMD, se em uma época de decisão uma ação a é escolhida em um estado $i \in E$ então o tempo, o estado e o custo até a próxima época de decisão dependem somente dos estado corrente e da ação escolhida.

Na confeção de um PSMD além dos elementos já introduzidos como o estado $i \in E$, ação $a \in A(i)$, custo $c_i(a)$ e probabilidade $p_{ij}(a)$ é importante também a definição do tempo esperado até a próxima época de decisão, $\tau_i(a)$, se a ação a é tomada no estado corrente i.

Considerando o custo médio por unidade de tempo a horizonte infinito como critério de otimalidade é possível utilizar o algoritmo de iteração de valores para obter a política ótima, desde que, aplica-se o método de uniformização (69).

Esse método converte um modelo de decisão Markoviano a tempo contínuo em ouro a tempo discreto tal que para cada política estacionária os custos médios por unidade de tempo são os mesmos em ambos modelos (82). A uniformização é feita da seguinte forma:

- Escolher um número $0 < \tau < \min_{i,a} \tau_i(a)$
- Considerar o PSMD cujo o espaço de estados, espaço de ações, probabilidade, custo e tempo entre as transições são dados respectivamente por: E, $A(i)$, $p_{ij}(a)$, $c_i(a)$ e $\tau_i(a)$
- Considerar o PMD cujo espaço de estados, espaço de ações, probabilidade, custo sejam dados por:

$$
\begin{aligned}
\overline{E} &= E, \\
\overline{A(i)} &= A(i), && i \in \overline{E} \\
\overline{c_i(a)} &= \frac{c_i(a)}{\tau_i(a)}, && i \in \overline{E} \text{ e } a \in \overline{A(i)} \\
\overline{p_{ij}(a)} &= \begin{cases} \frac{\tau}{\tau_i(a)} p_{ij}(a), & i \neq j, i \in \overline{E} \text{ e } a \in \overline{A(i)} \\ \frac{\tau}{\tau_i(a)} p_{ij}(a) + [1 - \frac{\tau}{\tau_i(a)}], & i = j, i \in \overline{E} \text{ e } a \in \overline{A(i)} \end{cases}
\end{aligned}
$$

Utiliza-se a pertubação τ para garantir que toda política ótima induza a uma cadeia de Markov aperiódica, condição necessária para garantir $\lim_{n\to\infty} m_n = \lim_{n\to\infty} M_n = g^*$, onde g^* é o custo médio mínimo por unidade de tempo (82).

Capítulo 3

Modelagem da Alocação de Recursos usando *Statecharts*

3.1 Preliminares

Neste capítulo serão apresentados e discutidos os resultados obtidos através da modelagem da alocação de recursos em uma rede GSM/GPRS. O enfoque abordado consiste na modelagem analítica dessa interface aérea, a qual, dentro das etapas do processo de modelagem apresentadas no capítulo anterior, destaca o uso de uma especificação de alto nível, para representar o compartilhamento dos canais de rádio. A partir dessa especificação é gerada uma cadeia de Markov de onde são derivadas as medidas de desempenho previamente apresentadas na seção 1.3.4.

Para validar o modelo analítico, *Statecharts*/Markov, os seus resultados são comparados com outros obtidos através de simulações feitas usando uma extensão da linguagem de programação C chamada SMPL (84). Esse simulador foi desenvolvido pelo Eng. Roberto Menezes Rodrigues e resultou na sua Dissertação de Mestrado dada na referência (94).

Os resultados apresentados neste capítulo foram publicados nas referências (86) e (89).

3.2 Modelagem da Interface Aérea da rede

Nesta seção são apresentados os modelos analíticos correspondentes a três esquemas de alocação de recursos. No primeiro, os canais de rádio são completamente compartilhados

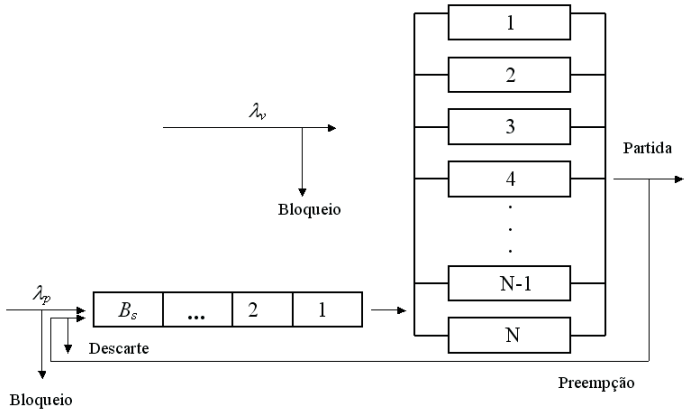


Figura 3.1: Sistema de fila do esquema de alocação de recursos 1

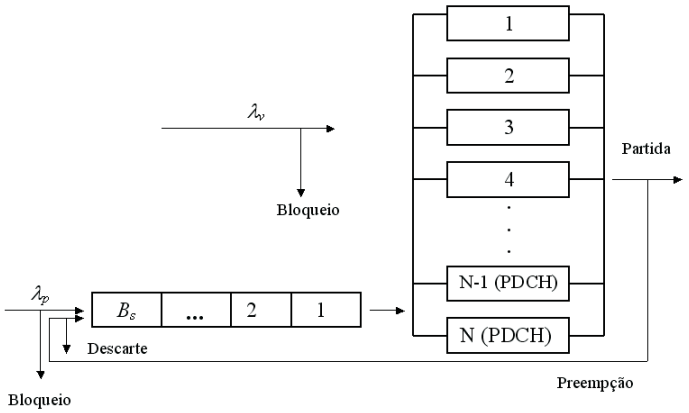


Figura 3.2: Sistema de fila do esquema de alocação de recursos 2

entre o serviço de voz e os pacotes GPRS. Uma prioridade preemptiva é atribuída para o serviço de voz sobre o GPRS. Para minimizar o efeito dessa alta prioridade, os pacotes de dados são enfileirados no *buffer* enquanto aguardam por serviço. O sistema de fila que representa esse esquema de alocação de recursos é mostrado na Fig.3.1. Nele estão indicados os elementos e eventos de interesse na análise de desempenho do sistema, isto é, os canais de rádio (servidores), o *buffer*, as chegada e partida dos serviços de voz e dados, os bloqueios de voz e pacotes GPRS, a preempção e o descarte dos pacotes GPRS.

O segundo esquema de alocação de recursos é similar ao primeiro, contudo, alguns canais de rádio (PDCH) são dedicados para o uso exclusivo do GPRS. Essa reserva de canais tem como objetivo garantir a QoS dos serviços GPRS para altas demandas de tráfego de GSM. O sistema de fila usado para representar o seu comportamento é mostrado na Fig.3.2. A única diferença entre os sistemas de fila das Fig.3.1 e Fig.3.2 são os PDCHs reservados.

No último esquema não há prioridade de voz sobre dados. Além disso, as chamadas

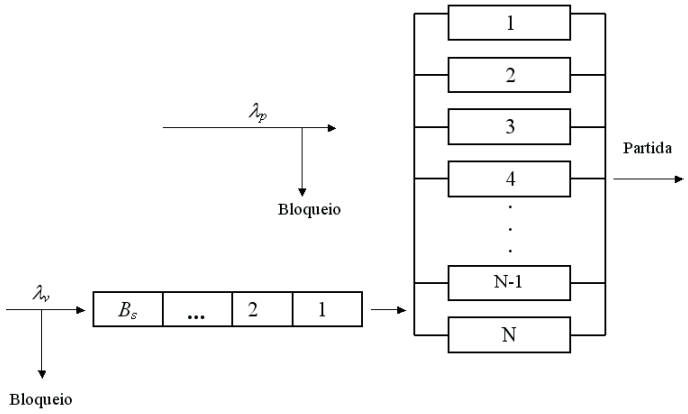


Figura 3.3: Sistema de fila do esquema de alocação de recursos 3

de voz são enfileiradas no *buffer* enquanto aguardam por serviço. A prioridade de voz sobre dados não é considerada nesse esquema, uma vez que, o tempo médio de serviço (tempo médio de transmissão) de um pacote GPRS é muito curto (44). Na Fig.3.3 é mostrado o sistema de fila correspondente à esse esquema. Diferente dos demais, a chamada de voz é encaminhada para o *buffer*, enquanto que, os pacotes GPRS são encaminhados para os canais de rádio.

Em todos os esquemas de alocação de recursos apresentados, as chegadas das chamadas de voz e dos pacotes GPRS são processos de Poisson independentes, enquanto que, os tempos de serviços de uma chamada de voz e de um pacote GPRS seguem distribuições exponenciais.

3.2.1 Esquema de alocação de recursos 1

Por fins didáticos, a modelagem analítica referente a esse esquema será apresentada passo a passo. Inicialmente será mostrada a especificação do comportamento sistêmico em *Statecharts*. Em seguida, um exemplo de pequeno porte será usado para ilustrar a associação entre a especificação *Statecharts* e a cadeia de Markov. Posteriormente, serão derivadas as medidas de desempenho. Para os demais esquemas serão enfatizadas somente as suas especificações e as suas medidas de desempenho.

A especificação do esquema de alocação de recursos 1 é mostrada na Fig.3.4. Ela é formada por um estado não básico do tipo AND, contendo cinco subestados, os quais são descritos a seguir:

- Source: Esse *template*[1] possui um único subestado chamado **Ready** Fig.(3.5.a) (55)(56).

[1]Neste trabalho, um *template* é um conjunto de estados e eventos que realizam uma determinada tarefa.

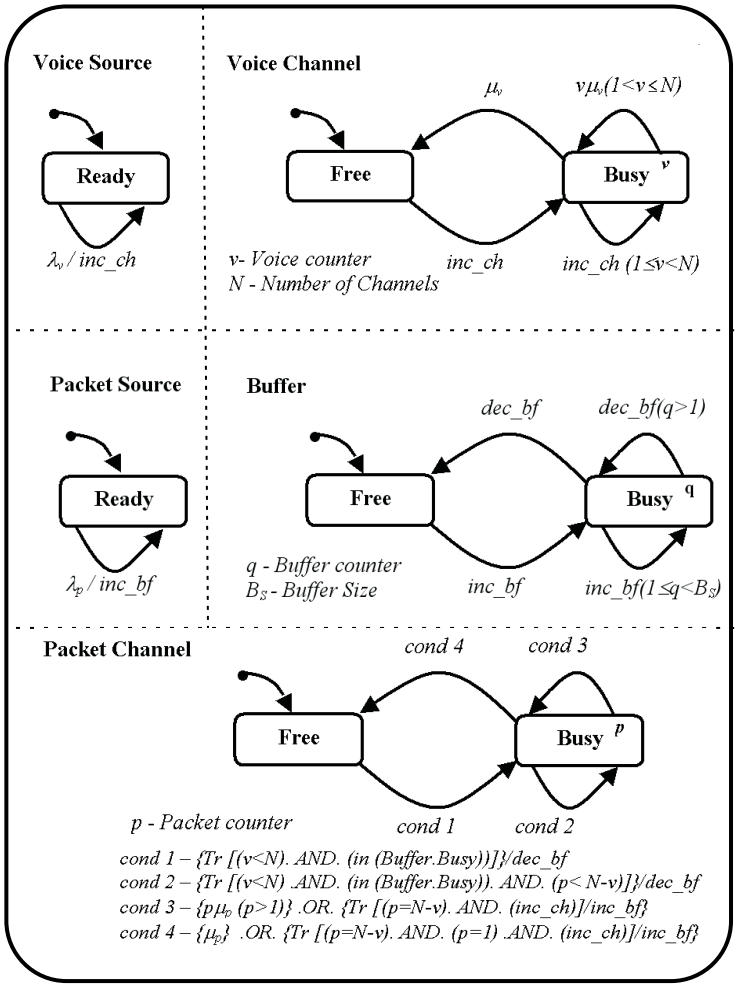
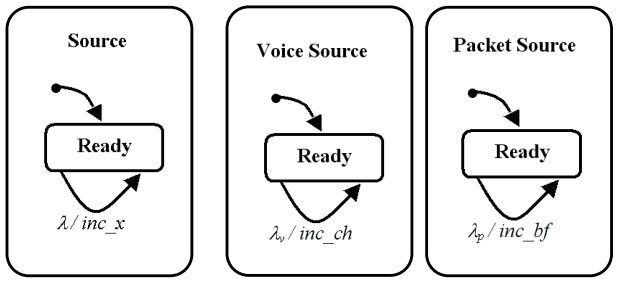


Figura 3.4: Especificação Statecharts para o esquema de alocação de recursos 1

Cada serviço é gerado obedecendo um processo de Poisson com média λ. Associado ao evento λ está a ação inc_x, que dispara uma nova chegada de um determinado serviço em um componente ortogonal. Neste trabalho, esse *template* é usado para gerar chamadas de voz GSM e pacotes GPRS.

Nesse esquema de alocação de recursos, as chamadas de voz são imediatamente servidas, caso haja disponibilidade de canais de rádio, e os pacotes GPRS são enfileirados no *buffer* enquanto aguardam por serviço (*vide* Fig.3.1). Assim, os eventos e ações λ_v/inc_ch e λ_p/inc_bf disparam uma nova chegada de uma chamada de voz ou de um pacote GPRS nos componentes ortogonais Voice Channel e Buffer que serão descritos a seguir. Os *templates* Source para a geração de voz e de pacotes GPRS são mostrados nas Fig.(3.5.b) e Fig.(3.5.c), respectivamente.

- Voice Channel: Como mostrado na Fig.(3.6), esse *template* possui dois subestados chamados **Free** e **Busy**. **Free** significa que não existem chamadas de voz no sistema,



(a) (b) (c)

Figura 3.5: *Templates*: (a) Source, (b) Voice Source e (c) Packet Source.

enquanto que, **Busy** significa que existe pelo menos uma chamada de voz no sistema. Como há N canais de rádios na BTS, o sistema pode estar ocupado com 1,2,3, ..., N chamadas. A ação *inc_ch*, descrita anteriormente, dispara uma nova chegada de uma chamada de voz nesse componente, mudando sua configuração de **Free** para **Busy** ou de **Busy** com v para $v+1$ chamadas de voz. v é a variável que representa o número de canais de rádios ocupados por serviços de voz.

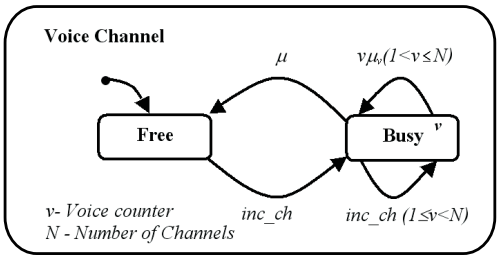


Figura 3.6: *Template* Voice Channel

- *Buffer*: Esse *template* possui dois subestados chamados **Free** e **Busy**, como mostrado na Fig.(3.7). **Free** significa que não existem pacotes GPRS no *buffer*, enquanto que, **Busy** significa que existe pelo menos um pacote GPRS no *buffer*. B_s é a capacidade de armazenamento do *buffer*. Esse último pode estar ocupado com 1,2,3, ..., B_s pacotes GPRS. A ação dec_bf decrementa um pacote GPRS por meio dos eventos *cond* 1 e *cond* 2, enquanto que, a ação inc_bf incrementa um pacote GPRS através dos eventos *cond* 3 e *cond* 4. Os eventos *cond* 1 até *cond* 4 serão descritos posteriormente.

- Packet Channel: Esse *template* possui dois subestados chamados **Free** e **Busy**, como mostrado na Fig.(3.8). **Free** significa que não existem pacotes GPRS em serviço, enquanto que, **Busy** significa que existe pelo menos um pacote em serviço. Essa informação é guardada na variável p. Quatro eventos que controlam a dinâmica desse *template*, são eles:

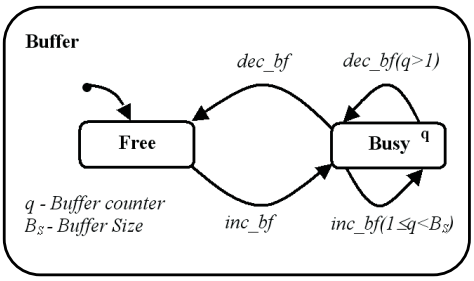


Figura 3.7: *Template Buffer*

Evento 1 (cond 1): Esse evento é reponsável pela transição **Free** para **Busy**, **se** é verdade que existem recursos de rádio disponíveis para servir um pacote GPRS ($v < N$), **e** existe pelo menos um pacote GPRS no *buffer* ($Buffer.Busy$). Por meio da ação dec_bf esse pacote entrará em serviço sendo, assim, retirado do *buffer*. . Esse evento é mostrado abaixo:

$$cond\ 1 = \{Tr[(v < N).AND.(in(Buffer.Busy))]\}/dec\_bf$$

Evento 2 (cond 2): Esse evento é reponsável pela transição de **Busy** com p para **Busy** com $p + 1$, **se** é verdade que existem recusos de rádio disponíveis para servir um pacote GPRS ($v < N$), **e** existe pelo menos um pacote no *buffer* ($Buffer.Busy$), **e** o número de pacotes em serviço é menor que a capacidade de rádio disponível ($p < N - v$). Assim, esse pacote será retirado do *buffer* e entrará em serviço por meio da ação dec_bf. Esse evento é mostrado abaixo:

$$cond\ 2 = \{Tr[(v < N).AND.(in(Buffer.Busy)).AND.(p < N - v)]\}/dec\_bf$$

Evento 3 (cond 3): Esse evento é reponsável pela transição de **Busy** com $p + 1$ para **Busy** com p, **se** é verdade que um pacote GPRS terminou seu serviço e deixou o sistema ($p\mu_p(p > 1)$), **ou se** é verdade que todos os canais de rádio estão ocupados ($p = N - v$), **e** uma chamada de voz chega no sistema (inc_ch). Essas duas últimas condições indicam que aconteceu uma preempção de um pacote. Tal pacote GPRS, será levado para o *buffer* por meio da ação inc_bf. Esse evento é mostrado abaixo:

$$cond\ 3 = \{p\mu_p(p > 1)\}.OR.\{Tr[(p = N - v).AND.(inc\_ch)]/inc\_bf\}$$

Evento 4 (cond 4): Esse evento é reponsável pela transição de **Busy** com 1 para **Free**, **se** é verdade que o único pacote terminou seu serviço e deixou o sistema (μ_p), **ou se** é verdade que todos os canais de rádio estão ocupados ($p = N - v$), **e** o número

de pacotes GPRS em serviço é igual a um ($p = 1$), **e** uma chamada de voz chega no sistema (inc_ch). Essas três últimas condições indicam que aconteceu a preempção do único pacote GPRS em serviço. Novamente, esse pacote será levado para o *buffer* por meio da ação inc_bf. Nos eventos 3 (cond 3) e 4 (cond 4), o pacote que sofreu preempção, somente será aceito no *buffer* caso exista espaço. O evento 4 é mostrado abaixo:

$$cond\ 4 = \{\mu_p\}.OR.\{Tr[(p = N - v).AND.(p = 1).AND.(inc\_ch)]/inc\_bf\}$$

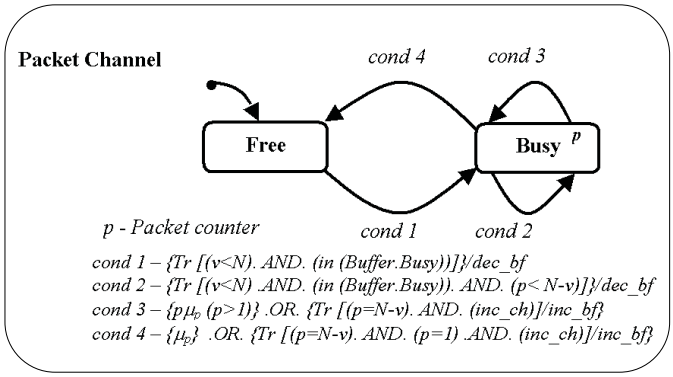


Figura 3.8: Template Packet Channel

Cada subestado da especificação descrita pode ser encontrado em uma determinda configuração. Os estados Source (Voice e Packet) estão sempre no estado **Ready** (R), o Buffer pode estar em **Free** (F) ou **Busy** com q pacotes GPRS ($B(q)$), o Voice Channel pode estar em **Free** (F) ou **Busy** com v chamadas de voz ($B(v)$), o Packet Channel pode estar em **Free** (F) ou **Busy** com p pacotes GPRS ($B(p)$). Conseqüentemente, o estado da cadeia de Markov gerada é

$$SC = (Voice\ Source,\ Packet\ Source,\ Buffer,\ Voice\ Channel,\ Packet\ Channel). \quad (3.1)$$

Porém, os estados Source (Voice e Packet) podem ser omitidos do estado da cadeia de Markov, uma vez que, eles nunca mudam a sua configuração. Dessa forma, o estado final da cadeia é dado por:

$$SC = (Buffer,\ Voice\ Channel,\ Packet\ Channel). \quad (3.2)$$

O processo de geração de uma cadeia de Markov a partir de uma especificação consiste em montar o gerador infinitesimal Q, o qual contém todas as transições (taxas) entre

os estados dessa cadeia. Para ilustrar esse procedimento, um exemplo de pequeno porte considerando $N = B_s = 2$ é mostrado a seguir. A partir do conjunto de eventos e ações descritas na especificação, e de acordo com Eq.(3.2) é gerado o grafo da Fig.(3.9). A partir dele é, então, construído o gerador infinitesimal correspondente à especificação *Statecharts* (55), o qual é mostrado em (3.3).

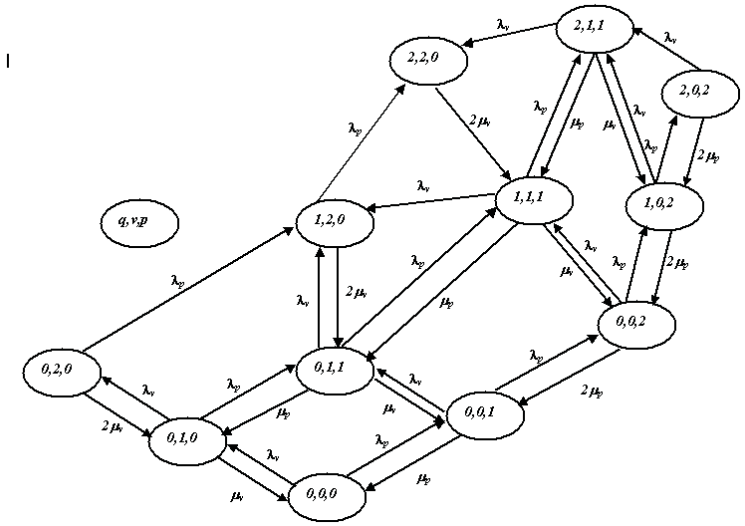


Figura 3.9: Grafo correspondente à especificação Statecharts para $N = B_s = 2$

$$Q = \begin{bmatrix} -(\lambda_p + \lambda_v) & \lambda_p & \lambda_v & 0 & . & . & . & 0 \\ \mu_p & -(\lambda_p + \lambda_v + \mu_p) & \lambda_p & \lambda_v & 0 & . & . & 0 \\ . & . & . & . & . & . & . & . \\ . & . & . & . & . & . & . & . \\ . & . & . & . & . & . & . & . \\ . & . & . & . & . & . & . & . \\ . & . & . & . & . & . & . & . \\ 0 & 0 & 0 & 0 & 0 & 2\mu_v & . & -2\mu_v \end{bmatrix} \tag{3.3}$$

A probabilidade do estado de equilíbrio $\pi = (\pi_0, \pi_1, \cdots, \pi_{max})$ pode ser calculada usando as Eqs(2.4).

A probabilidade de bloqueio de voz é dada pela Eq.(3.4). Devido à prioridade preemptiva do serviço de voz sobre o serviço GPRS, essa medida de desempenho deve concordar

com a fórmula de Erlang-B. Uma chamada de voz será bloqueada, toda vez que, mediante sua chegada, todos os canais de rádio estiverem ocupados por chamadas de voz.

$$P_{bv} = \sum_{q=0}^{B_s} \pi_{q,N,F}. \tag{3.4}$$

A probabilidade de bloqueio de um pacote GPRS é dada pela Eq.(3.5). O bloqueio de um pacote GPRS acontecerá sempre que o *buffer* estiver com sua capacidade esgotada.

$$P_{bd} = \sum_{q+v+p=N+B_s} \pi_{q,v,p}. \tag{3.5}$$

A probabilidade de preempção de um pacote é dada pela Eq.(3.6). Como descrito nos eventos 3 (*cond*3) e 4 (*cond*4), um pacote GPRS sofrerá preempção, toda vez que, uma chamada de voz chegar no sistema e encontrar todos os canais de rádio ocupados por voz e pacotes GPRS.

$$P_{pd} = \frac{1}{\lambda_p(1-P_{bd})} \sum_{\substack{v+p=N \\ \&p\geq 1}} \lambda_v \pi_{q,v,p}. \tag{3.6}$$

A probabilidade de descarte de um pacote é dada pela Eq.(3.7). Isso acontecerá, quando após ter sofrido preempção, ele não consegue ser admitido pelo sistema devido à indisponibilidade de espaço no *buffer*.

$$P_{dd} = \frac{1}{\lambda_p(1-P_{bd})} \sum_{\substack{v+p=N \\ \&p\geq 1}} \lambda_v \pi_{B_s,v,p}. \tag{3.7}$$

A fórmula de Little é usada para derivar o tempo médio de espera por serviço de um pacote GPRS (atraso médio), o qual é dado pela Eq.(3.8). Em sistemas de perda, a fórmula de Little deve ser corrigida para incorporar somente os clientes que não sofrem perda ao longo do sistema. Assim, o dividendo da Eq.(3.8) é dado pelo número médio de pacotes no *buffer*, enquanto que, o divisor é a taxa média de pacotes que o acessam. Como é mostrado na Fig.3.1, a população total que chega no *buffer* é dada pelos novos pacotes GPRS e por aqueles que sofreram preempção, mas, não foram descartados. O mesmo raciocínio se aplica no cálculo da vazão média dada na Eq.(3.9). A única diferença é que o pacote GPRS que não é bloqueado e não sofre preempção é considerado.

$$W_{qd} = \frac{\sum_{q=1}^{B_s} \sum_{v=0}^{N} \sum_{p=0}^{N-v} q\pi_{q,v,p}}{\lambda_p(1-P_{bd})[1+P_{pd}(1-P_{dd})]}. \tag{3.8}$$

$$T_h = \lambda_p(1-P_{bd})[(1-P_{pd}) + P_{pd}(1-P_{dd})]. \tag{3.9}$$

3.2.2 Esquema de alocação de recursos 2

Como mostrado na Fig.3.2, esse esquema de alocação de recursos é similar ao anterior. A única diferença é o número de canais de rádio (PDCH) que são reservados para o uso exclusivo do GPRS. Assim, o número de canais de rádios disponíveis para o escoamento do tráfego de voz é $N_v = N - N_d$, onde N_d é o número de PDCH.

A especificação correspondente a esse esquema é mostrada na Fig.3.10. Ela é muito similar a especificação da Fig.3.4. A diferença está na política de admissão de chamadas de voz no template Voice Channel. Assim, uma chamada é admitida pelo sistema somente se $v < N_v$. Da mesma forma, há uma mudança na política de admissão de um pacote GPRS. Assim, um pacote GPRS é imediatamente admitido e colocado em serviço, se não existirem pacotes em serviço (evento *cond* 1). Um pacote GPRS será ainda admitido e colocado em serviço, enquanto houver disponibilidade de canais de rádio (evento *cond* 2). Os demais eventos seguem o mesmo raciocínio da especificação anterior.

A probabilidade de bloqueio de uma chamada de voz é dada por

$$P_{bv} = \sum_{p=0}^{N-N_v} \sum_{q=0}^{B_s} \pi_{q,N_v,p}. \tag{3.10}$$

A probabilidade de bloqueio de uma pacote GPRS é dada por

$$P_{bd} = \sum_{q+v+p=N+B_s} \pi_{q,v,p}. \tag{3.11}$$

A preempção, claramente expressa no evento *cond* 3, acontecerá sempre que uma chamada de voz chegar no sistema e encontrar os canais de rádio ocupados por voz e pacotes GPRS. Essa medida é dada por

$$P_{pd} = \frac{1}{\lambda_p(1 - P_{bd})} \sum_{\substack{v+p=N \\ \&p \geq N_d}} \lambda_v \pi_{q,v,p}. \tag{3.12}$$

A probabilidade de descarte de um pacote GPRS é dada por

$$P_{dd} = \frac{1}{\lambda_p(1 - P_{bd})} \sum_{\substack{v+p=N \\ \&p \geq N_d}} \lambda_v \pi_{B_s,v,p}. \tag{3.13}$$

O atraso médio de um pacote é dado pela Eq.(3.14), enquanto que, a vazão é dada novamente pela Eq.(3.9).

$$W_{qd} = \frac{\sum_{q=1}^{B_s} \sum_{v=0}^{N_v} \sum_{p=0}^{N-v} q\pi_{q,v,p}}{\lambda_p(1 - P_{bd})[1 + P_{pd}(1 - P_{dd})]}. \tag{3.14}$$

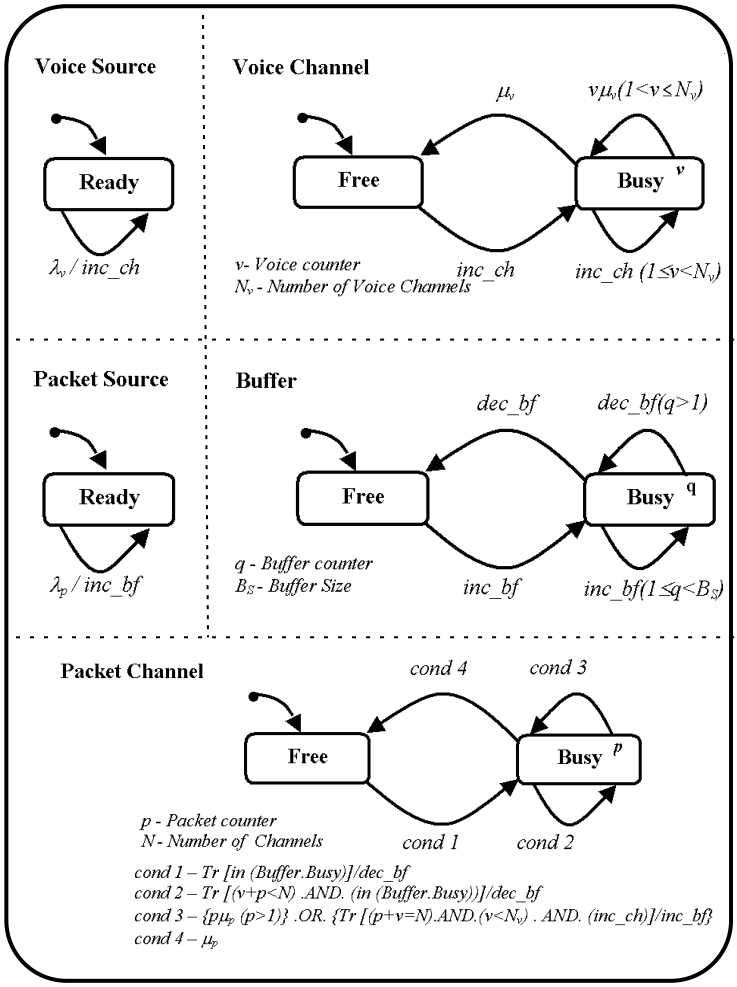


Figura 3.10: Especificação Statecharts para o esquema de alocação de recursos 2

3.2.3 Esquema de alocação de recursos 3

Nesse esquema, as chamadas de voz são encaminhadas para o *buffer*, enquanto que, os pacotes são admitidos se houver disponibilidade de recursos de rádio. Além disso, como mostrado na Fig.(3.3), não há prioridade das chamadas de voz sobre os pacotes GPRS. A especificação desse esquema de alocação de recursos é mostrada na Fig.(3.11).

As chamadas de voz sendo encaminhadas para o *buffer*, e os pacotes GPRS buscando por serviço são claramente representados nos templates Voice e Packet Source, respectivamente, por meio dos eventos λ_v/inc_bf e λ_p/inc_ch. A admissão de uma chamada de voz está condicionada ao número de pacotes GPRS em serviço $Tr(p < N)$ e $Tr(p + v < N)$, e vice e versa $inc\_ch(v < N)$ e $inc\_ch(p + v < N)$, como mostram os templates Voice e Packet Channel.

A probabilidade de bloqueio de uma chamada de voz é dada pela Eq.(3.15). O

bloqueio ocorrerá sempre que o *buffer* estiver cheio.

$$P_{bv} = \sum_{q+v+p=N+B_s} \pi_{q,v,p}. \tag{3.15}$$

A probabilidade de bloqueio de um pacote GPRS é dada por

$$P_{bd} = \sum_{q=0}^{B_s} \sum_{v+p=N} \pi_{q,v,p}. \tag{3.16}$$

A vazão é dada pela Eq.(3.17). Novamente, essa medida deve levar em consideração somente os pacotes aceitos no sistema.

$$T_h = \lambda_p(1 - P_{bd}). \tag{3.17}$$

Uma medida de desempenho de muita relevância nesse esquema é o tempo médio de espera por serviço de uma chamada de voz Eq.(3.18). Isso porque, as chamadas de voz que não são imediatamente servidas são acomodadas no *buffer*.

$$W_{qv} = \frac{\sum_{q=1}^{B_s} \sum_{v=0}^{N} \sum_{p=0}^{N-v} q\pi_{q,v,p}}{\lambda_v(1 - P_{bv})}. \tag{3.18}$$

3.3 Resultados

Para a análise de desempenho da rede GSM/GPRS segundo os parâmetros de QoS mostrados para os esquemas de alocação de recursos apresentados, utilizou-se, por simplicidade, um ambiente com uma única célula. Assim, o efeito da mobilidade não foi considerado. Além disso, somente uma portadora GSM foi usada. Contudo, um canal de rádio é dedicado para a sinalização, restando assim 7 canais TDMA para o escoamento do tráfego oferecido. No esquema de alocação de canal 2, o número de canais dedicados para o GPRS (PDCH) é 2. Para reduzir o impacto da prioridade preemptiva dos serviços de voz, o tamanho do *buffer* (B_s) é igual a 7. Esse mesmo valor de B_s é usado no esquema de alocação de recurso 3. Todos os pacotes GPRS possuem a mesma prioridade, além disso, é considerada somente a operação *singleslot*. Uma vez em serviço, um pacote GPRS, caso não sofra preempção, ocupa o recurso de rádio até o final da sua transmissão. O tamanho da mensagem GPRS é distribuído exponencialmente com média 2 x 13.4 Kbit, resultando em um tempo médio de serviço de $2s$ (CS-2). O tempo de serviço de voz também segue uma distribuição exponencial e possui média $180s$. Para validar os resultados do modelo analítico, seus resultados são comparados com os

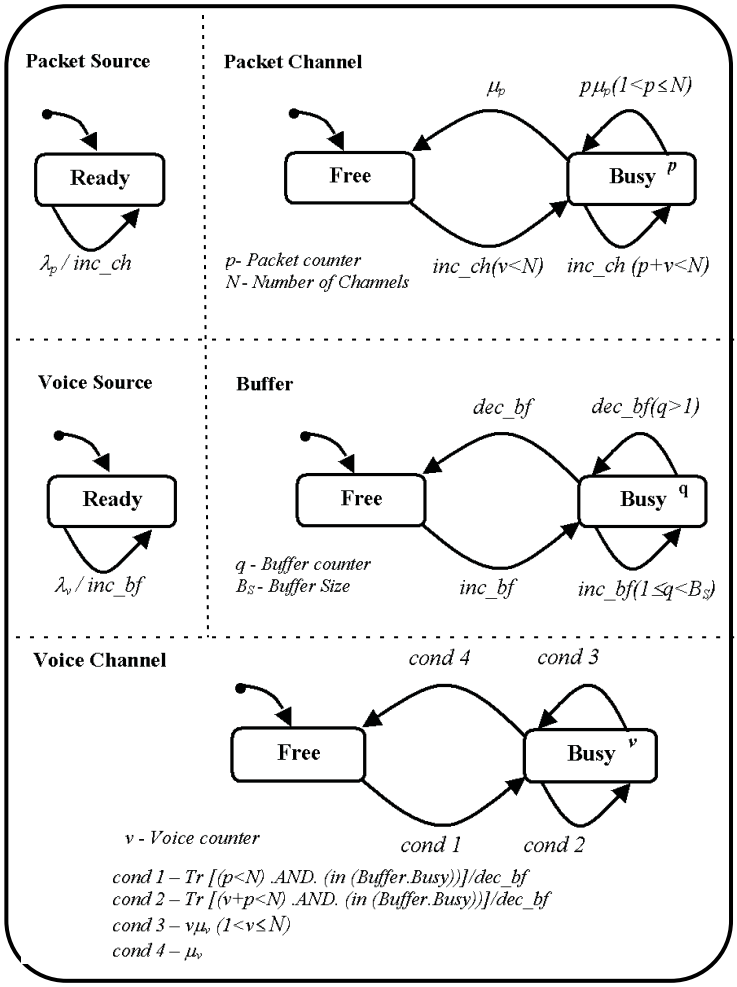


Figura 3.11: Especificação Statecharts para o esquema de alocação de canal 3

resultados simulados usando uma extensão do C chamada SMPL. O intervalo de confiança usado é de 95%.

Nesse caso o tráfego oferecido GSM e GPRS aumentara de 2,5 até 7,5 Erlang, resultando em tráfego oferecido de voz e dados variando de 5 até 15 Erlang. De acordo com a fórmula de Erlang-B, o valor de 2,5 Erlang para 7 canais de rádio fornece uma GoS de 1% para o serviço de voz.

A probabilidade de bloqueio de uma chamada de voz para os três esquemas de alocação de recursos é mostrada na Fig.(3.12.a). O melhor desempenho foi obtido pelo esquema 3, seguido do esquema 1, e depois do esquema 2. Pode-se observar que, o enfileiramento de chamadas de voz no *buffer* consegue suprir a falta da prioridade preemptiva fazendo com que o desempenho do esquema de alocação de recursos 3 seja melhor que dos demais. Essa figura mostra ainda que a reserva de canais de rádio para o GPRS degrada o desempenho do sistema de voz, fazendo com que a probabilidade de bloqueio de voz do esquema 2 seja maior que a

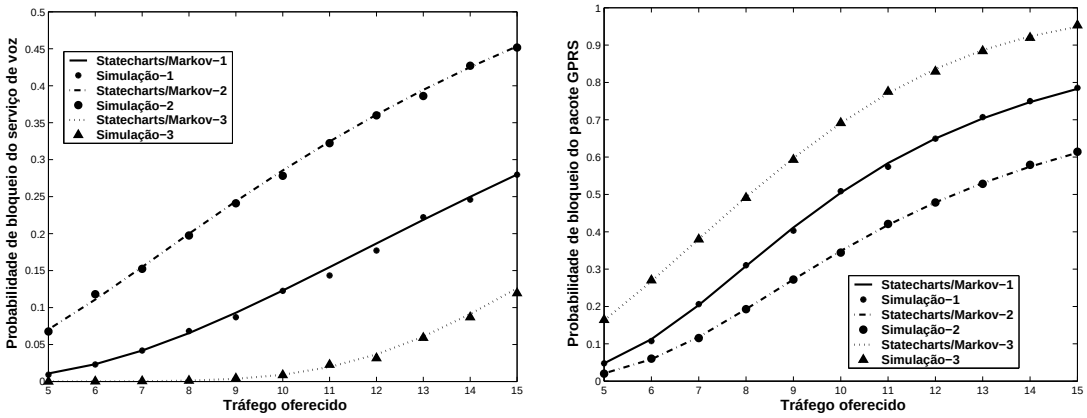
do esquema 1.

Contudo, como pode ser observado nas Fig.(3.12.b)-Fig.(3.14), o esquema de alocação de recursos 2 obtêm o melhor desempenho para a prestação de serviço de dados. Isso ocorre devido à reserva dos canais de rádio para o uso exclusivo do GPRS.

Na Fig.(3.12.b), observa-se que a ausência do *buffer* no esquema 3, tem um grande impacto do desempenho do GPRS aumentando a sua probabilidade de bloqueio de pacotes. Da mesma forma, a ausência da reserva de recursos no esquema 1 faz com que a sua probabilidade de bloqueio de dados aumente significativamente para altos valores de tráfego de voz.

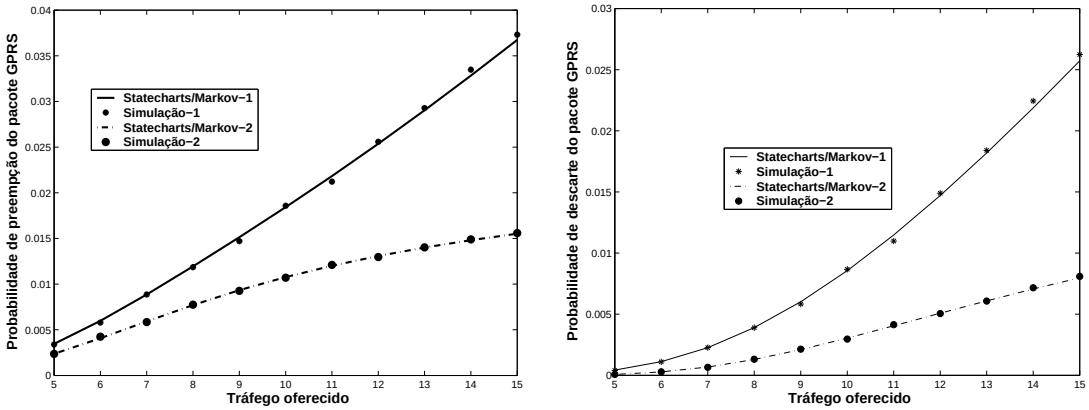
A reserva de canais de rádio diminui a probabilidade de preempção, e conseqüentemente a probabilidade de descarte de um pacote GPRS, como pode ser observado na Fig.(3.13). Além disso, ela proporciona o menor atraso médio e a maior vazão de um pacote GPRS, Fig.(3.14).

Na Fig.(3.15), nota-se que para altos valores de tráfego oferecido, um usuário de voz tem que aguardar pelo serviço no *buffer* aproximadamente um minuto.



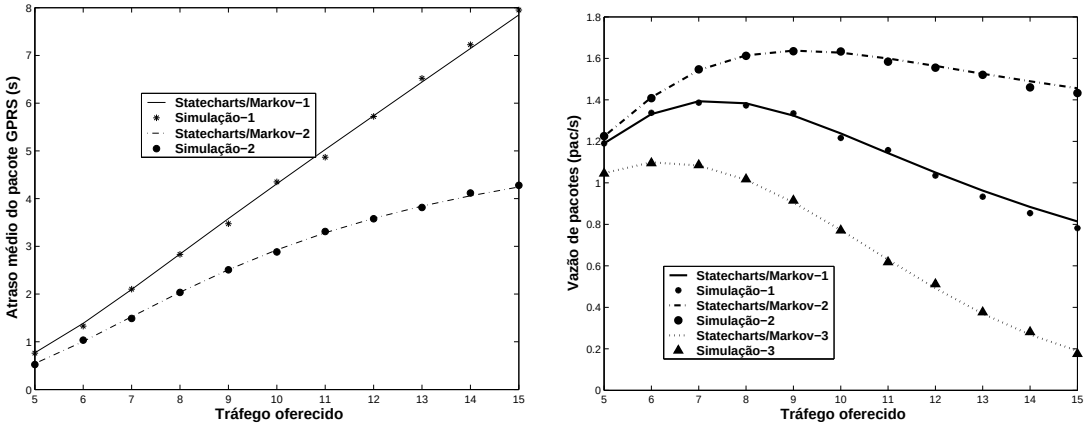
(a) (b)

Figura 3.12: Probabilidades de bloqueio: (a) Voz e (b) dados



(a) (b)

Figura 3.13: Probabilidades de: (a) Preempção e (b) Descarte



(a) (b)

Figura 3.14: Pacote GPRS: (a) Atraso médio e (b) Vazão

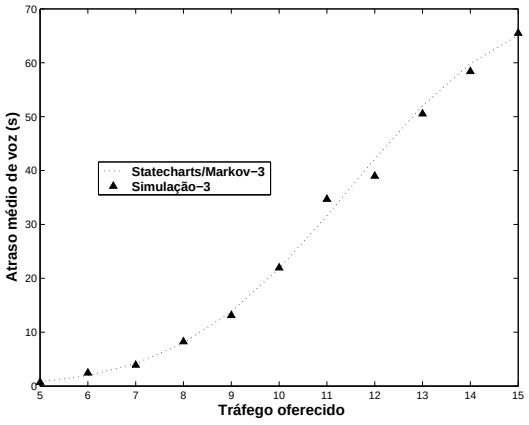


Figura 3.15: Tempo médio de espera de uma chamada de voz

Capítulo 4

Análise de desempenho de redes móveis hierárquicas

O objetivo deste capítulo é prover modelos analíticos para a avaliação de desempenho de redes móveis celulares hierárquicas. Dois modelos serão apresentados. O primeiro já publicado em (87)(88), é um *benchmark*, portanto, serve como uma base de comparação. O segundo, proposto neste trabalho, é uma forma alternativa de prover o atendimento de chamadas de dados de modo a não impactar no serviço de voz. Os resultados mostrarão que o esquema proposto melhora consideravelmente a QoS do serviço de dados.

Outro aspecto importante é que neste trabalho é utilizada a tecnologia GSM/(E)GPRS para a parametrização dos modelos. Contudo, eles podem ser aplicados à outras tecnologias.

4.1 Modelagem

4.1.1 Tráfego de voz

Os processos de chegada das novas chamadas de voz e sessões (E)GPRS são processos de Poisson mutuamente independentes com médias iguais a $\lambda_{n,GSM}$ e $\lambda_{n,(E)GPRS}$, respectivamente (41). Da mesma forma, os processos de chegada de *handovers* GSM e (E)GPRS são processos de Poisson mutuamente independentes com médias iguais a $\lambda_{h,GSM}$ e $\lambda_{h,(E)GPRS}$. Assim, os tráfegos oferecidos de voz e dados são também processos de Poisson com médias dadas por:

$$\lambda_{GSM} = \lambda_{n,GSM} + \lambda_{h,GSM}. \tag{4.1}$$

$$\lambda_{(E)GPRS} = \lambda_{n,(E)GPRS} + \lambda_{h,(E)GPRS}. \tag{4.2}$$

Os tempos de residência e duração de uma chamada nas microcélulas e na macrocélula são distribuídos exponencialmente com parâmetros $1/\mu_{h,GSM}$, $1/\mu_{d,GSM}$, $1/\mu^{M}_{h,GSM}$ e $1/\mu^{M}_{d,GSM}$, respectivamente. Da mesma forma, os tempos de residência e duração de uma sessão de dados nas micros e macrocélula são distribuídos exponencialmente com médias $1/\mu_{h,(E)GPRS}$, $1/\mu_{d,(E)GPRS}$, $1/\mu^{M}_{h,(E)GPRS}$ e $1/\mu^{M}_{d,(E)GPRS}$. Assim, o tempo de retenção de canal para o serviço de voz e dados são também variáveis aleatórias distribuídas exponencialmente com médias:

$$1/\mu_{GSM} = 1/(\mu_{h,GSM} + \mu_{d,GSM}). \tag{4.3}$$

$$1/\mu_{(E)GPRS} = 1/(\mu_{h,(E)GPRS} + \mu_{d,(E)GPRS}). \tag{4.4}$$

$$1/\mu^{M}_{GSM} = 1/(\mu^{M}_{h,GSM} + \mu^{M}_{d,GSM}). \tag{4.5}$$

$$1/\mu^{M}_{(E)GPRS} = 1/(\mu^{M}_{h,(E)GPRS} + \mu^{M}_{d,(E)GPRS}). \tag{4.6}$$

4.1.2 Tráfego de dados

O modelo de tráfego de Internet usado é definido pelo 3GPP, o qual consiste em uma seqüência de chamadas de pacotes (*packet calls*) e tempos de leitura (*reading times*) como mostra a Fig.(4.1) (85). O usuário inicia uma chamada de pacotes quando solicita uma determinada informação. Durante essa chamada, vários pacotes IP podem ser gerados. Uma sessão de dados (*packet session*) pode possuir, dependendo da aplicação, várias chamadas de pacotes. Após o *download*, o usuário consumirá algum tempo analisando o documento solicitado. Esse tempo é chamado de tempo de leitura.

O número de chamadas de pacotes dentro de uma sessão de dados é distribuído geometricamente com média N_{pc}. O tempo de leitura é uma v.a. distribuída exponencialmente com média dada por D_{pc}. O número de pacotes IP dentro de uma chamada de pacotes é distribuído geometricamente com média N_d, enquanto que, o tempo entre chegadas desses

pacotes é também distribuído exponencialmente com média D_d[1]. É importante observar que a utilização de distribuições exponenciais e geométricas é de fundamental importância para a construção de modelos Markovianos, uma vez que, elas possuem a propriedade do esquecimento.

A Tabela 4.1 mostra alguns valores médios típicos usados na caracterização do tráfego *Web*. O tempo entre chegada dos pacotes IP é ajustado para diferentes níveis de atividade da fonte. Na referência (85), o tamanho de um pacote IP segue uma distribuição de Pareto com média 480 *bytes*. Porém, neste trabalho será considerado que essa distribuição é exponencial. Isso, de acordo com a literatura (43), garante a geração de modelos analíticos.

Tabela 4.1: Característica do modelo de tráfego de dados *ON-OFF*.

Taxa média de bits da fonte	N_{pc}	D_{pc}	N_d	D_d
8 kbit/s	5	412	25	0,5
32 kbit/s	5	412	25	0,125
64 kbit/s	5	412	25	0,0625

Lindemann e Thummler em (41) representaram esse modelo de tráfego usando um processo de Poisson Interrompido (IPP) de dois estados onde, durante o estado *ON*, os pacotes IP são gerados de acordo com uma distribuição exponencial com parâmetro $\lambda_{packet} = 1/D_d$. No estado *OFF* não há geração de pacotes. Os tempos de permanência nos estados *ON* e *OFF* são distribuídos exponencialmente com parâmetros $\alpha = 1/(N_d D_d)$ e $\beta = 1/D_{pc}$. O tempo de serviço de uma sessão *Web* é distribuído exponencialmente com média $1/\mu_{d,(E)GPRS} = N_{pc}(D_{pc} + N_d D_d)$.

Para simplificar a notação, uma chamada de pacotes será denominada de um documento *Web* e uma sessão de dados será chamada de uma sessão (E)GPRS.

4.1.3 Rede hierárquica

A rede móvel celular hierárquica sob análise possui duas camadas, uma inferior contendo ψ microcélulas, e outra superior correspondente à macrocélula. Essa rede é considerada homogênea. Desse modo, todas as células pertencentes a mesma camada são estatisticamente idênticas. Assim, no estado de equilíbrio, o comportamento geral de uma camada pode ser

[1]Em (85) as distribuições temporais apresentadas são consideradas geométricas pois foi usada uma escala de tempo discreta.

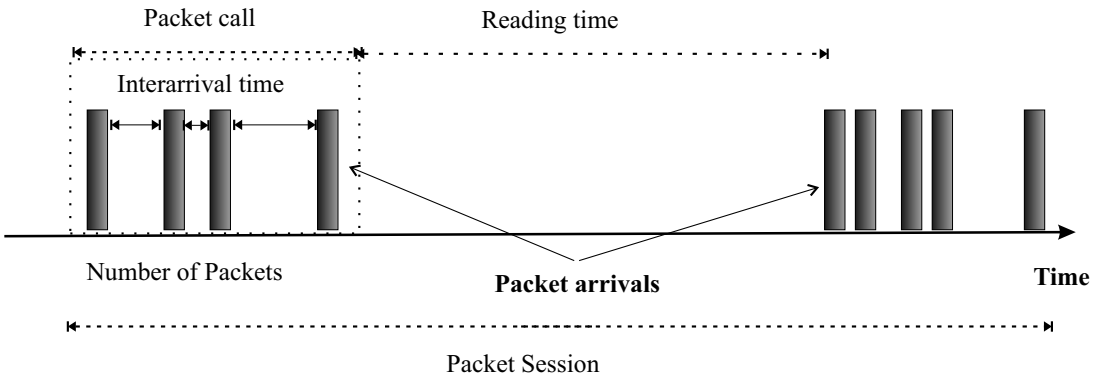


Figura 4.1: Modelo do tráfego de dados

analisado considerando apenas uma célula (2)(60). Além disso, é considerado um transbordo unidirecional, isto é, somente será atendido o transbordo do tráfego de voz das microcélulas para a macrocélula.

Cada microcélula da rede possui N canais de rádio que são usados para escoar o tráfego de voz e dados. As chamadas de voz chegam no sistema de acordo com uma distribuição de Poisson com taxa dada pela Eq.(4.1), e são atendidas imediatamentes se existirem recursos de rádio disponíveis na microcélula, caso contrário, transbordam para a macrocélula. É atribuído a esse serviço uma prioridade preemptiva sobre o serviço de dados. Assim, o seu desempenho pode ser medido através da fila M/M/N/N. As sessões (E)GPRS chegam no sistema de acordo com um processo de Poisson com média dada pela Eq.(4.2). Após suas chegadas, têm-se início a geração dos pacotes IP. No *buffer* são armazenados os pacotes IP pertencentes a um documento *Web*. O *buffer* possui uma capacidade de armazenamento igual a B_s. No modelo proposto é adicionado a esse sistema um *threshold* (T_h), que é usado como referencial para o roteamento das sessões de dados. A Fig.(4.2) mostra o sistema de fila usado para representar as microcélulas.

No esquema prioridade de voz, a macrocélula atenderá somente o tráfego transbordado de voz. Por outro lado, no esquema proposto, a macrocélula servirá além desse tráfego, aquele referente à sessões roteadas. Dessa forma, ela apresenta o mesmo comportamento da microcélula, sendo que, deve-se atentar para a diferença dos tráfegos oferecidos de voz e dados que na macrocélula são os tráfegos de transborado e roteado.

4.1.3.1 Esquema de alocação de recursos prioridade de voz

O sistema de fila correspondente à microcélula é modelado por uma cadeia de Markov com o seguinte estado:

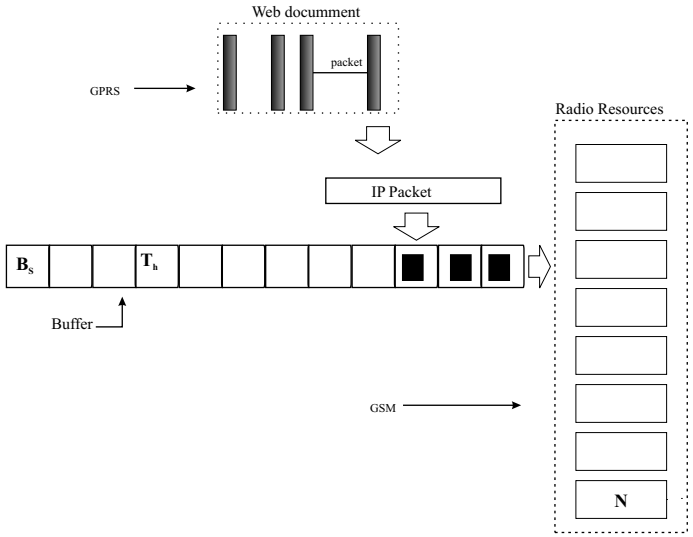


Figura 4.2: Sistema de fila usado na integração dos serviços de voz e dados

$$S = \{(v, k, m, r)/0 \leq v \leq N, 0 \leq k \leq B_s, 0 \leq m \leq M, 0 \leq r \leq m\} \tag{4.7}$$

onde para cada estado: v é o número de chamadas de voz em serviço; k é o número de pacotes IP no *buffer*; m é o número de sessões (E)GPRS ativas ; e r é o número de sessões (E)GPRS no estado OFF. Assim, se existirem m sessões ativas, então r estão no estado OFF e $m - r$ no estado ON (41).

Na Tabela 4.2 são mostradas as possíveis transições a partir de cada estado $S = (v, k, m, r)$, juntamente com as condições, as taxas e os eventos. As mudanças na variável v são determinadas pelas chegadas e partidas das chamadas de voz. A chegada de uma sessão (E)GPRS incrementa a variável m. Uma sessão pode chegar no sistema de duas formas: no estado ON ou OFF com probabilidades $\frac{\beta}{\alpha+\beta}$ e $\frac{\alpha}{\alpha+\beta}$, respectivamente. A partida de uma sessão de dados causa o decremento nas variáveis m e possivelmente r da seguinte forma:

- se todas as sessões estiverem estado ON, isto é, $r = 0$, então somente a variável m é decrementada;
- se todas as sessões estiverem estado OFF, isto é, $r = m$, então, as variáveis m e r serão decrementadas;
- se $m > 0$ e $0 < r < m$, então, uma sessão no estado ON deixa o sistema com probabilidade $\frac{m-r}{m}$;
- da mesma forma, uma sessão no estado OFF deixa o sistema com probabilidade $\frac{r}{m}$.

A chegada e a partida de um pacote IP, provenientes da fragmentação do conteúdo Web das sessões de dados, incrementa e decrementa, respectivamente, o valor da variável k. A transmissão desse pacote é feita usando todos os recursos ociosos da rede. Neste trabalho, essa transmissão pode ocupar no máximo uma portadora, ou seja, 7 *timeslots*. O aumento ou a diminuição da rajada motiva as mudanças da variável r, a qual com taxa $(m-r)\alpha$ diminui a rajada do processo IPP, enquanto que, com taxa $r\beta$ aumenta a rajada. Isso significa, respectivamente, que uma sessão passou do estado *ON* para o *OFF* e vice-versa.

Tabela 4.2: Possíveis transições a partir do estado $S = (v, k, m, r)$ da microcélula do esquema prioridade de voz.

Estado Sucessor	Condição	Taxa	Evento
$(v{+}1,k,m,r)$	$v{<}N$	λ_{GSM}	Chegada de uma chamada de voz
$(v{-}1,k,m,r)$	$v{>}0$	$v\mu_{GSM}$	Partida de uma chamada de voz
$(v,k,m{+}1,r)$	$m{<}M$	$\frac{\beta}{\alpha+\beta}\lambda_{GPRS}$	Chegada de uma sessão GPRS no estado *ON*
$(v,k,m{+}1,r{+}1)$	$m{<}M$	$\frac{\alpha}{\alpha+\beta}\lambda_{GPRS}$	Chegada de uma sessão GPRS no estado *OFF*
$(v,k,m{-}1,r)$	$(m{>}0)\wedge(r{=}0)$	$m\mu_{GPRS}$	Partida de uma sessão GPRS
$(v,k,m{-}1,r{-}1)$	$(m{>}0)\wedge(r{=}m)$	$m\mu_{GPRS}$	
$(v,k,m{-}1,r{-}1)$	$(m{>}0)\wedge(0{<}r{<}m)$	$\frac{r}{m}m\mu_{GPRS}$	
$(v,k,m{-}1,r)$	$(m{>}0)\wedge(0{<}r{<}m)$	$\frac{m-r}{m}m\mu_{GPRS}$	
$(v,k{+}1,m,r)$	$(k{<}B_s)\wedge(m{>}0)\wedge(r{<}m)$	$(m{-}r)\lambda_{packet}$	Chegada de um Pacote IP
$(v,k{-}1,m,r)$	$(min(\theta,k){>}0)\wedge(k{>}0)$ $\theta{=}min(N{-}v,7)$	$min(\theta,k)\mu_{service}$	Transmissão de um Pacote IP
$(v,k,m,r{+}1)$	$r{<}m$	$(m{-}r)\alpha$	Diminuição da rajada
$(v,k,m,r{-}1)$	$r{>}0$	$r\beta$	Aumento da rajada

O valor da taxa de chegada do *handover* é calculado através do equilíbrio do fluxo de chegada e saída dos usuários de voz e dados, os quais são dados por (59)(41):

$$\lambda_{h,GSM} = \mu_{h,GSM} \sum_{v=1}^{N} \sum_{k=0}^{B_s} \sum_{m=0}^{M} \sum_{r=0}^{m} v\pi_{v,k,m,r} \tag{4.8}$$

$$\lambda_{h,GPRS} = \mu_{h,(E)GPRS} \sum_{v=0}^{N} \sum_{k=0}^{B_s} \sum_{m=1}^{M} \sum_{r=0}^{m} m\pi_{v,k,m,r} \tag{4.9}$$

onde $\{\pi_{v,k,m,r}/(v,k,m,r) \in S\}$ são as probabilidades do estado de equilíbrio da cadeia.

A escolha do processo estocástico usado na representação do tráfego de transbordo impacta diretamente na precisão dos resultados obtidos na análise. Duas opções são encontradas na literatura. A primeira considera que esse tráfego segue uma distribuição de Poisson,

e a segunda um processo IPP (2). Por simplicidade, assim como em (59), será considerada a primeira opção, pelas seguintes razões:

- Embora mais simples, ela mostra com uma determinada precisão as principais tendências apresentadas pelo sistema (2);
- A macrocélula do esquema proposto escoa os serviços de voz e dados. Assim, o uso de um processo IPP levaria a sua cadeia à um espaço de estados que a tornaria intratável matematicamente, pois, teriam-se processos IPPs representando os modelos de voz e dados;
- Por fim, o objetivo deste trabalho é propor um modelo que melhore a provisão da QoS em ambientes hierárquicos. Dessa forma, que a condição sobre o comportamento do transbordo será relaxada.

Todo o tráfego transbordado de voz das microcélulas será escoado pela macrocélula se existirem recursos. Esse tráfego é dado por:

$$\lambda_{tof} = \psi \lambda_{GSM} \sum_{k=0}^{B_s} \sum_{m=0}^{M} \sum_{r=0}^{m} \pi_{N,k,m,r} \tag{4.10}$$

onde ψ é o número de microcélulas.

A probabilidade de um pacote IP não ser aceito pelo sistema é dado por:

$$P_D = \sum_{v=0}^{N} \sum_{m=0}^{M} \sum_{r=0}^{m} \pi_{v,B_s,m,r} \tag{4.11}$$

A vazão média de pacotes IP na microcélula é dado pela Eq.(4.12), enquanto que, o tempo médio de espera por serviço de um pacote IP no *buffer* é dado pela Eq.(4.13):

$$X = \mu_{service} \sum_{v=0}^{N} \sum_{k>0}^{B_s} \sum_{m=0}^{M} \sum_{r=0}^{m} min(min(N-v,7),k)\pi_{v,k,m,r} \tag{4.12}$$

$$W_q = \frac{\sum_{v=0}^{N} \sum_{k=1}^{B_s} \sum_{m=0}^{M} \sum_{r=0}^{m} k\pi_{v,k,m,r}}{X}. \tag{4.13}$$

A macrocélula é modelada usando uma fila $M/M/N^M/N^M$, onde de acordo com a notação de Kendall N^M é o número de canais de rádio. Com isso, a probabilidade de uma chamada de voz ser bloqueada, P_{bv}^M, é dada pela fórmula de Erlang-B, e a probabilidade de uma chamada de voz ser bloqueada na microcélula e não conseguir serviço na macrocélula é dada por:

$$P_B = P_{bv}^M \sum_{k=0}^{B_s} \sum_{m=0}^{M} \sum_{r=0}^{m} \pi_{N,k,m,r}. \tag{4.14}$$

4.1.3.2 Esquema de alocação de recursos proposto

Esse modelo é similar ao mostrado anteriormente, sendo a principal diferença entre eles o roteamento das sessões de dados das microcélulas para a macrocélula. O critério empregado para o roteamento é baseado na ocupação do *buffer* e dos canais de rádio. Assim, quando a ocupação do *buffer* for superior a um dado *threshold* e não existirem recursos de rádio disponíveis, um documento *Web* (chamada de pacotes) é roteado da micro para macrocélula, o que, em outras palavras representa o roteamento de uma sessão (E)GPRS.

Na Tabela 4.3 são mostradas somente as transições relacionadas ao acontecimento desse evento, visto que, as demais são as mesmas. O primeiro evento, é a chegada de um pacote IP quando não há o roteamento da sessão. Para que isso aconteça basta que:

1. existam recursos de rádio disponíveis ou
2. a ocupação do *buffer* seja menor ou igual ao *threshold*.

Caso essas condições sejam satisfeitas, a sessão será mantida na microcélula na qual foi originada. Caso contrário, ou seja, não existam recursos de rádio disponíveis e a ocupação do *buffer* for superior ao *threshold*, esse documento *Web* e juntamente a sessão será roteada para a macrocélula.

A variável γ indica quantas sessões concorrentes possuem pacotes IP no *buffer*. A variável η representa o número de sessões que serão roteadas. Como só podem ser roteadas as sessões ativas, ela é o mínimo entre γ e $m - r$. ϕ é o número de pacotes IP retirados do *buffer* pertencentes à essas sessões.

Novamente, o tráfego de transbordo de voz das microcélulas para a macrocélula é dado pela Eq.(4.10). Do mesmo modo, a probabilidade de bloqueio de um pacote IP é dado pela Eq.(4.11), enquanto que, a vazão média e o tempo médio de espera por serviço de pacotes IP na microcélula são dados por Eq.(4.12) e Eq.(4.13), respectivamente.

Considerando que o tráfego roteado possui um comportamento Poissoniano, a taxa na qual uma sessão (E)GPRS é roteada da micro para a macrocélula é dada por:

$$\mu_{rot} = \sum_{k=T_h+1}^{B_s} \sum_{m=1}^{M} \sum_{r=0}^{m} \eta \lambda_{packet} (m - r) \pi_{N,k,m,r} \tag{4.15}$$

assim, o tráfego oferecido roteado de sessões é dado por $O_{rot} = \psi \mu_{rot}$.

Tabela 4.3: Possíveis transições a partir do estado $S = (v, k, m, r)$ da microcélula com o esquema proposto.

Estado Sucessor	Condição	Taxa	Evento
$(v,k+1,m,r)$	$(v=N)\wedge(m>0)\wedge(k\leq T_h)\wedge(r<m)\vee$ $(v<N)\wedge(m>0)\wedge(k<B_s)\wedge(r<m)$	$(m-r)\lambda_{packet}$	Chegada de um Pacote IP sem roteamento
$(v,\phi,m-\eta,r)$	$(v=N)\wedge(m>0)\wedge(T_h<k<B_s)\wedge(r<m)$ e $\gamma=\lfloor\frac{k-T_h+N_d-1}{N_d}\rfloor$ onde $\lfloor x\rfloor$ é o maior inteiro $\leq x$ e $\phi=max(k-\eta N_d,0)$ e $\eta=min(\gamma,m-r)$	$(m-r)\lambda_{packet}$	Chegada de um Pacote IP com roteamento

Observe que se poderia utilizar um processo IPP para modelar o tráfego roteado de sessões. Contudo, no modelo final teria-se um processo IPP, como tráfego roteado de sessões, que durante os seus momentos de atividades regeria um outro processo IPP representando o tráfego *ON-OFF*. Desse modo, o modelo da macrocélula teria um espaço de estados imenso dificultando a sua manipulação e solução.

Nesse esquema uma macrocélula agrega as classes de serviço que são oriundas das microcélulas. Assim, através dela é escoado o tráfego de transbordo de voz e o tráfego roteado (E)GPRS. Novamente, os recursos de rádio são completamente compartilhados. O serviço de voz possui prioridade sobre o (E)GPRS. Assim, novamente uma fila $M/M/N^M/N^M$ pode ser usada para representar o seu comportamento, onde N^M é o número de canais da macrocélula. Um *buffer* de capacidade B_s^M é usado para armazenar os pacotes IP pertencentes aos documentos *Web* que são roteados das microcélulas.

Essa cadeia é similar da Tabela 4.2. As diferenças entre elas são mostradas na Tabela abaixo, e estão relacionadas ao tráfego transbordado de voz (λ_{tof}) e roteado de sessões (O_{rot}).

Tabela 4.4: Possíveis transições a partir do estado $S^M = (v, k, m, r)$ da macrocélula.

Estado Sucessor	Condição	Taxa	Evento
$(v+1,k,m,r)$	$v<N^M$	λ_{tof}	Chegada de uma chamada de voz
$(v,k,m+1,r)$	$m<M^M$	$\frac{\beta}{\alpha+\beta}O_{rot}$	Chegada de uma sessão GPRS roteada no estado *ON*
$(v,k,m+1,r+1)$	$m<M^M$	$\frac{\alpha}{\alpha+\beta}O_{rot}$	Chegada de uma sessão GPRS roteada no estado *OF*

A probabilidade de uma chamada de voz ser transbordada da microcélula e não ser escoada pela macrocélula é novamente dada pela Eq.(4.14). A probabilidade de um pacote IP

não ser aceito pelo sistema devido ao transbordo do *buffer* da macrocélula, a vazão média dos pacotes IP na macrocélula e o tempo médio de espera por serviço de um pacotes IP no *buffer* são dados por Eq.(4.11), Eq.(4.12) e Eq.(4.13), respectivamente.

4.2 Resultados

Na Tabela 4.5 são fornecidos os valores usados na obtenção dos resultados que serão apresentados a seguir. Se não for especificada qualquer mudança, eles serão tomados como base em todos os experimentos. A proporção entre os tráfegos de voz e dados é de 90% e 10%, respectivamente. O Tempo médio de serviço do pacote IP é dado como aquele no qual um pacote IP é escoado por um canal usando um dado esquema de codificação. Assim, ele é dado por:

$$t_{service} = \frac{1}{\mu_{service}} = \frac{\frac{480*8}{1024}(kbits)}{th_{CS}(kbits/s)}. \tag{4.16}$$

onde, th_{CS} é o *throughput* do esquema de codificação de canal usado.

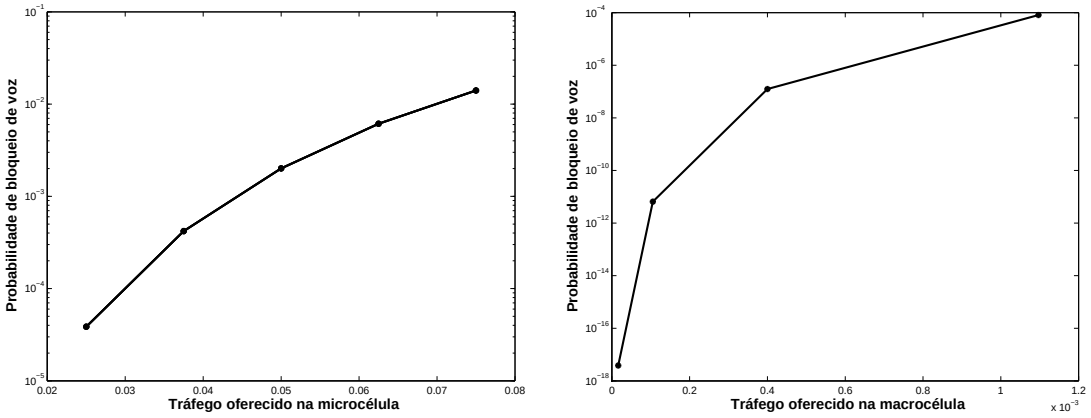
Tabela 4.5: Valores usados para a obtenção dos resultados.

Parâmetro		Valor
Número de canais na microcélula e macrocélula	N=N^M	7
Número médio de sessões (E)GPRS na microcélula e macrocélula	M=M^M	10
Tamanho do *buffer*	B_s=B_s^M	50
Threshold	T_h	25,35,40
Número de Microcélulas	ψ	19
Tempo médio de duração de uma chamada GSM (s)	$1/\mu_{d,GSM}$	120
Tempo médio de residência de uma estação móvel GSM (s)	$1/\mu_{h,GSM}$	60
Tempo médio de residência de uma estação móvel (E)GPRS (s)	$1/\mu_{h,(E)GPRS}$	120
Tempo médio de duração de uma chamada GSM na Macrocélula (s)	$1/\mu_{d,GSM}^M$	120
Tempo médio de residência de uma estação móvel GSM na Macrocélula (s)	$1/\mu_{h,GSM}^M$	$8/\mu_{d,GSM}$
Tempo médio de residência de uma estação móvel (E)GPRS na Macrocélula (s)	$1/\mu_{h,(E)GPRS}^M$	$8/\mu_{d,(E)GPRS}$
Tempo médio de leitura (s)	D_{pc}	41,2
Throughput do esquema de codificação (kbit/s)	$CS{-}2$	13,4

4.2.1 Desempenho do serviço de voz

É importante que a introdução do serviço de dados não altere o desempenho do serviço de voz, pois, sendo a aplicação preponderante na rede, ele é a fonte mais rentável das operadoras. Assim, desde que não haja reserva de recursos para o escoamento do tráfego de dados e seja mantida a prioridade preemptiva dos serviços de voz, o desempenho desse serviço não é afetado pelo serviço (E)GPRS.

Na Fig.(4.3) são mostradas as probabilidades de bloqueio de voz na microcélula e a probabilidade de bloqueio total dada pela Eq.(4.14). Na Fig.(4.3.a) observa-se que o bloqueio se mantêm dentro dos valores normalmente usados de 1% e 2%. Como já esperado, constata-se na Fig.(4.3.b) que a utilização de uma estrutura hierárquica é uma ótima alternativa para o escoamento do tráfego de voz que seria bloqueado em uma célula congestionada da rede.



(a) (b)

Figura 4.3: Probabilidade de bloqueio de voz (a)Microcélula; (b) Total

4.2.2 Efeito do *threshold*

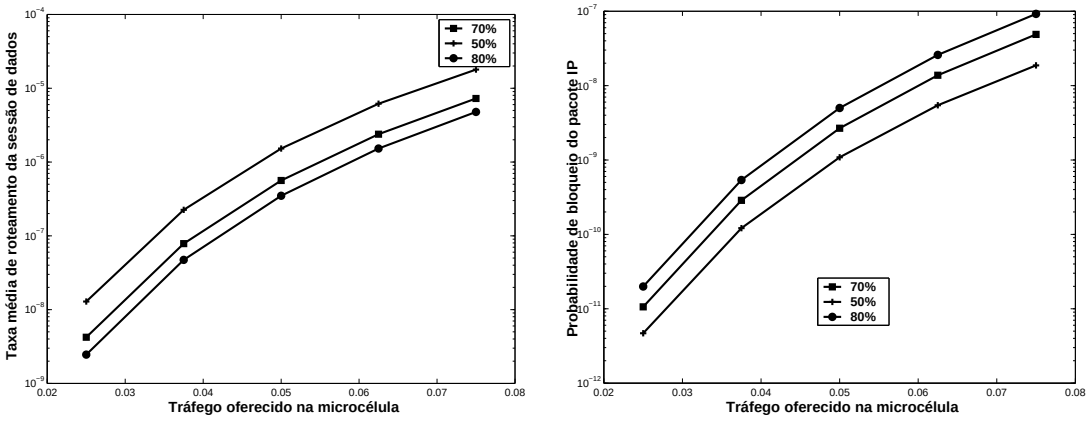
Nesta seção é analisado o efeito do *threshold* no desempenho do sistema. O seu valor tem um papel importante no esquema proposto, já que, sua escolha deve ser feita de forma a balancear a carga de tráfego entre as microcélulas e a macrocélula. Para esse experimento foi usado uma fonte de tráfego de 8kbits/s e valores de *threshold* de 25, 35 e 40 que correspondem à 50%, 70% e 80% de ocupação do *buffer*.

Na Fig.(4.4.a) observa-se que quanto menor o *threshold*, maior é a taxa média de roteamento. Esse comportamento é esperado, pois, com um *threshold* maior, o *buffer* da microcélula suporta mais pacotes IP, e conseqüentemente, um número maior de sessões de

dados concorrentes. Esse efeito é ratificado nas figuras Fig.(4.4.b) e Fig.(4.4.c), onde nota-se que o bloqueio e o atraso médio são maiores na célula com maior *threshold*, uma vez que, seu *buffer* acomoda mais pacotes.

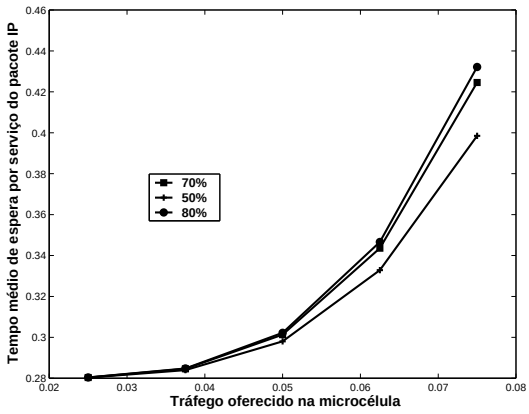
Com pode-se observa na Fig.(4.5), na macrocélula esse efeito é contrário, isto é, a probabilidade de bloqueio e o tempo médio de espera por serviço do pacote IP diminuem com o aumento do *threshold*, pois, a microcélula suporta um tráfego maior de documentos *Web*. Note ainda que os tempos médios de espera por serviço de um pacote IP são praticamente os mesmos para os três *thresholds* usados.

Assim, através da seleção do valor do *threshold* é possível inferir na interação entre as microcélulas e a macrocélula balanceando a carga de tráfego de dados entre elas.



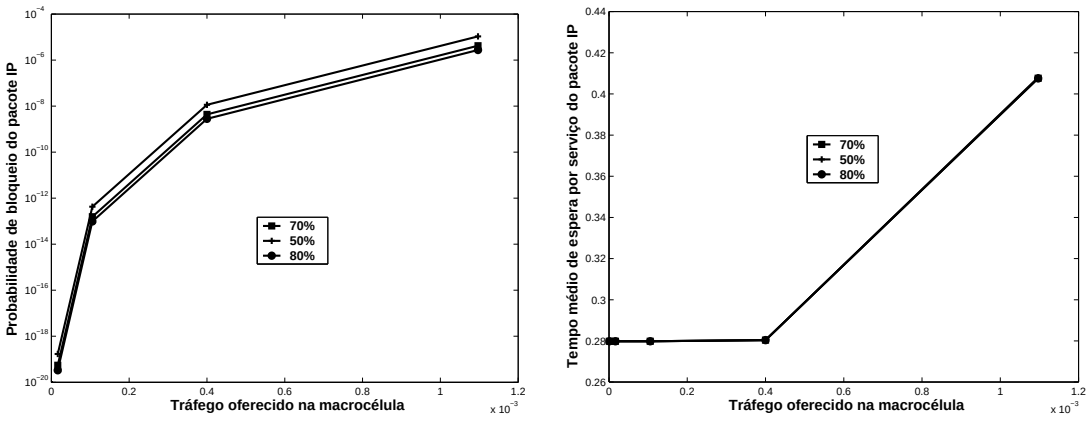
(a)

(b)



(c)

Figura 4.4: Microcélula:(a) taxa média de roteamento de uma sessão de dados (b) Probabilidade de bloqueio de pacotes IP (c) Tempo médio de espera por serviço



(a) (b)

Figura 4.5: Macrocélula:(a)Probabilidade de bloqueio de pacotes IP (b) Tempo médio de espera por serviço

4.2.3 Comparação entre os esquemas de alocação de recursos

Nesta seção é apresentada uma análise comparativa do desempenho dos dois esquemas de alocação de recursos descritos anteriormente . O valor escolhido do *threshold* foi T_h=35, ou seja, uma ocupação de 70% do tamanho do *buffer*. A análise é feita considerando os fatores de atividades da fonte de 8 kbits/s, 32 kbits/s e 64 kbits/s de modo que possa analisar o impacto dessa atividade no desempenho do sistema.

A Fig.(4.6.a) mostra que o esquema proposto melhora o desempenho da microcélula em relação ao atendimento do serviço de dados para todos os três fatores de atividade da fonte. Na Fig.(4.6.b) ilustra-se a redução na probabilidade de bloqueio conseguida pelo esquema proposto em relação ao prioridade de voz. Note que a vantagem do esquema proposto chega a 100% no caso de uma fonte de 8kbits/s. Para as demais fontes ela também é bastante significativa. O mesmo ganho em desempenho se observa para o tempo médio de espera por serviço, Fig.(4.6.c) e Fig.(4.6.d).

Na Fig.(4.7) é mostrada a característica do atendimento das sessões de dados roteadas das micros para as macrocélulas. Nota-se que esse tráfego roteado é facilmente escoado, mantendo valores bastante aceitáveis para a probabilidade de bloqueio e o tempo médio de espera por serviço de um pacote IP.

Os resultados apresentados até então quantificaram a melhora já esperada no desemempenho do sistema obtido pelo emprego do esquema proposto, pois, ao se rotear uma sessão de dados da microcélula para a macrocélula durante uma sobrecarga, libera-se o espaço no *buffer* dos pacotes IPs pertencentes a essa sessão. Assim, os os pacotes IP que chegarão no

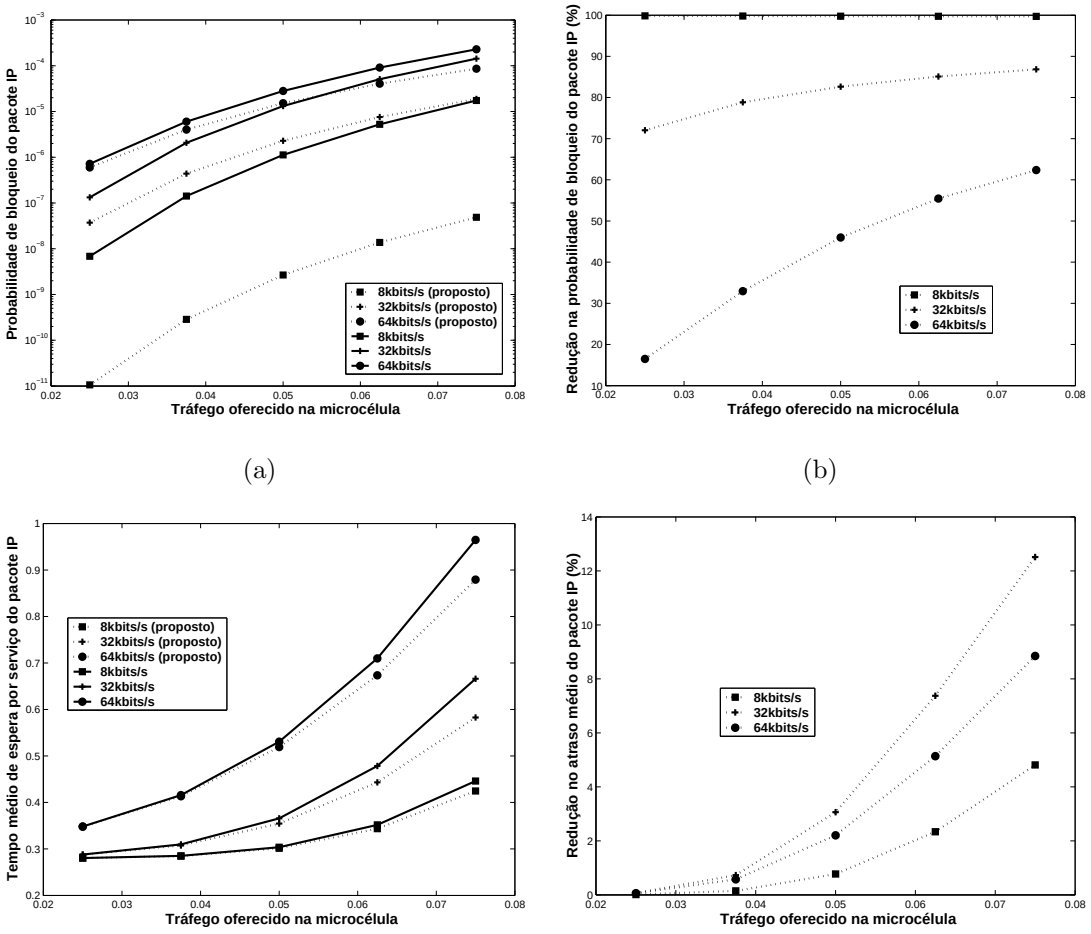
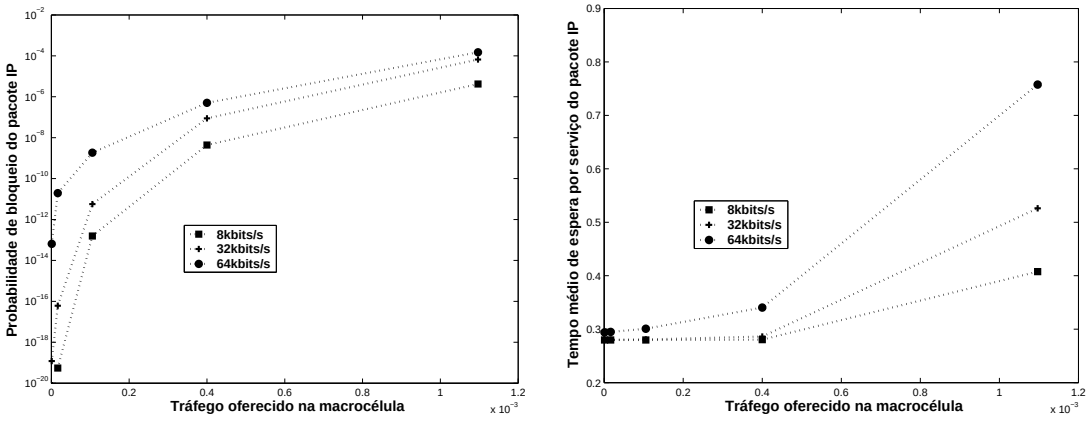


Figura 4.6: Microcélula:(a) Taxa média de roteamento da sessão de dados (a) Probabilidade de bloqueio do pacote IP, (b) Tempo médio por serviço do pacote IP

sistema encontrarão uma maior capacidade de armazenamento e assim experimentarão um menor bloqueio e atraso.

Porém, o que é importante analisar é se os pacotes IPs das sessões roteadas são melhores servidos do que seriam se ficassem na microcélula. Esse resultado é apresentado na Fig(4.8) o qual mostra o ganho no antedimento do esquema proposto em relação ao prioridade de voz. A Fig(4.8.a) diz que para os três fatores de atividade da fonte, o pacote IP roteado será atendido com uma probabilidade 100% maior na macrocélula. Para o atraso esse ganho chega aproximadamente à 50%, 40% e 20% para as fontes de 64 kbits/s, 32 kbits/s e 8 kbits/s; no pior caso, esses valores chegam, respectivamente, paras as fontes de 8 kbits/s, 32 kbits/s e 64 kbits/s à 75%, 53% e 34% para o bloqueio e 0,19% 2% e 15% para o atraso.

Essa última análise constata que o esquema proposto melhora consideravelmente o desempenho da provisão de QoS para o serviço de dados. Adicionalmente, ele não impacta na



(a) (b)

Figura 4.7: Macrocélula: (a)Probabilidade de bloqueio do pacote IP, (b)Tempo médio de espera por serviço do pacote IP

QoS dos serviços de voz. Assim, conclue-se a sua viabilidade, e recomenda-se o seu emprego em ambientes hierárquicos.

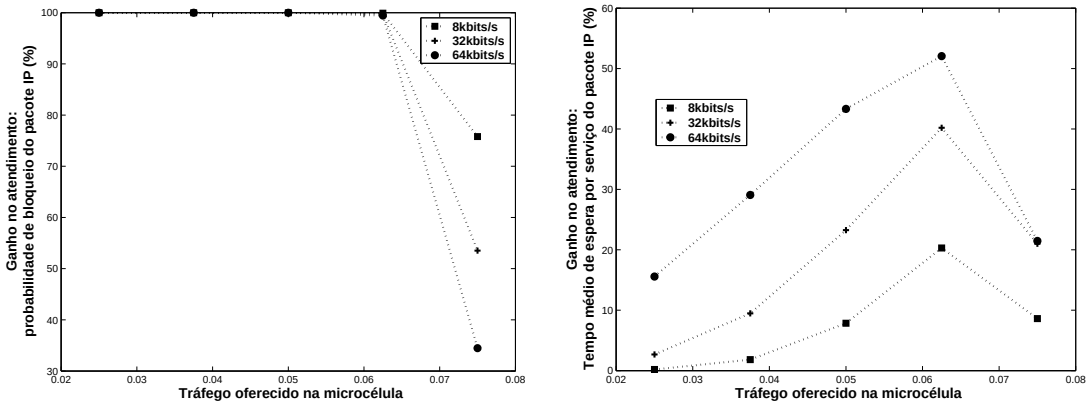
Um ponto observado em todos os resultados mostrados anteriormente mostra que o fator de atividade da fonte degrada consideravelmente a QoS dos serviços de dados. Para os dois esquemas de alocação de recursos, tanto na micro quanto na macrocélula, observa-se uma grande disparidade entre os valores assumidos das medidas de desempenho em relação à característica da fonte.

4.2.4 Comparação entre GPRS e EGPRS

Nesta seção é apresentada uma análise comparativa do desempenho entre o GPRS e EGPRS. O objetivo desse estudo é observar o ganho conseguido pelo EDGE através da modulação e dos esquemas de codificação de canal. Uma vez que, o emprego do esquema proposto resultou em um melhor desempenho, ele será usado em todos os experimentos. Foi considerado ainda: um *threshold* igual a T_h=35; fator de atividade da fonte de 32 kbits/s.

Considerou-se uma ótima qualidade de canal de forma que se possa usar na microcélula a modulação 8-PSK e o esquema MCS-9A (59,2 kbist/s) para o EGPRS, enquanto que, para o GPRS, CS-4 (21,4 kbits/s). Na macrocélula se usou uma modulação GMSK para ambos, e esquemas de codificação MCS-2B (11,2 kbist/s) e CS-2 (13,4 kbist/s) para o EGPRS e GPRS, respectivamente.

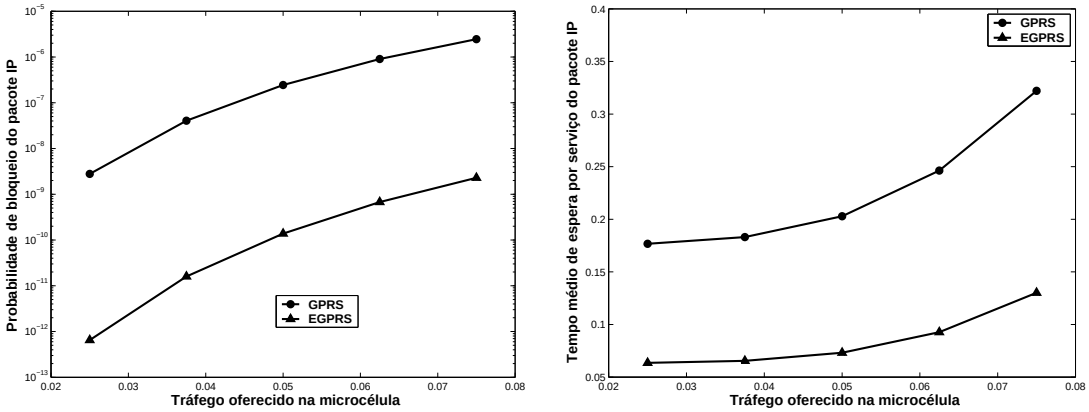
Na Fig.(4.9) se observa que a grande vantagem do EGPRS em relação a sua ca-



(a) (b)

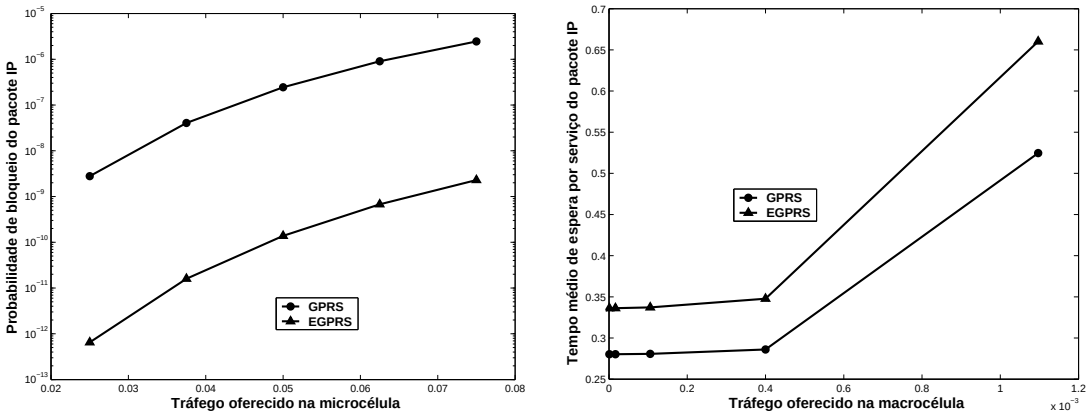
Figura 4.8: Ganho no atendimento: (a)Probabilidade de bloqueio do pacote IP (b)Tempo médio de espera por serviço do pacote IP

pacidade de escoamento do tráfego de dados sobre GPRS. Assim, por exemplo, para uma demanda de tráfego de voz e dados de 0.05 chamadas/s, a probabilidade de bloqueio de dados atinge níveis de 10^{-10} no EGPRS para 10^{-7} no GPRS. Além disso, os pacotes IP esperam um tempo menor no *buffer* para serem atendidos. Para a mesma demanda de tráfego citada anteriormente, um pacote IP espera cerca de 0,07 s no EGPRS contra 0,2 s no GPRS. Na macrocélula, Fig.(4.10), como os esquemas de codificação de canais usados apresentam *throughput* na mesma ordem de grandeza, os desempenhos são aproximadamente o mesmo em relação a probabilidade de bloqueio. Porém, quando considerado o tempo de atendimento, o GPRS consegue atender os pacotes ligeiramente mais rápidos.



(a) (b)

Figura 4.9: Microcélula: (a)Probabilidade de bloqueio de pacotes IP, (b) Tempo médio por serviço de um pacote IP



(a) (b)

Figura 4.10: Macrocélula (a)Probabilidade de bloqueio de pacotes IP, (b)Tempo médio de espera por serviço de um pacote IP

Capítulo 5

Análise de desempenho de esquemas de alocação de recursos adaptativos

A provisão da QoS em redes sem fio tem sido beneficiada pelo desenvolvimento de aplicações multimídia adaptativas que mudam dinamicamente sua largura de banda de acordo com as condições da rede. Tais serviços serão, assim, fortemente explorados de modo que será mandatório o emprego do mecanismo de adaptação de largura de banda juntamente ao Controle de Admissão de Chamadas (CAC) em redes móveis celulares. De fato, as redes de 3G já foram desenvolvidas para suportar esse procedimento (35).

Neste capítulo são apresentados dois modelos analíticos para a análise de desempenho de redes móveis celulares que empregam o CAC e o mecanismo de adaptação de largura de banda. Para mostrar a vantagem no atendimento conseguida por meio do emprego do procedimento de largura de banda juntamente ao CAC, é apresentado um esquema não adaptativo.

5.1 Modelagem

5.1.1 Rede

O ambiente multiserviço considerado suporta duas classes de serviço: Classe I e II. A classe I representa as classes de QoS Conversacional e *Streaming* definidas pelo 3GPP. Cada usuário dessa classe está equipado com *codecs* de taxas ajustáveis, como por exemplo MPEG-2 ou MPEG-4 (28)(35). Assim, é permitido que durante um congestionamento suas taxas possam ser ajustadas de acordo com a carga da rede (34). A Classe II representa a classe de QoS Interativa. Novamente, assim como no capítulo anterior, o *Web browsing* será

considerado como o principal serviço dessa classe.

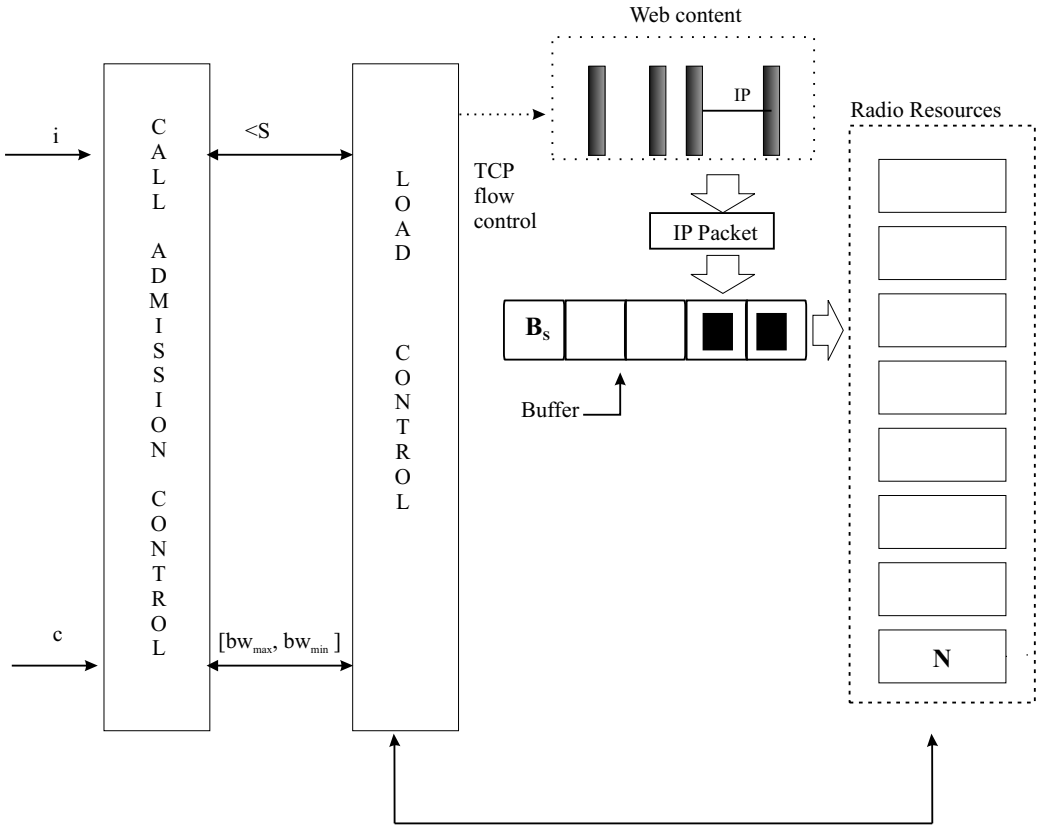


Figura 5.1: Sistema de Gerência de Recursos

A Fig.5.1 mostra o sistema de gerência de recursos usado. O módulo de Controle de Carga (*Load Control*-LC) monitora a carga de tráfego e informa periodicamente o CAC à respeito da condição da rede. O CAC aceitará ou rejeitará um serviço baseado nessa informação. Esse tipo de CAC que toma a sua decisão baseado na ocupação da rede é chamado, como já mencionado no capítulo 1, de CAC baseado em medidas.

Quando um serviço da classe I faz uma solicitação de uma conexão na rede especificando seu perfil de QoS em termos de de largura de banda [bw_{max}, bw_{min}], o CAC decide, baseado na informação proveniente do LC, se é possível aceitar esse serviço. De acordo com a política adotada, uma das possíveis situações pode ocorrer:

- Sem adaptação (SA): nesse esquema a chamada somente será aceita se existirem recursos de rádio suficiente para acomodá-la com a largura de banda máxima, bw_{max}.
- Com Adaptação (CA): nesse esquema, a largura de banda será negociada entre o usuário (aplicação) e a rede durante a etapa de conexão. Se existirem recursos de rádio disponíveis, o serviço será acomodado com a largura de banda máxima, se não, admitir-se-á com a largura de banda mínima. Se a rede estiver sobrecarregada e não existirem

recursos de rádio suficientes para acomodá-lo com largura de banda mínima, ele será bloqueado. Considera-se que, uma vez atendido os requerimentos de largura de banda, todos os demais parâmetros de QoS são satisfeitos. Quando uma chamada com largura de banda máxima deixa a rede, uma, com menor largura de banda, é promovida de forma a aumentar a satisfação do cliente e a utilização dos recursos de rádio.

- Adaptação Justa (AJ): nesse esquema, todos os serviços serão aceitos se possível com largura de banda máxima. Caso contrário, a largura de banda de todas as chamadas com banda máxima em serviço será reduzida para admitir um novo cliente com largura de banda mínima. Quando um cliente com essa largura de banda deixa o sistema, o CAC verifica se é possível promover o perfil de QoS de todos os cliente da rede. Esse esquema de alocação de recursos é denominado de justo, pois, a taxa de transmissão de todos os serviços em tempo real são reduzidas e elevadas igualmente de forma a beneficiar um novo serviço.

O LC também monitora o número de sessões *Web on line* na rede e informa ao CAC que tomará a decisão de aceitar ou não novas solicitações. Se for possível, o conteúdo dessa página *Web* é fragmentado em pacotes IP com taxa λ_{IP}. Esses pacotes aguardarão em um *buffer* com capacidade B_s, até serem transmitidos através da interface aérea. Durante um congestionamento na rede, o TCP utiliza o mecanismo de controle de fluxo, reduzindo a taxa de envio de pacotes da fonte. Embora não considerado neste trabalho, o módulo LC poderia opcionalmente auxiliar na realização dessa função.

5.1.2 Tráfego

A chegada de novos serviços das classes I e II são processos de Poisson mutuamente independentes com taxas $\lambda_{n,c}$ e $\lambda_{n,i}$. Os processos de chegada dos pedidos de *hand off* também são Poissonianos mutuamente independentes com taxas $\lambda_{h,c}$ e $\lambda_{h,i}$. Dessa forma, o tráfego oferecido dessas classes também são processos de Poisson mutuamente independentes com taxas $\lambda_c = \lambda_{n,c} + \lambda_{h,c}$ e $\lambda_i = \lambda_{n,i} + \lambda_{h,i}$.

Os tempos de residência e duração de uma chamada da classe I são v.a. distribuídas exponencialmente com parâmetros $1/\mu_{h,c}$ e $1/\mu_{d,c}$, respectivamente. Igualmente para os serviços da classe II, porém, com parâmetros $1/\mu_{h,i}$ e $1/\mu_{d,i}$. Os tempos de retenção de canal para ambos serviços são assim v.a. exponencialmente distribuídas com parâmetros $1/\mu_c = 1/(\mu_{h,c} + \mu_{d,c})$ e $1/\mu_i = 1/(\mu_{h,i} + \mu_{d,i})$.

O modelo de tráfego de Internet usado é o mesmo descrito no capítulo anterior. Porém, vale as seguintes mudanças de notação: os tempo entre chegada de pacotes IP (λ_{packet})

agora é λ_{IP}; o tempo de duração de uma sessão de dados, $(1/\mu_{d,(E)GPRS})$ é $1/\mu_{d,i}$; e o tempo de serviço de um pacote IP $(\frac{1}{\mu_{service}})$ é agora dado por $\frac{1}{\mu_s}$

5.1.3 Esquema sem Adaptação de Largura de Banda (SA)

Nesse esquema, um serviço da classe I solicitará sempre a largura de banda máxima, bw_{max}. Se esse recurso estiver disponível o serviço será aceito. Caso contrário, bloqueado. Uma cadeia de Markov multidimensional a tempo contínuo com estado dado pela Eq. (5.1) é usada para modelar esse esquema.

$$E = \{(c, k, m, r)/0 \leq c \leq \varrho, 0 \leq k \leq B_s, 0 \leq m \leq S, 0 \leq r \leq m\} \tag{5.1}$$

onde c é o número de chamadas da classe I em serviço. $\varrho = \lceil \frac{N}{bw_{max}} \rceil$ é o número máximo de chamadas pertencentes a essa classe aceitos pelo sistema. $\lceil x \rceil$ denota o menor número inteiro $\leq x$. k é o número de pacotes IP *buffer*. Como no capítulo anterior, m é o número de sessões de dados ativas. r representa o número de sessões de dados no estado OFF. A Tabela 5.1 são mostradas as possíveis transições da cadeia de Markov a partir de cada estado $E = (c, k, m, r)$, juntamente com as condições, as taxas e os eventos que ocasionam as transições dos estados.

As mudanças na variável c são determinadas pelas chegadas e partidas das chamadas da classe I. Os demais comportamentos seguem os mesmos princípios estipulados anteriormente.

A probabilidade de bloqueio de um serviço da classe I é dada pela Eq.(5.2).

$$P_{BC} = \sum_{\forall i \in E; c=\varrho} \pi_i \tag{5.2}$$

onde $\{\pi_{c,k,m,r}/(c, k, m, r) \in E\}$ é a probabilidade do estado de equilíbrio da cadeia de Markov.

A probabilidade de bloqueio de uma sessão de *Web* é dada pela Eq.(5.3). É importante mencionar que essa medida depende unicamente da especificação do número máximo de sessões concorrentes adotado pela Operadora de Serviço. Em outras palavras, ela independe do esquema de alocação de recursos empregado.

$$P_{BS} = \sum_{\forall j \in E; m=S} \pi_j \tag{5.3}$$

O tráfego oferecido IP é:

$$O = \lambda_{IP} \sum_{\forall i \in E} (m - r)\pi_i. \tag{5.4}$$

Tabela 5.1: Transições a partir do estado $E = (c, k, m, r)$.

Estado Sucessor	Condição	Taxa	Evento
$(c{+}1,k,m,r)$	$N{-}cbw_{max}{\geq}bw_{max}$	λ_c	Chegada de uma chamada da Classe I
$(c{-}1,k,m,r)$	$c{>}0$	$c\mu_c$	Partida de uma chamada da Classe I
$(c,k,m{+}1,r)$	$m{<}S$	$\frac{\beta}{\alpha+\beta}\lambda_i$	Chegada de uma sessão *Web* no estado ON
$(c,k,m{+}1,r{+}1)$	$m{<}S$	$\frac{\alpha}{\alpha+\beta}\lambda_i$	Chegada de uma sessão *Web* no estado OFF
$(c,k,m{-}1,r)$	$(m{>}0){\wedge}(r{=}0)$	$m\mu_i$	Partida de uma sessão *Web*
$(c,k,m{-}1,r{-}1)$	$(m{>}0){\wedge}(r{=}m)$	$m\mu_i$	
$(c,k,m{-}1,r{-}1)$	$(m{>}0){\wedge}(0{<}r{<}m)$	$\frac{r}{m}m\mu_i$	
$(c,k,m{-}1,r)$	$(m{>}0){\wedge}(0{<}r{<}m)$	$\frac{m-r}{m}m\mu_i$	
$(c,k{+}1,m,r)$	$(k{<}B_s){\wedge}(m{>}0){\wedge}(r{<}m)$	$(m{-}r)\lambda_{IP}$	Chegada de um pacote IP
$(c,k{-}1,m,r)$	$(min(\theta,k){>}0){\wedge}(k{>}0)$ $\theta{=}N{-}cbw_{max}$	$min(\theta,k)\mu_s$	Transmissão de um pacote IP
$(c,k,m,r{+}1)$	$r{<}m$	$(m{-}r)\alpha$	Diminuição da rajada
$(c,k,m,r{-}1)$	$r{>}0$	$r\beta$	Aumento da rajada

A probabilidade de bloqueio do pacote IP é

$$P_{BIP} = \sum_{\forall i \in E; k = B_s} \pi_i \tag{5.5}$$

A vazão média de um pacote IP é dada pela Eq.(5.6), enquanto que, o atraso médio é dado pela Eq.(5.7). A utilização do canal de rádio é dado pela Eq.(5.8)

$$X = O(1 - P_{BIP}). \tag{5.6}$$

$$W_q = \frac{\sum_{\forall i \in E} k\pi_i}{X}. \tag{5.7}$$

$$U_{SA} = bw_{max}\frac{\sum_{\forall i \in E} c\pi_i}{N} + \frac{\sum_{\forall i \in E; k>0} min(N - cbw_{max}, k)\pi_i}{N}. \tag{5.8}$$

5.1.4 Esquema com Adaptação de Largura de Banda (CA)

O estado da cadeia de Markov a tempo contínuo usada para representar esse esquema é dado por:

$$E = \{(c, \omega, k, m, r)/0 \leq c \leq \varrho, 0 \leq \omega \leq \vartheta, 0 \leq k \leq B_s, 0 \leq m \leq S, 0 \leq r \leq m\} \quad (5.9)$$

onde ω é o número de serviços da classe I com largura de banda mínima em serviço. $\vartheta = \lceil \frac{N - \varrho bw_{max}}{bw_{min}} \rceil$ é o número máximo de serviços com largura de banda mínima admitido pela rede. As outras variáveis de estado são as mesmas descritas anteriormente. A Tabela 5.2 mostra somente as transições, condições, taxas e eventos do estado $E = (c, \omega, k, m, r)$ correspondente à nova variável da cadeia.

A mudanças na variável ω são determinadas pelas chegadas, admissão e partida das chamadas da classe I com largura de banda mínima. A partida de uma chamada com largura de banda máxima, c, promove a largura de banda de uma chamada da classe I em serviço de bw_{min} para bw_{max} se $(\omega > 0)$. Novamente um pacote IP será transmitido usando toda a largura de banda disponível.

Tabela 5.2: Transições a partir do estado $E = (c, \omega, k, m, r)$.

Estado sucessor	Condição	Taxa	Evento
$(c,\omega+1,k,m,r)$	$(N-cbw_{max}<bw_{max})\wedge$ $[N-(cbw_{max}+\omega bw_{min})]\geq bw_{min}$	λ_c	Chegada de uma chamada da classe I com largura de banda mínima
$(c-1,\omega,k,m,r)$	$(c>0)\wedge(\omega=0)$	$c\mu_c$	Partida de uma chamada da classe I sem promoção
$(c,\omega-1,k,m,r)$	$(c>0)\wedge(\omega>0)$	$c\mu_c$	Partida de uma chamada da classe I com promoção
$(c,\omega-1,k,m,r)$	$\omega>0$	$\omega\mu_c$	Partida de uma chamada da classe I com menor largura de banda
$(c,\omega,k-1,m,r)$	$(min(\theta,k)>0)\wedge(k>0)$ $\theta=N-(cbw_{max}+\omega bw_{min})$	$min(\theta,k)\mu_s$	Transmissão de um pacote IP

Nesse esquema, o bloqueio acontecerá sempre que não houver disponibilidade de recurso para acomodar um serviço da classe I com banda mínima. Assim,

$$P_{BC} = \sum_{\forall i \in E; \omega = \vartheta} \pi_i \quad (5.10)$$

Todas as medidas de desempenho relativas ao serviço de dados são as mesmas do esquema sem adaptação. Porém, o estado a ser comsiderado é dado pela Eq.5.9). A utilização é dada pela Eq.(5.11).

$$U_{CA} = \frac{\sum_{\forall i \in E} \{cbw_{max} + \omega bw_{min}\}\pi_i}{N} + \frac{\sum_{\forall i \in ; k>0} min(N - (cbw_{max} + \omega bw_{min}), k)\pi_i}{N}. \quad (5.11)$$

5.1.5 Esquema com Adaptação de Largura de Banda Justa (AJ)

Esse esquema é muito similar ao anterior, contudo, quando a rede está congestionada, ele reduz a largura de banda de todas as chamadas em serviço e admite um novo cliente com largura de banda mínima. Além disso, quando um cliente com largura de banda mínima deixa o sistema, o LC verifica a carga da rede e informa ao CAC se é possível promover a largura de banda de todas as chamadas para largura de banda máxima.

O estado da cadeia de Markov a tempo contínuo usada para representar esse esquema é dado por:

$$E = \{(c, \omega, k, m, r)/0 \leq c \leq \varrho, 0 \leq \omega \leq \zeta, 0 \leq k \leq B_s, 0 \leq m \leq S, 0 \leq r \leq m\} \quad (5.12)$$

onde ω é o número de serviços da classe I com largura de banda mínima em serviço. $\zeta = \lceil \frac{N}{bw_{min}} \rceil$ é o número de chamadas de largura de banda mínima permitido pela rede. As demais variáveis de estados são as mesmas. Tabela 5.3 mostra as transições, condições, taxas e eventos tomados pelo CAC. Para melhor ilustrar o procedimento de promoção e redução de largura de banda são mostrados ainda os estados atual e sucessor.

A mudanças na variável c são determinadas pelas chegadas, redução da largura de banda e partida de uma chamada da classe I com largura de banda máxima. A redução da largura de banda ocorre no AJ quando não existem recursos de rádio para escoar uma chamada com banda máxima. Nesse caso, todas as chamadas com bw_{max} serão reduzidas para bw_{min} e uma nova chamada da classe I será admitida. Isso corresponde à uma mudança das variáveis c e ω para 0 e $c + 1$, respectivamente.

A mudanças na variável ω são determinadas pelas chegada e partida das chamadas da classe I com largura de banda mínima. Essa partida pode, se existirem recursos disponíveis, ocasionar a promoção da largura de banda de todas as chamadas da classe I em serviço, incorrendo na seguinte mudança de variável: $c : 0 \rightarrow \omega - 1$; e $\omega : \omega \rightarrow 0$. Novamente, um pacote IP será transmitido usando toda a largura de banda disponível.

Tabela 5.3: Transições da Cadeia de Markov do AJ.

Estado presente	Estado sucessor	Condição	Taxa	Evento
$(c,0,k,m,r)$	$(c+1,0,k,m,r)$	$(N-cbw_{max} \geq bw_{max})$	λ_c	Chegada de uma chamada da Classe I sem redução
$(c,0,k,m,r)$	$(0,c+1,k,m,r)$	$(N-cbw_{max} < bw_{max}) \wedge (N-cbw_{min}) \geq bw_{min}$	λ_c	Chegada de uma chamada da Classe I com redução
$(0,\omega,k,m,r)$	$(0,\omega+1,k,m,r)$	$(N-\omega bw_{min} \geq bw_{min})$	λ_c	Chegada de uma chamada da Classe I de menor largura de banda
$(c,0,k,m,r)$	$(c-1,0,k,m,r)$	$(c>0)$	$c\mu_c$	Partida de uma chamada da Classe I
$(0,\omega,k,m,r)$	$(0,\omega-1,k,m,r)$	$(\omega>0) \wedge N-(\omega-1)bw_{max}<0$	$\omega\mu_c$	Partida de uma chamada da Classe I sem promoção
$(0,\omega,k,m,r)$	$(\omega-1,0,k,m,r)$	$(\omega>1) \wedge N-(\omega-1)bw_{max} \geq 0$	$\omega\mu_c$	Partida de uma chamada da Classe I com promoção
$(c,0,k,m,r)$	$(c,0,k-1,m,r)$	$(min(\theta,k)>0) \wedge (k>0)$ $\theta=N-cbw_{max}$	$min(\theta,k)\mu_s$	Transmissão de um pacote IP
$(0,\omega,k,m,r)$	$(0,\omega,k-1,m,r)$	$(min(\theta,k)>0) \wedge (k>0)$ $\theta=N-\omega bw_{min}$	$min(\theta,k)\mu_s$	
$(0,0,k,m,r)$	$(0,0,k-1,m,r)$	$(min(N,k)>0) \wedge (k>0)$	$min(N,k)\mu_s$	

Nesse esquema, uma chamada da classe I será bloqueada se após a redução de largura de banda de bw_{max} para bw_{min}, o LC informar ao CAC que não existe recursos para admitir uma nova chamada. Assim:

$$P_{BC} = \sum_{\forall i \in E; c=0, \omega=\zeta} \pi_i \tag{5.13}$$

Novamente, todas as medidas de desempenho relativas ao serviço de dados são as mesmas do esquema sem adaptação. Porém, o estado a ser comsiderado é dado pela Eq.(5.12). A utilização é dada pela Eq.(5.14).

$$U_{AJ} = bw_{max}\frac{\sum_{\forall i\in E;c>0,\omega=0} c\pi_i}{N} + bw_{min}\frac{\sum_{\forall i\in E;c=0,\omega>0} \omega\pi_i}{N} + \frac{\sum_{\forall i\in E;c>0,\omega=0,k>0} min(N-cbw_{max},k)\pi_i}{N} + \frac{\sum_{\forall i\in E;c=0,\omega>0,k>0} min(N-\omega bw_{min},k)\pi_i}{N} + \frac{\sum_{\forall i\in E;c=\omega=0,k>0} min(N,k)\pi_i}{N}. \quad (5.14)$$

5.2 Resultados

Na Tabela 5.4 são mostrados os valores usados para obtenção dos resultados que serão apresentados a seguir. Um dado esquema de alocação de recurso adaptativo será referenciado como: A-bw_{max},bw_{min}; onde A pode ser CA ou AJ. No caso do esquema SA, será somente especificado a largura de banda máxima.

Tabela 5.4: Parâmetros usados nos experimentos.

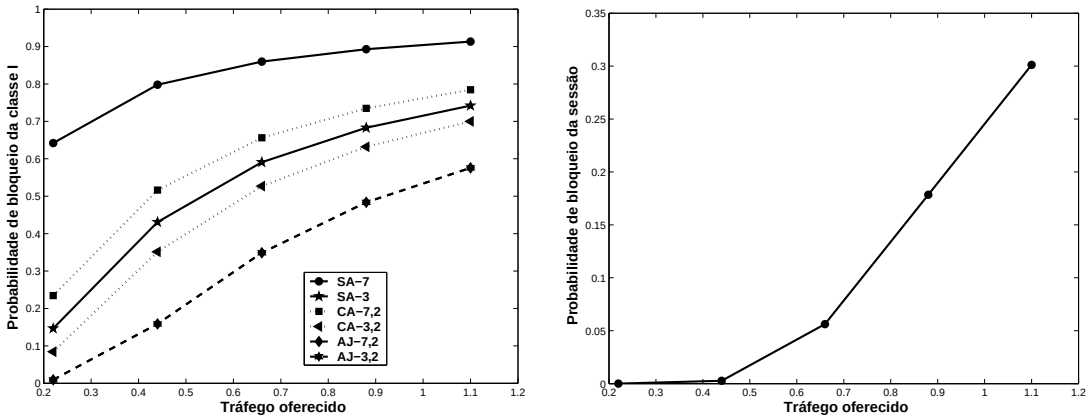
Parâmetros		**Valor**
Número de canais de rádio	N	20
Número de sessões *Web*	S	20
Tamanho do *buffer*	B_s	50
Tempo médio de duração de uma chamada da classe I(s)	$1/\mu_{d,c}$	120
Tempo médio de residência de uma chamada da classe I(s)	$1/\mu_{h,c}$	60
Tempo médio de residência de uma chamada da classe II(s)	$1/\mu_{h,i}$	120
Tempo médio de leitura (s)	D_{pc}	41,2
Tempo médio de serviço do pacote IP (s)	$1/\mu_s$	0,0375
Percentagem de *hand off* para chamadas das classes I e II (%)		10
Taxa média de *bits* da fonte (kbits/s)		32
Largura de banda máxima	bw_{max}	7,5,3
Largura de banda mínima	bw_{min}	3,2

5.2.1 Comparação entre os Esquemas

Como esperado, os desempenhos dos esquemas de alocação de recursos adaptativos são superiores ao do sem adaptação, Fig.(5.2). Além disso, dentre os esquemas adaptativos, o AJ possui um melhor desempenho. Nesse esquema, o bloqueio de uma chamada multimídia em tempo real da classe I somente ocorre quando não há possibilidade de admitir um cliente com banda mínima. Assim, quanto menor for essa especificação, maior será a probabilidade de uma chamada ser aceita. Como conseqüência, a probabilidade de bloqueio do AJ não depende da especificação da banda máxima, e sim, da mínima. Dessa forma, as probabilidades de bloqueio da classe I das configurações AJ-7,2 e AJ-3,2 são iguais.

Devido à capacidade de negociação de largura de banda durante a etapa de conexão de uma chamada a probabilidade de bloqueio da classe I do esquema CA é menor que do esquema SA.

A mesma figura mostra a probabilidade de bloqueio da sessão *Web*. Note que, com o aumento da carga, isto é, do número de sessões concorrentes em serviço, a probabilidade de bloqueio da sessão aumenta. Como mencionado anteriormente, a probabilidade independe do esquema de alocação de recurso usado. Portanto, esse comportamento não mudará no decorrer dos demais experimentos.



(a) (b)

Figura 5.2: Probabilidade de bloqueio: (a)Classe I, (b) Sessão

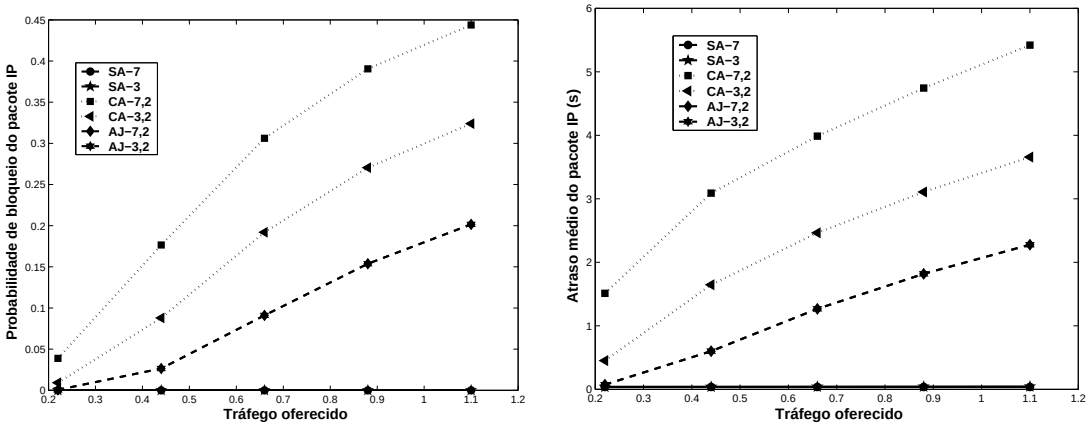
Nos esquemas SA sempre existem recursos ociosos, para os valores usados nos experimentos, o que garante o escoamento do tráfego IP de maneira que a sua probabilidade de bloqueio e atraso se mantenham sempre baixos, Fig(5.3). Por outro lado, nos esquemas CA, em um dado momento, todos os recursos são usados. Assim, a largura de banda disponível usada para escoar o tráfego IP decresce com o aumento do tráfego oferecido, degradando o

atendimento do serviço de dados. Em suma, a capacidade de negociação de recursos durante a etapa de conexão, fazendo com que a chamada multimídia se adapte às condições da rede causa uma degradação no desempenho dos serviços da classe II quando nenhum mecanismo de reserva de recursos é empregado.

No que tange o escoamento do tráfego IP, o desempenho do esquema CA-3,2 é superior ao do CA-7,2. Note na Fig(5.3) que o primeiro possui a probabilidade de bloqueio e o atraso médio do pacote IP menores que a do segundo. Isso ocorre, pois, durante uma sobrecarga, o CA-3,2 libera dois canais de rádio a uma taxa de $7\mu_c$, contra um canal, a uma taxa de $5\mu_c$ do CA-7,2.

Nos esquemas AJ, após a redução da largura de banda para a admissão de uma nova chamada, há uma grande disponibilidade de recursos para o escoamento do tráfego IP. Essa disponibilidade de recursos diminui com o aumento do tráfego, ocasionando um aumento na probabilidade de bloqueio e no atraso médio dos pacotes IP.

A Fig.(5.4) mostra que a utilização dos esquema SA é baixa quando comparada as utilizações dos esquemas adaptativos CA e AJ para uma carga de tráfego média e alta. O que já era esperado devido à ociosidade de canais de rádio. Para as configurações usadas, nota-se que as utilizações dos recursos dos esquemas CA são superiores as dos esquemas AJ. Isso ocorre, pois, uma vez que, uma chamada é admitida com largura de banda máxima, não há redução da mesma. Por sua vez, a utilizações dos esquemas AJ são, inicialmente, baixas. Entretanto, com o aumento do tráfego oferecido, e a admissão de mais chamadas, tem-se um aumento nessa utilização.



(a) (b)

Figura 5.3: Pacote IP: (a) Probabilidade de bloqueio e (b) Atraso.

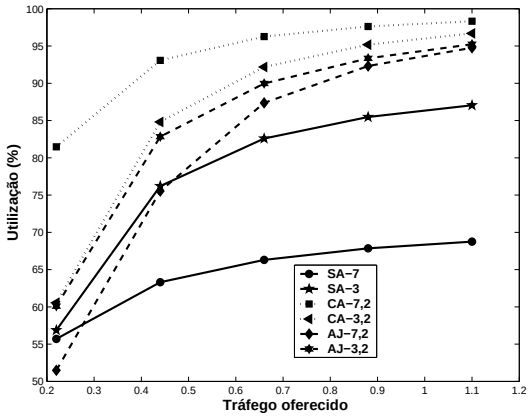


Figura 5.4: Utilização.

5.2.2 Multiplicidade entre os requerimentos de largura de banda e o número de canais

5.2.2.1 Esquema CA

A desvantagem do esquema CA ocorre quando a largura de banda máxima requerida por uma aplicação é múltipla do número de canais de rádio. Nesse caso, ele funcionará exatamente como o esquema SA, degradando o desempenho do sistema como um todo.

Isso pode ser evitado se a Operadora de Serviço durante a fase de desenvolvimento do seu *framework* de QoS, ou negociação do contrato de serviço, escolher apropriadamente os valores de largura de banda de modo a evitar esse problema .

5.2.2.2 Esquema AJ

Uma característica desejável em um ambiente multiserviço é que a provisão de garantias de QoS se estenda a todos os serviços de acordo com a suas características tal qual definido pelas classes de tráfego do 3GPP. De uma forma geral, sempre existirá um compromisso entre a provisão de QoS dos serviços das classes I e II, sendo que, a prioridade preemptiva naturalmente atribuída aos serviços da classe I pode exaurir os recursos de rádio, o que, certamente, ocasionará um aumento na probabilidade de bloqueio e no atraso médio dos pacotes IP. Porém, em células onde a demanda por serviços da classe II é grande, a mínima QoS deve ser garantida durante os momentos de congestionamento da rede. Uma forma tradicional de assegurar essa QoS é através da reserva de recursos. No esquema AJ, essa característica pode ser implementada implicitamente por meio da seleção apropriada dos valores de bw_{max} e bw_{min}. De uma maneira prática, a multiplicidade entre o número de canais e a largura de

banda máxima resulta em uma indisponibilidade de recursos antes da redução da banda. Da mesma forma, após essa redução, a multiplicidade entre e os recursos de rádio e a largura de banda mínima causa uma indisponibilidade de recursos.

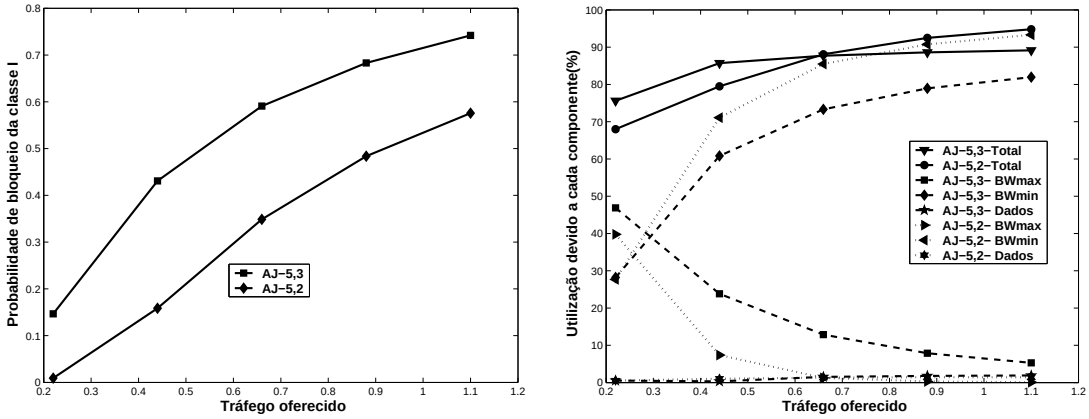
O efeito dessa multiplicidade no desempenho do esquema AJ será analisada nos experimentos subseqüentes para as configurações AJ-5,2 e AJ-5,3. Na primeira configuração, após a redução de largura de banda, a rede suporta no máximo dez chamadas concorrentes com banda mínima dois, o que ocupa todos os recursos de rádio da rede. Por outro lado, a outra configuração, AJ-5,3, suporta no máximo seis chamadas concorrentes com banda mínima três, restando dois canais para o escoamento do tráfego de dados. Como resultado direto dessa característica, nota-se na Fig.(5.5) que a configuração AJ-5,3 possui probabilidade de bloqueio do serviço da classe I maior.

Outra característica do esquema AJ é que para cada valor do tráfego de dados, existe uma porção de tempo que cada configuração gasta prestando o serviço com bandas máxima e mínima. A relação de tempo entre a prestação de serviço com bandas máxima e mínima tende a diminuir com o aumento do tráfego, o que é natural, visto que, quanto maior o tráfego, maior o número de chamadas com banda máxima em serviço, maior a taxa de consumo dos recursos e menor será o tempo de prestação de serviço com banda máxima.

Isso pode visualizado na Fig.(5.5), onde tem-se a utilização total de cada configuração, além da utilização devido a cada componente sua. Em outras palavras, a utilização devido ao escoamento do serviço com banda máxima, mínima e o escoamento do serviço da classe II. Nota-se para as duas configurações que, com o aumento do tráfego há predominância de uma componente em detrimento à outra. Assim, observa-se que com o aumento do tráfego a componente devido ao serviço com banda máxima diminui, liberando mais rapidamente os recursos para o escoamento do tráfego de dados, e a utilização devido a componente com banda mínima aumenta, sendo que esse efeito é proeminente na configuração AJ-5,2. Note ainda que, o tráfego IP compõe uma pequena fatia da utilização total.

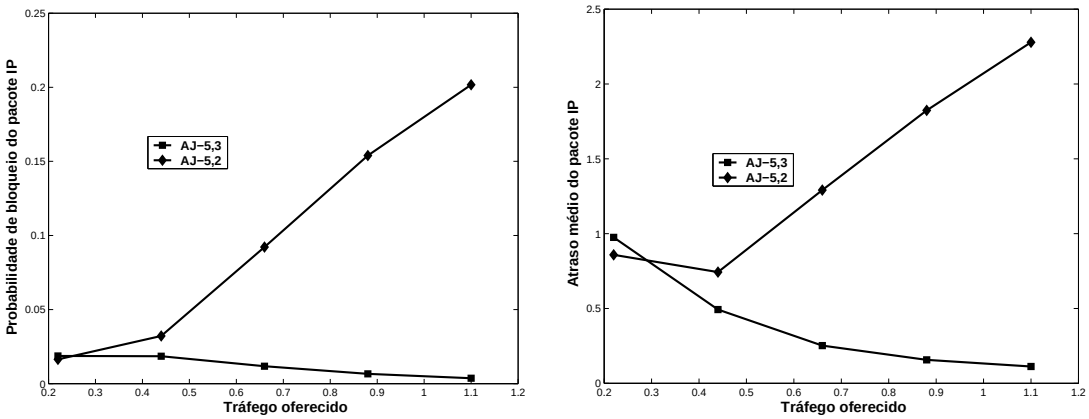
Com esses preceitos é possível explicar o porquê da diminuição do bloqueio e do atraso médio do pacote IP para a configuração AJ-5,3, e o perfil do atraso médio na configuração AJ-5,2, Fig.(5.6). No caso da configuração AJ-5,3 além da reserva de recursos implícita, ocasionada pela escolha do valor da largura de banda mínima, tem-se com o aumento do tráfego uma liberação mais rápida de recursos para o escoamento do tráfego da classe IP, diminuindo a sua probabilidade de bloqueio e o atraso médio com o aumento do tráfego. No caso da configuração AJ 5,2, observa-se que o atraso inicialmente decresce com o aumento do tráfego, e posteriormente, aumenta. Isso acontece pois, para uma carga baixa de tráfego há predominância do serviço da classe I com banda máxima, e por conseguinte,

uma indisponibilidade de recursos para o escoamento do tráfego IP, aumentando o atraso médio. Com o aumento do tráfego há predominância do serviço com banda mínima, disponibilizando mais rapidamente os recursos de rádio. Como, nesse caso, não há uma reserva de recursos implícita, toda a largura de banda disponível será consumida pelo serviço da classe I, aumentado a probabilidade de bloqueio e o atraso médio do pacote IP.



(a) (b)

Figura 5.5: (a) Probabilidade de bloqueio da classe I e (b) Utilização.



(a) (b)

Figura 5.6: Pacote IP: (a) Probabilidade de bloqueio e (b) Atraso.

Capítulo 6

Política ótima de alocação de recursos

No capítulo anterior foi mostrado que o mecanismo de adaptação de largura de banda desempenha um papel fundamental nos esquemas de alocação de recurso em redes de próxima geração. Sua atuação, balizada na promoção ou redução da largura de banda de uma chamada multimídia mediante as flutuações do tráfego, permite que a Operadora de Serviços atenda um número maior de clientes. Porém, essa operação acarreta um aumento na carga de sinalização consumindo recursos na rede sem fio e cabeada, tão bem como potência da bateria da estação móvel (35). Adicionalmente, freqüentes comutações de largura de banda podem ser piores que uma grande taxa de degradação. Assim, existe um compromisso entre os recursos da rede utilizado pelas chamadas, sinalização e o processamento incorrido pela operação de adaptação.

Outro aspecto concernente ao mecanismo de adaptação de largura de banda é a satisfação do cliente. Em momentos de sobrecarga, a Gerência dos Recursos de Rádio (*Radio Resource Management*- RRM) tende a reduzir a banda atribuída a cada cliente de forma a atender o maior número de assinantes. Porém, se somente a largura de banda mínima é disponível, alguns clientes da classe de QoS conversacional podem experimentar uma degradação na QoS. Note que usuários da classe *Streaming* podem lidar com essa quantidade de recurso disponibilizada pela rede usando técnicas de *buffering* em suas estações móveis de forma a neutralizar o *jitter* (4)(16).

Assim, é imperativo que a alocação de recursos em ambientes adaptativos seja feita de forma a estabelecer um compromisso entre a utilização dos canais, o controle na freqüência de comutação de largura de banda entre os diferentes níveis (freqüência de adaptação) e a satisfação dos usuários.

Neste capítulo essa questão é abordada pela construção de um modelo Semi-Markoviano de Decisão, o qual busca uma política estacionária ótima sob o critério de oti-

malidade do custo esperado médio a longo prazo. Com esse modelo é possível através de penalizações (custos) ponderar objetivos diferentes de forma a achar um política de custo mímino que satisfaça as condições dadas. A estrutura de custo proposta faz com que o modelo Semi-Markoviano de Decisão busque uma política estacionária ótima que pondere as probabilidade de bloqueio, a freqüência de adaptação de largura de banda e a satisfação do usuário.

6.1 Modelagem

No capítulo anterior constatou-se que o esquema justo obteve o melhor desempenho entre os analisados. Assim, ele foi escolhido para servir de base de comparação com o esquema de alocação de recursos adaptativo ótimo.

6.1.1 Modelo do esquema de alocação justa modificado

Por simplicidade, antes de apresentar o modelo com decisão, será introduzido o modelo justo descrito anteriormente com algumas simplificações no modelo de dados que foram feitas de forma a diminuir o tamanho do PSMD.

O novo modelo de dados agrega as mesmas características da fonte de tráfego proposta pelo 3GPP, contudo, diferente do modelo proposto em (41) onde as sessões de dados solicitavam a conexão, este, proposto em (59), já considera as sessões admitidas pelo sistema. O estado da cadeia de Markov a tempo Contínuo usado para representar esse esquema é dado por

$$E = \{(c, \omega, k, m)/0 \leq c \leq \varrho, 0 \leq \omega \leq \zeta, 0 \leq k \leq B_s, 0 \leq m \leq S\} \tag{6.1}$$

onde, assim como no capítulo anterior, c é o número de chamadas da classe I em serviço. $\varrho = \lceil \frac{N}{bw_{max}} \rceil$ é o número máximo de chamadas pertencentes a essa classe aceitos pelo sistema. ω é o número de serviços da classe I com largura de banda mínima em serviço. $\zeta = \lceil \frac{N}{bw_{min}} \rceil$ é o número de chamadas de largura de banda mínima permitido pela rede. k é o número de pacotes no *buffer* e m é o número de chamadas por pacotes ativas. A capacidade de armazenamento do *buffer* e o número máximo de sessões são dados por B_s e S. O esquema justo modificado é mostrado na Tabela (6.1). Note que esquema de adaptação de recursos permanece inalterado.

Novamente, a probabilidade de bloqueio de uma chamada multimídia em tempo

Tabela 6.1: Transições do modelo de alocação de recursos justa.

Estado corrente	Estado sucessor	Condição	Taxa	Evento
$(c,0,k,m)$	$(c+1,0,k,m)$	$(N-cbw_{max} \geq bw_{max})$	λ_c	Chegada de uma chamada em tempo real sem redução
$(c,0,k,m)$	$(0,c+1,k,m)$	$(N-cbw_{max} < bw_{max}) \wedge$ $(N-cbw_{min}) \geq bw_{min}$	λ_c	Chegada de uma chamada em tempo real com redução
$(0,\omega,k,m)$	$(0,\omega+1,k,m)$	$(N-\omega bw_{min} \geq bw_{min})$	λ_c	Chegada de uma chamada em tempo real e admissão com bw_{min}
$(c,0,k,m)$	$(c-1,0,k,m)$	$(c>0)$	$c\mu_c$	Partida de uma chamada em tempo real
$(0,\omega,k,m)$	$(0,\omega-1,k,m)$	$(\omega>0) \wedge$ $N-(\omega-1)bw_{max}<0$	$\omega\mu_c$	Partida de uma chamada em tempo real sem promoção
$(0,\omega,k,m)$	$(\omega-1,0,k,m)$	$(\omega>1) \wedge$ $N-(\omega-1)bw_{max} \geq 0$	$\omega\mu_c$	Partida de uma chamada em tempo real com promoção
(c,ω,k,m)	$(c,\omega,k,m+1)$	$m<S$	$(S-m)\beta$	Início de uma chamada por pacotes
(c,ω,k,m)	$(c,\omega,k,m-1)$	$m>0$	$m\alpha$	Término de uma chamada por pacotes
$(c,0,k,m)$	$(c,0,k+1,m)$	$m>0 \wedge k<B_s$	$m\lambda_{IP}$	Geração do pacote IP
$(c,0,k,m)$	$(c,0,k-1,m)$	$(min(\theta,k)>0) \wedge (k>0)$ $\theta=N-cbw_{max}$	$min(\theta,k)\mu_s$	Transmissão do pacote IP
$(0,\omega,k,m)$	$(0,\omega,k-1,m)$	$(min(\theta,k)>0) \wedge (k>0)$ $\theta=N-\omega bw_{min}$	$min(\theta,k)\mu_s$	
$(0,0,k,m)$	$(0,0,k-1,m)$	$(min(N,k)>0) \wedge (k>0)$	$min(N,k)\mu_s$	

real é dada por:

$$P_{BC} = \sum_{\forall i \in E; c=0, \omega=\zeta} \pi_i \tag{6.2}$$

De acordo com a referência (59) o tráfego oferecido IP é dado por:

$$O = \lambda_{IP} \frac{\beta}{\beta + \alpha} S \tag{6.3}$$

A probabilidade de bloqueio do pacote IP é

$$P_{BIP} = \sum_{\forall i \in E; k=B_s} \pi_i \tag{6.4}$$

Assim, a vazão média e o atraso médio são dados pelas Eq.(6.5) e Eq.(6.6).

$$X = O(1 - P_{BIP}). \tag{6.5}$$

$$W_q = \frac{\sum_{\forall i \in E} k\pi_i}{X}. \tag{6.6}$$

$$U_{AJ} = bw_{max}\frac{\sum_{\forall i\in E;c>0,\omega=0} c\pi_i}{N} + bw_{min}\frac{\sum_{\forall i\in E;c=0,\omega>0} \omega\pi_i}{N} + \frac{\sum_{\forall i\in E;c>0,\omega=0,k>0} min(N - cbw_{max}, k)\pi_i}{N} + \frac{\sum_{\forall i\in E;c=0,\omega>0,k>0} min(N - \omega bw_{min}, k)\pi_i}{N} + \frac{\sum_{\forall i\in E;c=0,\omega=0,k>0\in E} min(N, k)\pi_i}{N}. \quad (6.7)$$

6.1.2 Modelo Semi-Markoviano de Decisão

O estado do sistema é definido pelo conjunto de valores:

$$E = \{(c, b, ev, k, m)/ \quad 0 \leq c \leq \Theta, b \in \{0, 1\}, ev \in \{0, 1, 2\}, 0 \leq k \leq B_s, 0 \leq m \leq S; \quad (6.8)$$
$$\text{se b=0, } \Theta = \lceil \frac{N}{bw_{max}} \rceil$$
$$\text{se b=1, } \Theta = \lceil \frac{N}{bw_{min}} \rceil\}$$

onde c é o número de chamadas multimídia em tempo real pertencentes a Classe I (como definido no capítulo anterior) usando a largura de banda definida pelo valor de b. Isto é, se $b = 0$ todos os clientes da Classe I estão usando a largura de banda máxima, caso contrário, $b = 1$, todos estão usando banda mínima. Assim, o sistema pode admitir um número máximo de chamadas com qualidade máxima $\lceil \frac{N}{bw_{max}} \rceil$ e mínima $\lceil \frac{N}{bw_{min}} \rceil$. N é o número de canais de rádio disponível na célula para o escoamento do tráfego oferecido. ev é o último evento ocorrido no sistema. Para $ev = 0$ e 1 tem-se, respectivamente, a partida e chegada de uma chamada da classe I; para $ev = 2$ tem-se os demais eventos, isto é, a geração e transmissão de um pacote IP e ativação e desativação de uma sessão *Web*. Essa v.a. é definida no estado do sistema para estipular o conjunto de ações em cada estado. k é o número de pacotes IP no *buffer* e m é o número de sessões ativas no sistema. A capacidade de armazenamento do *buffer* e o número máximo de sessões são dados por B_s e S. Cada estado representa a configuração do sistema logo após a ocorrência de um evento e antes da tomada de decisão.

As épocas de decisão são os instantes de chegada e partida de uma chamada multimídia em tempo real. Quando o controle de admissão de chamadas é analisado isoladamente,

os instantes de partida são considerados fictícios, porém, quando analisado juntamente com a adaptação de largura de banda, os mesmos passam a ser reais.

Na chegada de uma chamada multimídia em tempo real, $(ev = 1)$, pode-se tomar as ações de rejeitá-la e não adaptar as chamadas em serviço (NN); de rejeitá-la e adaptar as chamadas em serviço (NA); de aceitá-la e não adaptar as chamadas em serviço (AN) e de aceitá-la e adaptar as chamadas em serviço (AA). Por outro lado, nos instantes de partida, $(ev = 0)$, pode-se tomar somente as ações relativas à adaptação de largura de banda das chamadas que permanecem em serviço, isto é, (NN) e (NA). Para os demais eventos no sistema, $(ev = 2)$, não se toma qualquer ação, (NN), isto é, "não faça nada !". Assim, para cada estado $i = (c, b, ev, k, m) \in E$ define-se o seguinte conjunto de ações:

$$A(i) = \begin{cases} 0 - NN, & \forall\ ev \in \{0,1,2\}. \\ 1 - NA, & ev = 0 \wedge b = 0; \\ & \vee ev = 0 \wedge b = 1 \wedge (c \leq \lceil \frac{N}{bw_{max}} \rceil); \\ & \vee ev = 1 \wedge (c \leq \lceil \frac{N}{bw_{max}} \rceil). \\ 2 - AN, & ev = 1 \wedge b = 0 \wedge (c < \lceil \frac{N}{bw_{max}} \rceil); \\ & \vee ev = 1 \wedge b = 1 \wedge (c < \lceil \frac{N}{bw_{min}} \rceil). \\ 3 - AA, & ev = 1 \wedge b = 0; \\ & \vee ev = 1 \wedge b = 1 \wedge (c < \lceil \frac{N}{bw_{max}} \rceil). \end{cases} \tag{6.9}$$

Note que cada ação corresponde a um determinado número $\in \{0, 1, 2, 3\}$, onde a ação NN corresponde ao número 0 e assim por diante. Esse artifício matemático é feito, pois, dessa forma a tomada de decisão relativa a aceitação de uma chamada acontece sempre que o quociente da divisão inteira entre $A(i)/2 = 1$, caso contrário a chamada é rejeitada. Da mesma forma, a tomada de decisão relativa a adaptação de largura de banda das chamadas em serviço é dada sempre que o resto da divisão inteira[1] $A(i)/2 = 1$, caso contrário, não há adaptação. Assim, o termo $ac = 1$ será usado sempre que a ação tomada aceitar uma nova chamada, e $ac = 0$, caso rejeite. Do mesmo modo, o termo $ad = 1$ será usado se a ação significar adaptação, caso contrário, $ad = 0$, quando não houver adaptação.

A ação NN pode ser tomada em qualquer instante de decisão. A ação NA pode ser tomada durante a partida sempre que as chamadas estiverem com banda máxima. Por outro lado, caso estejam com banda mínima, elas somente poderão ser promovidas se o sistema comportar as chamadas remanescentes com banda máxima, isto é, $(c \leq \lceil \frac{N}{bw_{max}} \rceil)$. Em instantes

[1]Essa operação é normalmente chamada de MOD.

de chegada, essa ação poderá ser tomada sempre que o sistema estiver com todas as chamadas com banda máxima. Caso contrário, se elas estiverem com banda mínima, a ação poderá ser tomada somente se o sistema puder acomodar as chamadas em serviço com banda máxima, ou seja, $(c \leq \lceil \frac{N}{bw_{max}} \rceil)$.

As ações AN e AA somente podem ser tomadas em instantes de chegada de uma chamada multimídia em tempo real. A primeira poderá ser escolhida sempre que o sistema poder aceitar chamadas com banda máxima,$(c < \lceil \frac{N}{bw_{max}} \rceil)$, ou mínima, $(c < \lceil \frac{N}{bw_{min}} \rceil)$. A última poderá ser tomada sempre que as chamadas estejam com banda máxima. Veja que é sempre possível aceitar e adaptar as chamadas de banda máxima para banda mínima. Por outro lado, se as chamadas estiverem com banda mínima, a ação AA somente poderá ser escolhida se o sistema puder comportar essas chamadas e mais a nova com banda máxima, isto é, $(c < \lceil \frac{N}{bw_{max}} \rceil)$.

Para esse PSMD dado que o sistema está no estado $i \in E$ e a ação $a \in A(i)$ é tomada, tem-se:

- O tempo esperado até o próximo instante de decisão, $\tau_i(a)$;
- A probabilidade de que no próximo instante de decisão o sistema esteja em $j \in E$, $p_{ij}(a)$;
- O custo esperado incorrido até o próximo instante de decisão, $C_i(a)$.

A partir das taxas de transição de cada estado, $\Lambda_{ij}(a)$, obtêm-se facilmente a taxa total de saída de cada estado dada por $\Lambda_i(a) = \sum_{j \neq i} \Lambda_{ij}(a)$. Onde $\Lambda_i(a)$ é o parâmetro da distribuição exponencial negativa que descreve o tempo de permanência tempo no estado i quando a ação a é tomada. Assim, o tempo entre as transições e as probabilidades de transição são dadas, respectivamente, por

$$\begin{aligned} \tau_i(a) &= 1/\Lambda_i(a), \\ p_{ij}(a) &= \frac{\Lambda_{ij}(a)}{\Lambda_i(a)} \end{aligned} \tag{6.10}$$

Abaixo se apresenta o algoritmo gerador de transições do modelo de decisão. Nele, se o estado do sistema é $i = (c, b, ev, k, m) \in E$, então, denomina-se $i.n$ a coordenada correspondente ao número de chamadas multimídia, e assim sucessivamente. Denominação análoga para os estados r e $t \in E$. Define-se como o *estado reagido a ação* ($r \in E$) como aquele observado logo após a tomada da ação. O sistema permanecerá nele até que o próximo evento ocorra.

1. INÍCIO
2. **Para cada estado e ação** $i \in E$ e $a \in A(i)$ **faça**
3. $r \leftarrow i$;
4. **se** $(ac == 1)$ **faça** $\{r.c \leftarrow r.c + 1\}$;
5. **se** $(ad == 1)$ **faça** $\{r.b \leftarrow 1 - r.b\}$;
6. Por *default* **faça** $r.ev \leftarrow 2$;
7. *// Chegada de uma chamada multimídia em tempo real*
8. $t \leftarrow r$;
9. $t.ev \leftarrow 1$;
10. **Faça a transição de** $i \rightarrow t$ **com probabilidade** $\frac{\lambda_c(a)}{\Lambda_i(a)}$;
11. *// Partida de uma chamada multimídia em tempo real*
12. **se** $(r.c > 0)$ **faça** $\{$;
13. $t \leftarrow r$;
14. $t.c \leftarrow t.c - 1$;
15. $t.ev \leftarrow 0$;
16. **Faça a transição de** $i \rightarrow t$ **com probabilidade** $\frac{r.c\mu_c}{\Lambda_i(a)}\}$;
17. *// Início de uma chamada por pacotes*
18. **se** $(r.m < S)$ **faça** $\{$;
19. $t \leftarrow r$;
20. $t.m \leftarrow t.m + 1$;
21. **Faça a transição de** $i \rightarrow t$ **com probabilidade** $\frac{(S - r.m)\beta}{\Lambda_i(a)}\}$;
22. *// Término de uma chamada por pacotes*
23. **se** $(r.m > 0)$ **faça** $\{$;
24. $t \leftarrow r$;

25. $t.m \leftarrow t.m - 1$;

26. **Faça a transição de** $i \rightarrow t$ **com probabilidade** $\frac{r.m\alpha}{\Lambda_i(a)}$};

27. *// Geração do pacote IP*

28. **se** $(r.k < B_s \wedge r.m > 0)$ **faça** {;

29. $t \leftarrow r$;

30. $t.k \leftarrow t.k + 1$;

31. **Faça a transição de** $i \rightarrow t$ **com probabilidade** $\frac{r.m\lambda_{IP}}{\Lambda_i(a)}$};

32. *// Transmissão do pacote IP*

33. **se** $(r.b == 0)$ **faça** $\{band \leftarrow bw_{max}\}$;

34. **Senão se** $(r.b == 1)$ **faça** $\{band \leftarrow bw_{min}\}$;

35. $trans \leftarrow \min(N - r.cband, r.k)$;

36. **se** $(trans > 0)$ **faça** {;

37. $t \leftarrow r$;

38. $t.k \leftarrow t.k - 1$;

39. **Faça a transição de** $i \rightarrow t$ **com probabilidade** $\frac{trans\mu_s}{\Lambda_i(a)}$};

40. FIM

Diferente dos demais trabalhos na literatura, o modelo de decisão proposto busca uma política estacionária ótima de alocação de recursos que minimize a probabilidade de bloqueio das chamadas multimídia em tempo real , controle a freqüência de adaptação e a insatisfação do usuário perante a sobrecarga. Para tal, usou-se uma estrutura de custo com a seguinte característica:

$$C_i(a) = C_B(i,a) + C_{AD}(i,a) + C_H(i,a) \tag{6.11}$$

onde $C_B(i,a)$, $C_{AD}(i,a)$ e $C_H(i,a)$ são, respectivamente, os custos imediatos de bloqueio, adaptação e permanência em uma dada banda. Dado o número de chamadas em serviço imediatamente antes da tomada de decisão como n_{ad}, o estado do sistema após a tomada da ação $a \in A(i)$ como $i = (c, b, ev, k, m) \in E$, as expressão para esses custos são:

$$
\begin{aligned}
C_B(i,a) &= c_b, \text{ se } ev = 1 \wedge a = ac = 0 \\
C_{AD}(i,a) &= c_a n_{ad}, \text{ se } ev \in \{0,1\} \wedge a = ad = 1 \\
C_H(i,a) &= \begin{cases} C_{max} = c_{max} n\tau_i(a), \text{ se } b = 0 \\ C_{min} = c_{min} n\tau_i(a), \text{ se } b = 1 \end{cases}
\end{aligned}
$$

Note que os custos relativos ao bloqueio e a adaptação de largura de banda são fixos, enquanto que, o custo de permanência é função do tempo no qual o sistema encontra-se servindo as chamadas com banda máxima ou mínima. Além disso, os custos de adaptação e permanência são funções da quantidade de chamadas presentes no sistema antes e depois da ação, respectivamente.

Depois de definido todos os elementos pertinentes ao modelo do PSMD, pode-se utilizar o algoritmo de iteração de valores juntamente com o método de uniformização apresentado no capítulo 2 para encontrar a política ótima.

6.1.2.1 Medidas de Desempenho

Uma vez que a política estacionária ótima é achada usando o AIV, resolve-se a cadeia de Markov a tempo contínuo, e então com a distribuição de equilíbrio extrai-se as medidas de desempenho.

O tráfego escoado no estado $i \in E$ dado que $ev = 1$ e a ação $a = ac = 1 \in A(i)$ foi escolhida é

$$
T_{ESC} = \sum_{\forall i \in E, ev=1, a=ac=1 \in A(i)} \Lambda_i(a)\pi_i. \tag{6.12}
$$

Assim, a probabilidade de bloqueio de uma chamada multimídia em tempo real é

$$
P_{BC} = 1 - \frac{T_{ESC}}{\lambda_c(a)}. \tag{6.13}
$$

Como no pseudo código para a geração das transições, considere a variável $band = bw_{max}$ se $(b = 0)$ ou $band = bw_{min}$ se $(b = 1)$. Considere ainda a variável $trans = \min(N - cband, k)$. De posse dessas variáveis, deriva-se a utilização dos recursos de rádio como:

$$
\begin{aligned}
U_{OT} &= bw_{max} \frac{\sum_{\forall i \in E, a \in A(i); c>0, b=0} c\pi_i}{N} + bw_{min} \frac{\sum_{\forall i \in E, a \in A(i); c>0, b=1} c\pi_i}{N} \\
&\quad + \frac{\sum_{\forall i \in E, a \in A(i)} trans\pi_i}{N}.
\end{aligned} \tag{6.14}
$$

As medidas de desempenho do serviço em dados são as mesmas mostradas anterirormente.

6.2 Resultados

Na Tabela 6.2 são mostrados os valores usados para obtenção dos resultados que serão apresentados a seguir. O esquema de alocação de recursos adaptativo ótimo será referenciado como (OT), enquanto que, o justo será novamente chamado de (AJ). Como os requerimentos de QoS são os mesmos, isto é, bw_{max} e bw_{min}, eles não serão colocados na referência ao esquema usado. A taxa de chegada das chamadas multimídia será aumentada de 0,011 até 0,55 chamadas/s para que se estude o comportamento da política ótima mediante ao aumento do tráfego. Uma fonte de 8 kbits/s será considerada na análise do serviço de dados.

A capacidade de armazenamento do *buffer* foi reduzida em relação à usada no capítulo anterior, visto que, o objetivo é analisar a política ótima para a classe de maior prioridade de serviço. A Tabela 6.3 mostra algumas medidas que justificam essa escolha. Observe que o custo médio a longo prazo por unidade de tempo permanece praticamente inalterado com o aumento do *buffer*. Além disso, o número médio de pacotes no *buffer* não cresce significativamente. Por outro lado, como mostram as últimas três colunas dessa tabela, o PSMD aumenta consideravelmente com o aumento da capacidade do *buffer*. Outro detalhe que será mostrado posteriormente é que política ótima não depende do comportamento do serviço de dados.

6.2.1 Análise da política ótima

Nesta seção será estudada a política ótima para os valores considerados na Tabela 6.2 e para a faixa de variação dos valores de tráfego. A Tabela 6.4 mostra o comportamento assumido pela política ótima, isto é, ação tomada para cada estado do processo quando o último evento é uma chegada ou partida e todos as chamadas em serviço estão com largura de banda máxima. Note que a política ótima é gulosa, pois, ela aceita a nova chamada sempre que existem recursos para tal. Além disso, ela nunca adapta a largura de banda das chamadas em serviço para a banda mínima. Isso garante a satisfação do cliente durante os momentos de congestionamento. Na partida nunca há a adaptação da largura de banda.

A Tabela 6.5 mostra o comportamento tomado pela política ótima quando o último evento é uma chegada e todos as chamadas em serviço estão com largura de banda mínima. A análise será realizada para as faixas de tráfego baixa ($0,011 \rightarrow 0,088$ chamadas/s), média

Tabela 6.2: Parâmetros usados nos experimentos.

Parâmetros		Valor
Número de canais de rádio	N	20
Número de sessões *Web*	S	5
Tamanho do *buffer*	B_s	20
Tempo médio de duração de uma chamada da classe I(s)	$1/\mu_{d,c}$	120
Tempo médio de residência de uma chamada da classe I(s)	$1/\mu_{h,c}$	60
Tempo médio de leitura (s)	D_{pc}	41.2
Tempo médio de serviço do pacote IP (s)	$1/\mu_s$	0.0375
Percentagem de *hand off* para chamadas multimídia (%)		10
Taxa média de *bit* da fonte (kbits/s)		8
Largura de banda máxima	bw_{max}	2
Largura de banda mínima	bw_{min}	1
Custo de bloqueio	c_b	100
Custo de adaptação	c_{ad}	5
Custo de banda máxima	c_{max}	1
Custo de banda mínima	c_{min}	2
Número de chamadas com banda máxima	$\lceil \frac{N}{bw_{max}} \rceil$	10
Número de chamadas com banda mínima	$\lceil \frac{N}{bw_{min}} \rceil$	20

$(0,11 \rightarrow 0,176$ chamadas/s) e alta $(0,22 \rightarrow 0,55$ chamadas/s). É interessante observar que para cada uma das cargas de tráfego a política ótima apresenta um comportamento diferente. Assim:

- Para uma carga de tráfego baixa, ela aceita e adapta para banda máxima sempre que possível ($c \leq 9$). Além disso, quando não é possível mais comportar novas chamadas com banda máxima, ela passa a aceitar todas as chamadas com banda mínima ($10 \leq c \leq 19$), a não ser no limite ($c = \lceil \frac{N}{bw_{min}} \rceil$) quando a ação (NN) é a única a ser tomada;

- Para a segunda faixa de tráfego considerada a política ótima continua aceitando e adaptando para banda máxima sempre que possível ($c \leq 9$), porém, no caso limite da quantidade de chamadas com banda máxima, ($c = \lceil \frac{N}{bw_{max}} \rceil$), ela prefere recusar uma nova chamada com banda mínima e adaptar todas as chamadas em serviço para a banda máxima. Uma outra característica dessa política é que ela cria uma *zona de rejeição* com o aumento do tráfego, isto é, ela deixa de aceitar novas chamadas mesmo quando existem recursos disponíveis de modo a melhorar o desempenho do sistema a longo prazo. Assim, pode-se observar que para o valor de tráfego $0,11$ chamadas/s a ação (NN) é tomada sempre que o estado $c = \lceil \frac{N}{bw_{max}} \rceil + 1$ é observado. Com o aumento do tráfego essa ação começa, então, a ser tomada nos estados $c = \lceil \frac{N}{bw_{max}} \rceil + i, i = 2, ..., \lceil \frac{N}{bw_{min}} \rceil - 2$

Tabela 6.3: Impacto do aumento da capacidade do *buffer* no PSMD.

B_s	$C_i(a)$	Número médio de pacotes	Número de estados	Número de pares estado-ação	Número de transições não nulas
20	$4,46128$	$0,156302$	11.844	23.688	125.876
30	$4,46128$	$0,167639$	17.484	34.968	186.776
50	$4,46128$	$0,177431$	28.764	57.528	308.576
100	$4,46129$	$0,182155$	56.964	113.928	613.076

Tabela 6.4: Comportamento da política com banda máxima.

Estado	Ação
$(c \leq 9, b = 0, ev = 1, k, m)$	AN
$(c = 10, b = 0, ev = 1, k, m)$	NN
$(0 \leq c \leq 9, b = 0, ev = 0, k, m)$	NN

de modo que o sistema rejeite sempre uma nova chamada quando todas as chamadas em serviços estão com banda mínina. Observe que no estado $c = \lceil \frac{N}{bw_{min}} \rceil$ a única ação possível é NN;

- Esse efeito é proeminente para altos valores de tráfego $\lambda_c = 0,22 \rightarrow 0,55$ chamadas/s quando o sistema sempre recusa uma chamada com largura de banda mínima em prol do benefício da satisfação do cliente.

A Tabela 6.6 mostra o comportamento da política ótima quando o último evento é uma partida e todos as chamadas em serviço estão com largura de banda mínima. Esse comportamento é o mesmo para todas as faixas de tráfego citadas anteriormente. Observa-se que o sistema adapta as chamadas remanescentes no sistema sempre que possível de forma a aumentar a satifação do usuário.

Tabela 6.5: Comportamento da política com banda mínima.

Tráfego	Estado	Ação	Tráfego	Estado	Ação
0,011→0,088	(c≤9,b=1,ev=1,k,m)	AA	0,154	(c≤9,b=1,ev=1,k,m)	AA
	(10≤c≤19,b=1,ev=1,k,m)	AN		(c=10,b=1,ev=1,k,m)	NA
	(c=20,b=1,ev=1,k,m)	NN		(11≤c≤16,b=1,ev=1,k,m)	NN
				(17≤c≤19,b=1,ev=1,k,m)	AN
				(c=20,b=1,ev=1,k,m)	NN
0,11	(c≤9,b=1,ev=1,k,m)	AA	0,165	(c≤9,b=1,ev=1,k,m)	AA
	(c=10,b=1,ev=1,k,m)	NA		(c=10,b=1,ev=1,k,m)	NA
	(c=11,b=1,ev=1,k,m)	NN		(11≤c≤17,b=1,ev=1,k,m)	NN
	(12≤c≤19,b=1,ev=1,k,m)	AN		(18≤c≤19,b=1,ev=1,k,m)	AN
	(c=20,b=1,ev=1,k,m)	NN		(c=20,b=1,ev=1,k,m)	NN
0,121	(c≤9,b=1,ev=1,k,m)	AA	0,176	(c≤9,b=1,ev=1,k,m)	AA
	(c=10,b=1,ev=1,k,m)	NA		(c=10,b=1,ev=1,k,m)	NA
	(11≤c≤12,b=1,ev=1,k,m)	NN		(11≤c≤18,b=1,ev=1,k,m)	NN
	(13≤c≤19,b=1,ev=1,k,m)	AN		(c=19,b=1,ev=1,k,m)	AN
	(c=20,b=1,ev=1,k,m)	NN		(c=20,b=1,ev=1,k,m)	NN
0,132	(c≤9,b=1,ev=1,k,m)	AA	0,22→0,55	(c≤9,b=1,ev=1,k,m)	AA
	(c=10,b=1,ev=1,k,m)	NA		(c=10,b=1,ev=1,k,m)	NA
	(11≤c≤13,b=1,ev=1,k,m)	NN		(11≤c≤20,b=1,ev=1,k,m)	NN
	(14≤c≤19,b=1,ev=1,k,m)	AN			
	(c=20,b=1,ev=1,k,m)	NN			
0,143	(c≤9,b=1,ev=1,k,m)	AA			
	(c=10,b=1,ev=1,k,m)	NA			
	(11≤c≤15,b=1,ev=1,k,m)	NN			
	(16≤c≤19,b=1,ev=1,k,m)	AN			
	(c=20,b=1,ev=1,k,m)	NN			

6.2.2 Análise comparativa entre a política ótima e o esquema de alocação adaptativa justa

Nesta seção o desempenho da política de alocação ótima será comparado com o do esquema justo. É importante ressaltar que o objetivo da política ótima é minimizar a probabilidade de bloqueio, a freqüência de adaptação e a insatisfação do usuário.

A Fig.(6.1.a) mostra o custo médio a longo prazo para os dois esquemas de alocação de recurso. Note que o esquema justo onera mais o sistema que a política ótima. A Fig.(6.1.b) destaca a redução no custo médio a longo prazo, calculado como $\frac{C_i^{AJ} - C_i(a)}{C_i^{AJ}} \times 100$, obtido pela política ótima. Note que ela chega a alcançar 22,59 % durante momentos de congestionamento. A Fig.(6.1.c) mostra uma comparação entre os custos médio de adaptação para ambos

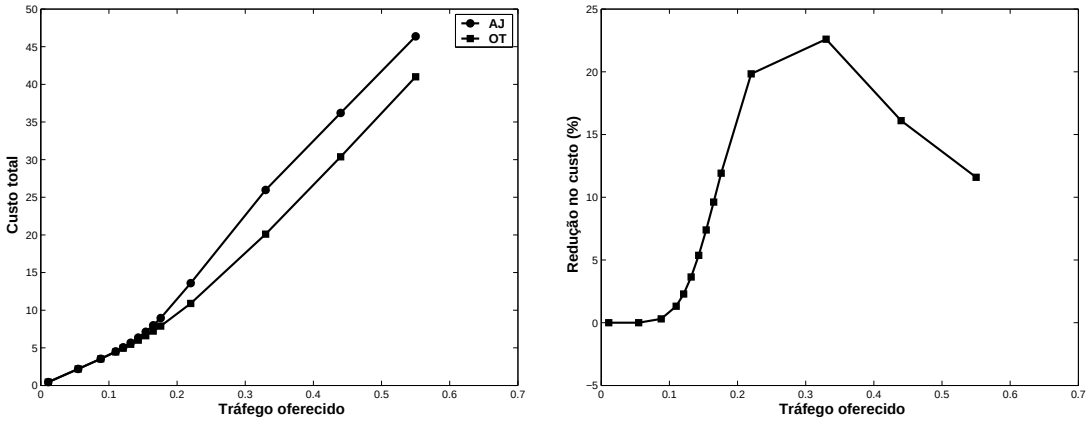
Tabela 6.6: Comportamento da política ótima quando o último evento é uma partida e os clientes estão na banda mínima.

Estado	Ação
$(1 \leq c \leq 10, b = 1, ev = 0, k, m)$	NA
$(c = 0$ e $11 \leq c \leq 19, b = 1, ev = 0, k, m)$	NN

esquemas. Novamente é possível observar a superioridade da política ótima sobre o esquema justo principalmente para médio e alto tráfego.

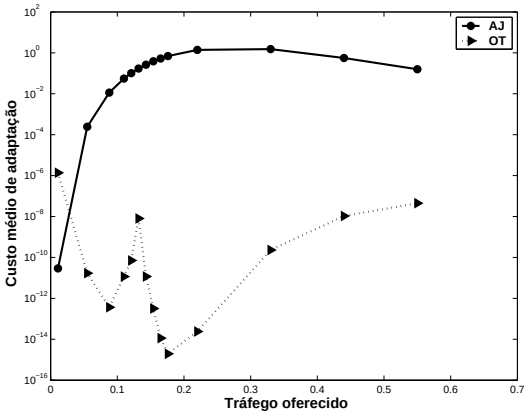
A Fig.(6.2.a) mostra que a probabilidade de bloqueio do esquema justo é inferior a da política ótima. Isso acontece, pois, embora a política ótima minimize a probabilidade de bloqueio, ela tenta a todo momento satisfazer o usuário. Isso significa atribuir e manter a largura de banda máxima mesmo durante os momentos de congestionamento. Assim, com um número maior de chamadas com a banda máxima, menor é a disponibilidade de recursos para as novas solicitações por serviço e, conseqüentemente, maior a probabilidade de bloqueio. Por outro lado, a Fig.(6.2.b) revela que a utilização dos recursos de rádio do esquema que emprega a política ótima é superior a do justo. A Fig.(6.2.c) ratifica essa informação destacando a diferença percentual entre essas utilizações. Nela se observa que o esquema justo consegue manter a mesma utilização para tráfego baixo, e uma superficial superioridade para uma parte do tráfego médio. Porém, com o aumento do tráfego, a política ótima mostra-se novamente superior chegando a melhorar a utilização dos recursos de rádio em quase 13 %. A Fig.(6.2.d) mostra que a política ótima mantém a largura de banda máxima durante todo o experimento, e como dito anteriormente, a maior utilização, enquanto que, o esquema justo somente segue o desempenho da política ótima sob uma carga de tráfego média. Para valores de tráfego alto, a maior parte das chamadas é servida com banda mínima.

A Fig.(6.3) mostra o desempenho do serviço de dados sob a política ótima e justa. Note que o preço a ser pago pela satisfação dos clientes mantendo a banda máxima sempre que possível, também se reflete no desempenho desse serviço. Assim, o esquema justo consegue escoar mais rapidamente o tráfego *Web*. Porém nota-se que o durante tráfego alto a política ótima não chega a bloquear 10% do tráfego oferecido *Web*.



(a)

(b)



(c)

Figura 6.1: Custos médios: (a) Total (b) Redução no Custo médio a longo prazo por unidade de tempo e (c) Custo médio de adaptação.

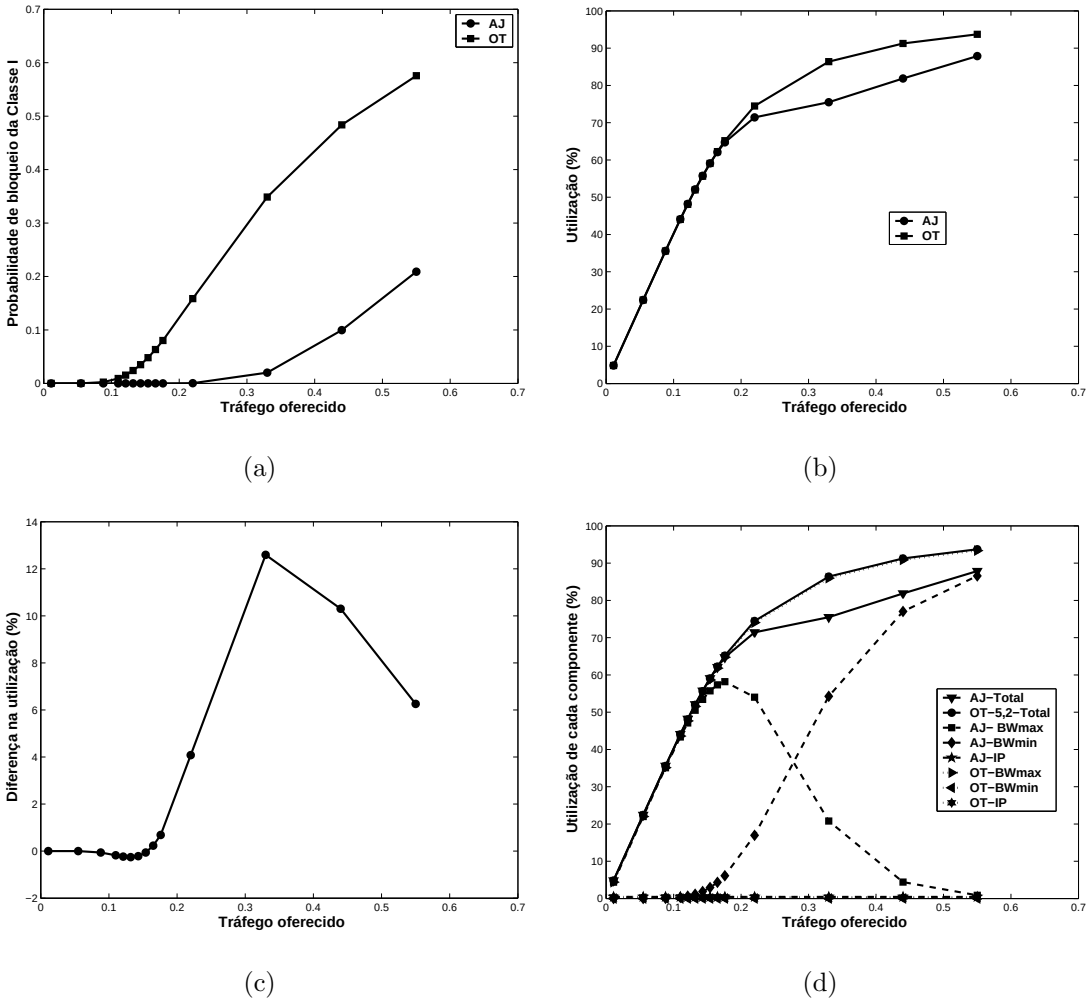
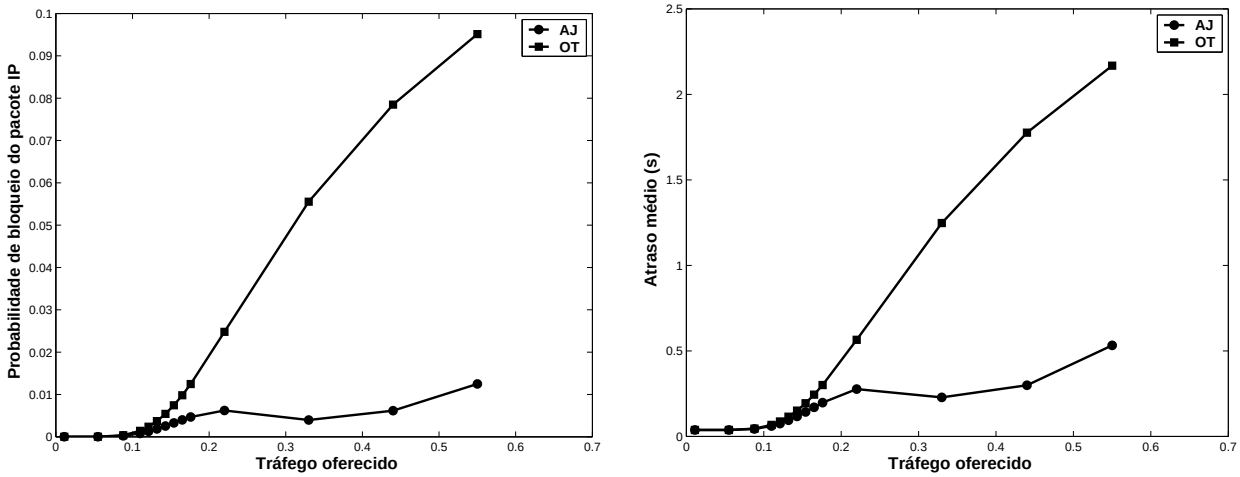


Figura 6.2: (a) Probabilidade de bloqueio de uma chamada multimídia em tempo real (b) Utilização (%), (c) Melhora na Utilização (%) e (d) Utilização devido a cada componente (%).



(a) (b)

Figura 6.3: Pacote IP:(a) Probabilidade de bloqueio e (b) Atraso médio.

Conclusão

Foi proposto neste trabalho um modelo denominado de *Statecharts/Markov*, o qual dentro das etapas do processo de modelagem destaca o uso de uma especificação formal para a descrição do comportamento do sistema.

Através dessa abordagem se obtém uma especificação clara e consistente do comportamento do sistema. É importante citar que o uso dessa metodologia torna o processo de verificação, validação e documentação mais consistente.

Um dos aspectos positivos do *Statecharts* do ponto de vista da modelagem é o seu apelo visual que permite uma especificação melhor do comportamento do sistema devido às suas peculiaridades na representação de atividades paralelas e da hierarquia.

A flexibilidade da técnica foi empregada na modelagem de três esquemas de alocação de recursos. Os modelos analíticos apresentados foram parametrizados de acordo com a rede GSM/GPRS. Além disso, os seus resultados foram comparados com resultados obtidos através de simulação. Os resultados mostraram a importância da reserva de recursos na manutenção da QoS do serviço de menor prioridade. Nesse sentido, deve-se salientar que há um aumento natural do bloqueio do serviço de maior prioridade.

É importante ressaltar que o enfileiramento de chamadas de voz consegue suprir a ausência da prioridade preemptiva desde que o tempo médio de serviço de dados seja curto.

O desenvolvimento de um ambiente computacional chamado de *PerformCharts* empregando a metodologia discutida no capítulo 3 é um pesquisa corrente dentro do grupo de desempenho de redes no Programa de Pós-Graduação em Engenharia Elétrica da UFPA. Através desse ambiente é possível modelar o sistema usando o *Statecharts*. Através da especificação, pode-se gerar, então, uma solução analítica ou de simulação. Nesse sentido, espera-se que futuros trabalhos já possam utilizar essa ferramenta para a modelagem, verificação, validação e documentação de modelos de desempenho.

No capítulo 4 foi proposto e analisado um novo esquema de alocação de recursos em redes móveis celulares hierárquicas. Seu desempenho foi investigado usando uma rede de

duas camadas GSM/(E)GPRS. Os resultados mostraram um desempenho superior do esquema proposto em relação ao sistema com prioridade de voz. Um ponto importante desse esquema é que ele não impacta na QoS do serviço de voz. Essa característica é muito importante já que esse serviço é preponderante na rede, e assim, a garantia do seu provimento deve ser a máxima possível. Um parâmetro importante no desempenho da rede é o valor do *threshold*, uma vez que, sua escolha permite balancear a carga de tráfego de dados entre as células das diversas camadas da rede.

Outros estudos indicaram que o fator de atividade da fonte de dados é mais um elemento responsável pela degradação da QoS do serviço de dados. Por fim, como esperado, verificou-se que o desempenho do (E)GPRS é superior ao desempenho do GPRS na microcélula e relativamente inferior na macrocélula pois foi usado um esquema de codificação de canal como menor vazão para o EGPRS. Isso aconteceu porque na geração do modelo analítico é necessário fixar uma modulação e um esquema de codificação de canal. Contudo, em um sistema real, as variações do canal possibilitariam mudanças na modulação e no esquema de codificação usado. Isso, na média resultaria em um desempenho superior do EGPRS quando comparado ao GPRS.

Embora tenham sido estudadas somente as configurações macro/microcélula EGPRS ou GPRS, outras combinações podem ser usadas como microcélulas EGPRS e macrocélula GPRS e vice-versa. Observe que a escolha da configuração apropriada deve refletir a demanda por serviços de dados e a viabilidade econômica.

O modelo proposto melhora a utilização do espectro de freqüência, visto que, o tráfego de dados utiliza os recursos de rádio ociosos das células em todas as camadas. Essa eficiência pode ser ainda melhorada considerando no modelo o procedimento de retorno. Assim, futuras pesquisas podem incluir esse tema. Uma outra questão que pode ser explorada é a modelagem do tráfego de roteamento das sessões de dados. Por simplicidade, foi considerado que esse tráfego assume um comportamento Poissoniano, porém, um estudo mais aprofundado pode ser feito para identificar qual é o verdadeiro comportamento desse tráfego.

Outra sugestão para trabalhos futuros é o estudo de políticas de admissão ótimas em ambientes hierárquicos usando um PSMD. Como primeiro passo, poder-se-ia investigar um valor de *threshold* ótimo.

No capítulo 5 foram propostos modelos analíticos Markovianos para a análise de desempenho de esquemas de alocação de recursos onde adaptação de largura de banda é empregada junto com o CAC de modo a melhorar a provisão da QoS. O emprego de esquemas adaptativos exerce um papel fundamental em redes móveis celulares de próxima geração onde a prestação de serviços multimídia em tempo real e Internet deverá ser provida com garantias

de QoS. Nesse ambiente multiserviço será imperativo que a utilização dos recursos de rádio seja eficiente, de maneira que, ambos serviços sejam atendidos satisfatoriamente.

De maneira geral, com relação aos esquemas investigados neste trabalho, observou-se que o emprego do esquema justo resultou em uma melhor provisão de serviço com garantias de QoS.

Um outro ponto importante concerne à multiplicidade entre o número de canais de rádio e as especificações de largura de banda. No esquema com adaptação de banda (CA), uma multiplicidade entre o número de canais de rádio e o valor da largura de banda máxima torna o seu comportamento igual ao de um esquema sem adaptação, o que não é interessante do ponto de vista da Operadora de Serviço. Isso pode ser evitado se a Operadora de Serviço durante a fase de desenvolvimento do seu *framework* de QoS, ou negociação do contrato de serviço, escolher apropriadamente os valores de largura de banda

No esquema justo, essa multiplicidade pode ser feita de forma a estabelecer um compromisso entre o desempenho dos serviço das classes de maior e menor prioridade. Assim, através de uma seleção apropriada dos valores das larguras de bandas máxima e mínima, pode ser implementada uma reserva de recursos implícita, de maneira que, durante um congestionamento a mínima QoS da classe de menor prioridade seja garantida. Contudo, isso será conseguido em detrimento ao aumento da probabilidade de bloqueio e a diminuição da utilização do serviço de maior prioridade.

O desempenho do esquema justo é beneficiado pela redução total da largura de banda máxima de todas as chamadas da classe I em serviço. Contudo, esse comportamento pode causar uma certa insatisfação por parte dos assinantes que não desejam reduzir sua largura de banda em detrimento à admissão de outros. Por sua vez, o desempenho do CA é beneficiado pela negociação inicial de largura de banda entre a rede e a aplicação. Porém, uma vez saturada a admissão de clientes com banda mínima, a rejeição de chamadas tende a aumentar. Uma forma de melhorar o desempenho do sistema é a implementação de um esquema adaptativo híbrido que combine as características do AJ e do CA. Assim, por exemplo, enquanto houvesse largura de banda disponível o sistema se comportaria como o CA. Durante o congestionamento, ela poderia se comportar como o justo. Esse tema fica sugerido para trabalhos futuros.

No capítulo 6 foi estuda uma política ótima para uma esquema de alocação de recursos que busca minimizar a probabilidade de bloqueio, freqüência de adaptação e a insatisfação do usuário.

Para os valores considerados nos experimentos, verificou-se que, a política ótima apresenta um comportamento guloso no que diz respeito a aceitação de chamadas multimídia

em tempo real, pois, ela tenta manter a satisfação do cliente a todo momento. Por outro lado, quando as chamadas em serviço estão todas com banda mínima, o seu comportamento muda de acordo com a demanda por serviço tornando-a difícil de implementar na prática.

Os resultados mostraram que o desempenho do esquema justo consegue seguir a política ótima durante tráfego baixo e uma parte do tráfego médio. Sob tráfego alto, esse esquema reduz a largura de banda das chamadas em serviço de modo a aceitar novas chamadas. Esse fato pode, como já mencionado, causar uma insatisfação por parte dos usuários de serviços como vídeo conferência, videofone, etc. Assim, em células onde a demanda por serviço se encaixa nesse perfil de tráfego no qual o justo atua satisfatoriamente, pode-se usá-lo para a prestação de serviço.

Outro fato explorado pela política ótima é a freqüência de adaptação. Estudos recentes mostraram que ela consome muitos recursos da rede através da sinalização necessária para a sua manutenção. Além disso, a estação móvel também é penalizada pelo consumo extra de potência de sua bateria. Através dos resultados do custo médio de adaptação, observou-se que a política ótima consegue reduzir bastante a freqüência de comutação em os níveis de largura de banda.

Com sugestões para trabalhos futuros nessa linha de pesquisa pode-se a *priori*, considerar o fluxo de *hand off* diferente daquele referente as novas chamadas de voz. Veja que isso pode ser feito facilmente a partir dos modelos adaptativos e do modelo da política ótima. Neste trabalho, esses fluxos foram imersos no mesmo processo de Poisson, uma vez que, inicialmente a demanda por serviços que requerem uma grande quantidade de recursos tende a ser baixa, ou pelos menos, com um perfil de mobilidade quase estacionário. Porém, futuras aplicações multimídia estarão disponíveis mesmo em cenários de alta mobilidade. Assim, a separação entre esses fluxos deve ser feita. Neste caso, cada fluxo pode funcionar separadamente como um esquema justo dentro da sua classe de serviço.

Outro aspecto interessante, ainda nessa linha de pesquisa, é o estudo de políticas ótimas nesse cenário descrito. Recentemente, alguns autores buscaram soluções usando reforço de aprendizado através de uma rede neural. Essa solução se mostrou bastante promissora, uma vez, que não é necessário o conhecimento das probabilidades de transição. Além disso, o problema da dimensionalidade do PSMD também é resolvido. Assim, futuros trabalhos podem incluir essa aproximação como a solução do PSMD.

Outra sugestão para trabalhos futuros é o estudo de políticas de alocação de recursos adaptativas sub ótimas de fácil implementação. Nesse sentido, o busca-se através da combinação de mecanismos como múltiplos *thresholds*, prioridades, etc, comportamentos que possuam o seu desempenho próximos ao da política ótima.

Como visto nos resultados do capítulo 6, o comportamento da política ótima é independente do serviço de dados. Assim, sugere-se ainda que o *buffer* seja retirado do estudo da política ótima, uma vez que, seu impacto no tamanho do sistema é considerável. Sua inclusão pode ser posteriormente feita para quantificar o desempenho do serviço de dados após o estudo e implementação de políticas sub-ótimas.

Sugere-se ainda para trabalhos futuros o efeito da elasticidade das chamadas de dados como feito em (32). Essas chamadas melhoraram o desempenho do serviço de dados durante momentos de congestionamento.

De forma consensual, redes de 3G e 4G são baseadas na tecnologia CDMA. Assim, a modelagem de mecanismos de alocação de recursos visando soluções analíticas para essas redes podem ser desenvolvidos usando as ferramentas apresentadas neste trabalho. Vale novamente ressaltar que, embora já existam modelos para redes WCDMA, UTRA-TDD, CDMA2000 em sua grande maioria, eles utilizam uma solução via simulação.

Referências Bibliográficas

[1] M. Hännikäinen, T. D. Hämäläinen, M. Niemi e J. Saarinen. Trends in personal wireless data communications, Computer Communications, vol. 25, 2002, pp. 84–99.

[2] X. Wu, B.Mukherjee e D. Ghosal. Hierarchical architectures in the third-generation-cellular Network, Wireless communication, vol. 11, no. 3, 2004, pp. 62–71.

[3] K. Yeo e C.H. Jun. Modeling and analysis of hierarchical cellular networks with general distributions of call and cell residence times, IEEE Transactions on Vehicular Technology, vol. 51, no.6, Novembro, 2002, pp.1361 - 1374.

[4] T. Janevisk. Traffic analysis and design of wireless IP network, Artech House, 2003.

[5] R. Parry. Overlooking 3G, IEEE Potentials, vol. 21 , no. 4,Outubro-Novembro, 2002, pp.6–9.

[6] M. D. Yacoub. Wireless Technology: Protocols, Standards, and Techniques, CRC Press, 2002.

[7] http://www.gsmworld.com, outubro de 2004.

[8] http://www.3gamericas.org, outubro de 2004.

[9] C. Bettstetter, H.J. Vogel e J. Eberspacher. GSM Phase 2+, General Packet Radio Service (GPRS): architecture, protocols and air interface, IEEE Communications Surveys, vol.2, no.3, 1999, pp. 2–14,

[10] R. Kalden, I. Meirick e M. Meyer. Wireless Internet Access Based on GPRS, IEEE Personal Communications, vol. 7, no. 2, Abril, 2000, pp. 8-18.

[11] E. Seurre, P. Savelli e P. J. Pietri, EDGE for Mobile Internet, Artech House, 2003.

[12] A. Furuskar, S. Mazur, F. Muller e H. Olofsson. EDGE: enhanced data rates for GSM and TDMA/136 evolution , IEEE Personal Communications, vol. 6, no. 3, Junio, 1999, pp. 56–66

[13] D. Molkdar, W. Featherstone e S. Larnbotharan. An overview of EGPRS: the packet data component of EDGE. Electronics & Communication Engineering Journal, vol. 14, no. 1, Fevereiro, 2002, pp. 21–38.

[14] M. Zeng, A. Annamalai e V. K. Bhargava. Harmonization of Global Third-Generation Mobile Systems, IEEE Communications Magazine, vol. 38, no. 12, Dezembro 2000, pp. 94–104.

[15] J. Korhonen. Introduction to 3G mobile communications, Artech House, 2003.

[16] N. Baghaei e R. Hunt. Review of quality of service performance in wireless LANs and 3G multimedia application services, Computer Communications, vol. 27, no. 17, November, 2004, pp. 1684-1692.

[17] 3GPP TS 23.107. Techinical Specification Group Services and System Aspects; QoS Concept and Architecture (Release 5), V4.4.0.

[18] S.Dixit, Y. Guo e Z. Antoniou. Resource management and quality of service in third generation wireless networks. IEEE Communications Magazine, vol. 39, no. 2, Fevereiro, 2001, pp. 125–133.

[19] http://www.cdg.org/index.asp, outubro de 2004.

[20] J.-Z. Sun, J. Sauvola e D. Howie. Features in future: 4G visions from a technical perspective. IEEE Global Telecommunications Conference (GLOBECOM01), vol. 6, Novembro, 2001 , pp. 3533–3537.

[21] A. Young Kim e E. C. Kim. Key concept of radio access for the 4G mobile communication systems, The 6th International Conference on Advanced Communication Technology, vol. 1, 2004, pp. 245–248

[22] J. Y. Kim e E. C. Kim. Wired and wireless network integration for the 4G mobile communication systems, The 6th International Conference on Advanced communication Technology, vol. 1, 2004, pp. 249–252.

[23] I. Katzela e M. Naghshineh. Channel assignment schemes for cellular mobile telecommunication systems: A comprehensive survey. IEEE Personal Communications, vol. 2, no. 3, 1996, pp. 11–31.

[24] A. Hac e A. Armstrong. Resource allocation scheme for QoS provisioning in microcellular networks carrying multimedia traffic. Int. J. Netw. Manag. (John Wiley & Sons, Inc.), vol.11, no. 5, 2001, pp. 277–307.

[25] J.Kim e A. Jamalipour. Traffic management and QoS provisioning in future wireless IP networks, IEEE Personal Communications, vol. 8, no. 5, 2001, pp. 46–55.

[26] Z. Sahinoglu e S. Tekinay. On multimedia networks: self-similar traffic and network performance, IEEE Communications Magazine, vol. 37, no. 1, 1999, pp. 48–52.

[27] S. Shakkottai e R. Srikant. Scheduling real-time traffic with deadlines over a wireless channel, Wirel. Netw. (Kluwer Academic Publishers), vol. 8, no. 1, 2002, pp. 13–26.

[28] L. Huang, S. Kumar e C.C. Jay Kuo. Adaptive resource allocation for multimedia QoS management in wireless networks, IEEE Transactions on Vehicular Technology, vol. 53, no. 2, 2004, pp. 547 - 558.

[29] M. Ermel, T. Müller, J. Schüler, M. Schweigel e K. Begain. Performance of GSM networks with general packet radio services, Perform. Eval., vol. 48, no. 1–4, 2002, pp. 285–310.

[30] N.M. Mitrou, G. L. Lyberopoulos e A.D. Panagopoulou. Voice and data integration in the air-interface of a microcellular mobile communication system, IEEE Transactions on Vehicular Technology, vol. 42, no. 1, 1993, pp. 1–13.

[31] M. Naghshineh e A. S. Acampora. QoS provisioning in micro-cellular networks supporting multiple classes of traffic, Wirel. Netw. (Kluwer Academic Publishers), vol. 2, no. 3, 1996, pp. 195–203.

[32] B. Li, L. Li, Bo Li, K.M. Sivalingam e Xi-Ren Cao, Call admission control for voice/data integrated cellular networks: performance analysis and comparative study, IEEE Journal on Selected Areas in Communications, vol. 22, no. 4, 2004, pp. 706 - 718.

[33] D.-H. Shih e Z.-Q. Lin. Bandwidth saturation QoS provisioning for adaptive multimedia in wireless/mobile networks, Computer Standards & Interfaces, vol. 26, no. 4, 2004, pp. 279–288.

[34] C. Oliveira, J. B. Kim, T. Suda. An adaptive bandwidth reservation scheme for high-speed multimedia wireless networks, IEEE Journal on Selected Areas in Communications, vol. 16, no. 6, 1998, pp. 858–874.

[35] F. Yu, V. W.S Wong e V. C.M. Leung. A new QoS provisioning method for adaptive multimedia in cellula wireless networks, *in Proceedings* IEEE INFOCOM 2004, 2004.

[36] G. Min e M. Ould-Khaoua. Message latency in hypercubic computer networks with the bursty traffic pattern, Computers and Electrical Engineering, vol. 30, no. 3, 2004, pp. 207-222.

[37] A. Klemm, C. Lindemann e M. Lohmann. Modeling IP traffic using the batch Markovian arrival process, Perform. Eval., vol. 54, no. 2, 2003, pp. 149–173.

[38] P. Salvador, A. Pacheco e R. Valadas. Modeling IP traffic: joint characterization of packet arrivals and packet sizes using BMAPs, Comput. Networks, vol. 44, no. 3, 2004, pp. 335–352.

[39] R. Fantacci. Performance evaluation of prioritized handoff schemes in mobile cellular networks, IEEE Transactions on Vehicular Technology, vol. 49, no. 2, 2000, pp. 485–493.

[40] A. S. Alfa e W. Li. A homogeneous PCS network with markov call arrival process and phase type cell residence time, Wirel. Netw.(Kluwer Academic Publishers), vol. 8, no. 6, 2002, pp. 597–605.

[41] C. Lindemann e A. Thümmler. Performance analysis of the general packet radio service, Comput. Networks, vol. 41, no. 1, 2003, pp. 1–17.

[42] H.-W. Ferng e Y.-C. Tsai. Channel allocation and performance study for the integrated GSM/GPRS system, Wireless Communications and Networking 2003 (WCNC 2003), vol. 3, 2003, pp. 1861 -1865.

[43] M. Mahdavi e R. Tafazolli. Analysis of integrated voice and data for GPRS, International Conference on 3G Mobile Communication Technologies (IEE Conf. Publ. No. 471), 2000, pp. 436–440.

[44] P. Lin e Y.B.Lin. Channel Allocation for GPRS, IEEE Transaction on Vehicular Technology, vol.50, no. 2, 2001, pp. 375-387.

[45] X. Fang e D.Ghosal. Performance modeling and QoS evaluation of MAC/RLC layer in GSM/GPRS networks, IEEE International Conference on Communications (ICC03), vol. 1, 2003, pp. 271–275.

[46] W.Y. Chen, J.L.C.Wu e H.H.Liu. Performance Analysis of Radio Resource Allocation in GSM/GPRS Networks, Vehicular Technology Conference (VTC2002), vol.3, 2002, pp. 1461–1465.

[47] W.Y.Chen, J.L.C. Wu e L.L.Lu. Performance Comparision of Dynamic Resource Allocation With/Without Channel De-Alloccation in GSM/GPRS Networks. IEEE Communication Letters, vol.7, no. 1, Janeiro, 2003, 10–12.

[48] C. H. Foh, B. Meini, B.Wydrowski e M.Zukerman. Modeling and performance evaluation of GPRS, Vehicular Technology Conference, 2001 (VTC2001), vol. 3, 2001, pp. 2108–2112.

[49] J. M. Wing. A specifier's introduction to formal methods, Computer, vol. 23, Setembro, 1990, pp. 8, 10 -22, 24.

[50] G. Tremblay. Formal methods: mathematics, computer science or software engineering?, IEEE Transactions on Education, vol. 43, no. 4, novembro, 2000, pp. 377–382.

[51] P.E. Black, K.M. Hall, M.D. Jones, T.N. Larson e P.J.Windley. A brief introduction to formal methods, IEEE Custom Integrated Circuits Conference, 1996.

[52] J.P. Bowen e M.G. Hinchey. Ten commandments of formal methods, Computer, vol. 28, no. 4, Abril, 1995, pp. 56–63.

[53] D. Blyth, C. Boldyreff, C. Ruggles e N. Tetteh-Lartey. The case for formal methods in standards, IEEE Software, vol. 7, no. 5 , Setembro, 1990, pp. 65–67.

[54] K.J. Turner. The use of formal methods in communications standards, IEE Colloquium on Formal Methods for Protocols, 1991.

[55] N. L. Vijaykumar. Statecharts: Their use in specifying and dealing with Performance Models, Tese de Doutorado, Instituto Tecnológico de Aeronáutica (ITA), 1999, São José Dos Campos,Brasil.

[56] C. R. L. Francês. Statecharts Estocásticos e Queuing Statecharts - Novas Abordagens para Avaliação de Desempenho Baseadas em Especificação Statecharts, Tese de Doutorado, Universidade de São Paulo (USP), 2001, São Paulo, Brasil.

[57] G. Bolch, S. Greiner, H. Meer e K. S. Trivedi. Queuing Networks and Markov Chains: modeling and peformance evaluation with computer science applications, John Wiley & Sons, 1998.

[58] D. Harel. Statecharts: A visual formalism for complex systems, Sci. Comput. Program., vol. 8, no. 3, 1987, pp. 231–274.

[59] M.Meo e M. A. Marsan. Resource management policies in GPRS systems, Performance Evaluation, Vol. 56, no. 1-4, Março, 2004, pp. 73-92.

[60] B.Jabbari e F.Fuhrmann. Teletraffic Modeling and Analysis of Flexible Hierarchical Cellular Networks with Speed-Sensitive Handoff Strategy, IEEE Journal on Selected Areas in Communications, Vol. 15, no. 8, Outubro, 1997, pp. 1539-1548.

[61] F. Prihandoko, M. H. Habaebi e B. M. Ali. Adaptive call admission control for QoS provisioning in multimedia wireless networks, Computer Communications, vol. 26, no. 14, Setembro, 2003, pp. 1560-1569

[62] H. K. Pati, R. Mall e I. Sengupta. An efficient bandwidth reservation and call admission control scheme for wireless mobile networks, Computer Communications, vol. 25, no. 1, Janeiro, 2002, pp. 74-83.

[63] J.-H. Lee, T.-H. Jung, S.-U. Yoon, S.-K. Youm e C.-H. Kang. An adaptive resource allocation mechanism including fast and reliable handoff in IP-based 3G wireless networks. IEEE Personal Communications, vol. 7, no. 6, Dezembro, 2000, pp. 42-47.

[64] E. Geraniotis e W.-B. Yang. Admission policies for integrated voice and data traffic in CDMA packet radio networks, IEEE Journal onSelected Areas in Communications, vol. 12, no. 4, Maio, 1994 pp. 654–664.

[65] S. Singh, V. Krishnamurthy e H.V.Poor. Integrated voice/data call admission control for wireless DS-CDMA systems. IEEE Transactions on Signal Processing, vol. 50, no. 6, Junho, 2002 pp. 1483 - 1495

[66] Y. Xiao, C.L.P. Chen e Y. Wang. An optimal distributed call admission control for adaptive multimedia in wireless/mobile networks, *Proceedings* 8th International Symposium on Modeling, Analysis and Simulation of Computer and Telecommunication Systems, 2000, pp. 477– 482.

[67] K.-Min Ahn e S. Kim. Optimal bandwidth allocation for bandwidth adaptation in wireless multimedia networks, Computers & Operations Research, vol. 30, no. 13, Novembro, 2003, pp. 1917–1929.

[68] S. M. Ross. Applied Probability Models With Optimization Applications, Dover Dover Publications, 1992.

[69] H. C. Tijms. Stochastic models: an algorithmic approach, John Wiley & Sons, 1994.

[70] G.V. Zaruba, I. Chlamtac e S.K. Das. A prioritized real time wireless call degradation framework for optimal call mix selection. Mobile Networks and Applications, vol. 7, Abril, 2002, pp. 143–151.

[71] Y. Xiao, C. L. P. Chen e B. Wang. Bandwidth degradation QoS provisioning for adaptive multimedia in wireless/mobile networks, Computer Communications, vol. 25, no. 13, Agosto, 2002, pp.1153-1161.

[72] H. Jiang e W. Zhuang. Quality-of-service provisioning in future 4G CDMA cellular networks. IEEE Wireless Communications, vol. 11, no. 2, Abril, 2004, pp. 48–54.

[73] B. Moon e H. Aghvami. Diffserv extensions for QoS provisioning in IP mobility environments. IEEE Wireless Communications, vol. 10, no. 5, Outubro, 2003, pp.38–44.

[74] Y. Cheng e W. Zhuang. DiffServ resource allocation for fast handoff in wireless mobile Internet, IEEE Communications Magazine, vol. 40, no. 5, Maio, 2002, pp. 130–136.

[75] Y.Wei, C.Lin, F.Ren, R.Raad e E. Dutkiewicz. Dynamic handoff scheme in differentiated QoS wireless multimedia networks, Computer Communications, vol. 27, no. 10, Junio, 2004, pp. 1001-1011.

[76] Y. Guo e H. Chaskar. Class-based quality of service over air interfaces in 4G mobile networks, IEEE Communications Magazine, vol. 40, no. 3, Março, 2002, pp. 132–137.

[77] L. Martin Puterman. Markov Decision Processes: Discrete Stochastic Dynamic Programming. John Wiley & Sons, 1994.

[78] S. Osaki. Applied Stochastic System Modeling, Springer-Verlag,1992.

[79] R. K. Jain, The Art of Computer Systems Performance Analysis: Techniques for Experimental Design, Measurement, Simulation, and Modeling, John Wiley & Sons, 1991.

[80] C. R. L. Francês, M. J. Santana, R. H. C. Santana, R. C. G. S. Orlandi. Tools and Methodologies For Performance Evaluation of Distributed Computing Systems - A Comparison Study, *Proceedings of the* Summer Computer Simulation Conference (SCSC'97), 1997

[81] http://www.cs.bham.ac.uk/ dxp/prism/, fevereiro, 2005.

[82] R C. M. Rodrigues.Inserção de distribuiçãoo do tipo fase em modelos markovianos de decisão. Tese de doutorado, Instituto Nacional de Pesquisas Espaciais, INPE, Brasil, 1998.

[83] http://www.isi.edu/nsnam/ns/, fevereiro, 2005.

[84] http://www.csee.usf.edu/ christen/tools/toolpage.html#simulation, fevereiro, 2005.

[85] Universal Mobile Telecommunications System (UMTS); Selection procedures for the choice of radio transmission technologies of the UMTS (UMTS 30.03 version 3.2.0).

[86] G. H. S. Carvalho, R. M. Rodrigues; C. R. L. Francês, J. C. W. A. Costa, S. V.Carvalho. Modelling and Performance Evaluation of Wireless Networks, Lecture Notes in Computer Science, Heidelberg, Germany, v. 3124, 2004, pp. 595-600.

[87] G. H. S. Carvalho, J. C. W. A. Costa, C. R. L. Frânces, R. C. M. Rodrigues, S. V.Carvalho. Alocação de Recursos em Redes Móveis Celulares Hierárquicas GSM/GPRS, Simpósio Brasileiro de Telecomunicações, 2004, Belém, Pará.

[88] G. H. S. Carvalho, J. C. W. A. Costa, C. R. L. Frânces, R. C. M. Rodrigues, S. V.Carvalho. Modelagem Markoviana da Alocação de Recursos em Redes Móveis Celulares Hierárquicas GSM/GPRS, XXXVI SBPO - Simpósio Brasileiro de Pesquisa Operacional, 2004, São João del-Rei, MG.

[89] G. H. S. Carvalho, J. C. W. A.Costa, C. R. L.Frânces, R. M.Rodrigues, S. V.Carvalho. Modelling and Performance Evaluation of Wireless Networks, International Conference on Telecommunications, 2004, Fortaleza, Ceará.

[90] A. M.Cavalcante, A. M. L.Miranda, J. C. W. A.Costa, C. R. L.Francês, L. A. G. Oliveira, G. P. S.Cavalcante e G. H. S.Carvalho. Proposta para Integração de Sistemas Legados para Aprendizado a Distância: Estudo de Caso em Planejamento de Sistemas Móveis Celulares, Momag (XI Simpósio Brasileiro de Microondas e Optoeletrônica-SBMO e VI Congresso Brasileiro de Eletromagnetismo-CBMag) 2004, São-Paulo - SP.

[91] G. H. S.Carvalho, M.B.L.Santos, J. C. W. A.Costa, Analise de Desempenho da Rede GSM/GPRS,X Simpósio Brasileiro de Microondas e Optoeletrônica (SBMO), 2002, Recife.

[92] G. H. S.Carvalho, M.B.L.Santos, J. C. W. A.Costa, Performance Analysis of GSM-GPRS Network, IEEE International Telecommunications Symposium (ITS2002), 2002, Natal.

[93] G. H. S.Carvalho, A M.Cavalcante, E. S.Lelis, G. P. S.Cavalcante, J. C. W. A.Costa. Software Educacional para Dimensionamento de Sistemas Móveis Celulares, X Simpósio Brasileiro de Microondas e Optoeletrônica (SBMO), 2002, RECIFE

[94] R. M. Rodrigues. Análise da Qualidade de Serviço em Redes GSM/GPRS através de Esquemas de Compartilhamento de Recursos, Dissertação de Mestrado, Programa de Pós-Graduação de Engenharia Elétrica (PPGEE),Universidade Federal do Pará (UFPA), Brasil, 2004.